Wasit Journal of Engineering Sciences

Journal homepage: https://ejuow.uowasit.edu.iq

A novel technique for predicting streamflow: Combining neural network and metaheuristic algorithm

Baydaa Abdul Kareem [1, 2], and Salah L. Zubaidi [2]

Affiliations
[1] Department of Civil Engineering, University of Maysan, Maysan 57000, Iraq;

[2] Department of Civil Engineering, Wasit University, Wasit 52001, Iraq.

Correspondence
Baydaa Abdul Kareem, Department of Civil Engineering, University of Maysan, Maysan 57000, Iraq.
Email: bida.k.z@uomisan.edu.iq

Received
18-December-2022
Revised
28-December-2022
Accepted
25-March-2023

Doi: 10.31185/ejuow.Vol11.Iss2.407

Abstract

Precise streamflow forecasting is crucial when designing water resource planning and management, predicting flooding, and reducing flood threats. This study invented a novel approach for the monthly water streamflow of the Tigris River in Amarah city, Iraq, by integrating an artificial neural network (ANN) with the particle swarm optimization algorithm (PSO), depending on data pre-processing. Historical streamflow data were utilized from (2010 to 2020). The primary conclusions of this study are that data pre-processing enhances data quality and identifies the optimal predictor scenario. In addition, it was revealed that the PSO algorithm effectively forecasts the parameters of the suggested model. Also, the outcomes indicated that the suggested approach successfully simulated the streamflow according to multiple statistical criteria, including R^2, RMSE, and MAE.

Keywords: Streamflow, ANN, SSA, PSO, Amarah.

الخلاصة: يعد التنبؤ الدقيق بالتدفق أمرًا بالغ الأهمية عند تصميم إدارة موارد المياه والتخطيط لها، والتنبؤ بالفيضانات، وتقليل أضرار الفيضانات. ابتكرت هذه الدراسة نهجًا جديدًا للتدفق الشهري للمياه لنهر دجلة في مدينة العمارة، العراق، من خلال دمج شبكة عصبية اصطناعية (ANN) مع خوارزمية تحسين سرب الجسيمات (PSO)، اعتمادًا على معالجة البيانات المسبق. تم استخدام البيانات التاريخية للتدفق من (2010 إلى 2020). الاستنتاجات الأولية لهذه الدراسة هي أن المعالجة المسبقة للبيانات تعزز جودة البيانات وتحدد سيناريو التوقع الأمثل. بالإضافة إلى ذلك، تم الكشف عن أن خوارزمية PSO تتنبأ بفعالية بمعلمات النموذج المقترح. علاوة على ذلك، أشارت النتائج إلى أن النهج المقترح يحاكي بنجاح التدفق بناءً على معايير إحصائية متعددة (على سبيل المثال، R^2، RMSE، MAE).

1. INTRODUCTION

Along with the deepening of global climate change, frequent droughts and floods have negatively influenced social and economic growth. Therefore, increasing requirements for developing and utilizing water resources have been proposed [1]. Where climate change has raised the demand for water and reduced available water supplies, socioeconomic factors like population increase and urbanization will exacerbate freshwater demand issues. Consequently, water shortage is one of the concerns associated with water resources, a formidable obstacle for the world in the twenty-first century [2].

Forecasting of streamflow is one of the primary issues of hydrologists and engineers when designing and managing water resources and constructing projects involving water resources. Forecasts of streamflow over different periods can provide crucial information on the feasibility of developing and operating water facilities and the availability of water resources [3]. As a result, numerous approaches have been employed and effectively utilized for predicting streamflow. Traditional models are one of these approaches. The fundamental difficulty of traditional models is incapable of capturing the nonlinear data being captured, resulting in a weak level of accuracy of forecasting [4]. Zhang, Li [5] observed in their investigation of univariate streamflow forecasts that artificial intelligence (AI) models are extensively employed due to their simplicity and minimum data requirements. Also, it performs better than conventional models. The most often employed AI models in streamflow forecasting are multi-layer perceptron (MLP) [6], neural-based fuzzy inference systems (ANFIS) [7], and artificial neural networks (ANN) [8].

The ANN tool could be especially advantageous, and ANN can predict the parameters of a model with reasonable precision using historical data. It is capable of simulating any physical occurrence [9]. Hence, it is appropriate for a vast array of applications in hydrological, such as Tiu, Huang [10] for water level, and Ömer Faruk [11] forecasting water quality.

The necessity for enhanced dependability and data-driven approach precision and efficiency prompted the development of hybrid models. Typically, the hybrid model integrates more than one approach. One represents the primary model and the other represents post-or pre-processing procedures [12]. Hybrid models are reliable, and informative and were utilized effectively in various hydrological sectors, such as rainfall forecasting [13] and water demand forecasting [14].

Recently, Ibrahim, Huang [4] examined simulating the models of streamflow and indicated that machine learning (ML) methods must be optimized in combination in tandem to achieve the optimal result. In addition, the study suggested that:

1. It is advisable to do the data pre-processing phase as effectively as feasible to avoid data noise. It proposed that greater emphasis should be placed on identifying the optimal predictor combination.
2. The ANN's learning rate coefficient is one of the most influential factors in the outcomes and performance of the model. Therefore, it is advised that these hyperparameters be chosen and tuned using metaheuristic algorithms.

This research seeks to create a new approach to precisely predict monthly streamflow for the medium-term using historical streamflow data. The key objectives of the current research are to:

1. Utilizing the data pre-processing steps to denoise data by single spectrum analysis (SSA) and choose the best model input scenario by average mutual information (AMI).
2. The particle swarm optimization algorithm (PSO) optimizes the ANN model to determine the optimum ANN parameters.
3. Assess the performance of the PSO-ANN approach for simulating streamflow.
4. Provide policymakers with a better scientific perspective on streamflow forecasts.

2. Area of Study and Data Set

Al-Amarah is the southern Iraqi capital of the Maysan Governorate, located 400 km southeast of Baghdad (Figure 1). The region's area is 16,702 square kilometers and it has a population of (1,106,208) million [15]. The longitudes and latitudes for the study area are (46°20’–48°05′ E) and (31°10′–32°50′ N), respectively. Seasons in the Al- Amara region range from hot, dry summers and chilly, wet winters. The typical duration of the spring and fall seasons is two weeks [16]. The Iraqi meteorological agency stated that the winter continues for five months, from November until March. The rest months are deemed summer, where June, July, and August typically record the greatest temperatures [17].

The Directorate of Water Resources in Maysan city supplied historical monthly streamflow (m^3/s) data from 2010 to 2020 (11 years), which were used to develop and evaluate the model. The boxplot and raw time series for streamflow are shown in figure 2.

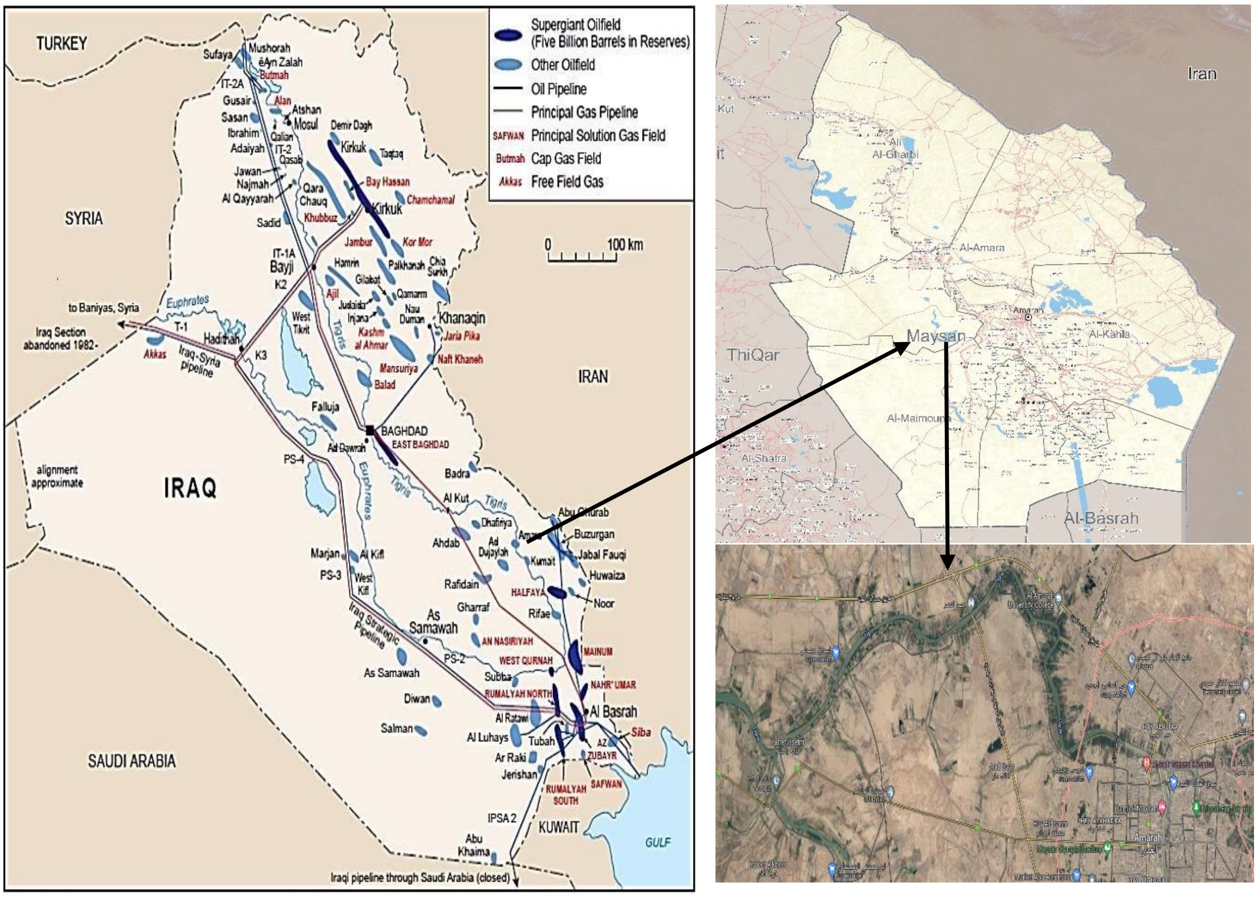

Figure 1 Study area, Amarah city, Iraq.

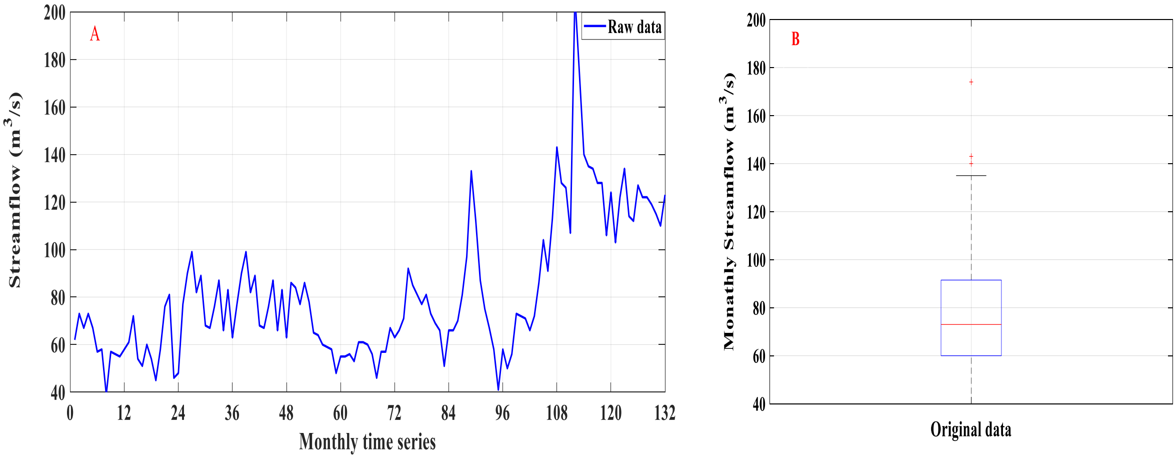

Figure 2 Streamflow (A) Monthly time series, (B) Box plot.

3. Methodology

3.1. Pre-processing Data

The pre-processing data includes three approaches: normalization, cleaning, and selecting the optimal model input.

3.1.1. Normalization

In this study, the goal of normalization is to minimize the effect of outliers and make the time series have a normal or near to normal distribution [18]. To normalize the time series the natural logarithm method was employed [19]. This study uses the SPSS 24 statistical software.

3.1.2. Cleaning

Cleaning data is critical for identifying and modifying outliers and ignoring time series noise. Consequently, this study employed the box and whisker technique to remove outliers from the data. In addition, the SSA was employed to denoise time series[18]. SSA is a valuable pre-processing approach for time series prediction when used with neural networks. It is used for linear and nonlinear data [20, 21]. This strategy was utilized successfully in various fields, such as hydrology [13] and stochastic process forecast [22]. More details on SSA can be found in Zhigljavsky [23].

3.1.3. Selection of Explanatory Factors

Choosing explanatory variables to represent the model input data is vital in developing a model for accurate prediction [24]. The Mutual Information (MI) method has been used in this study to determine the optimal explanatory variables. This method permits choosing components with the most significant mutual information and the highest correlation [25].

3.2.The Particle Swarm Optimisation (PSO)

The current literature trend is that nature-inspired metaheuristic algorithms can solve various issues and achieve highly effective outcomes [26]. PSO is a metaheuristic method that monitored the social and cooperative behavior of birds and fish. Additionally, PSO is a population-based algorithm that optimizes a parallel group of swarms by employing. PSO travels at a predetermined rate through the problem area in search of the optimal solution [24]. PSO is being implemented successfully in various fields; For example, wireless sensor networks Dash, Panigrahi [27], and single server optimization Alharkan, Saleh [28]. More information on PSO can be found in Poli [29].

3.3.Artificial Neural Network (ANN)

The ANN is a data-processing approach that mimics human brain operations by employing identical connections and behavior as biological neurons [25]. The multi-layer feedforward neural network (MLFFNN) was used that utilized the Levenberg–Marquardt (LM) technique for training the ANN model [12]. The presumed ANN structure includes four layers: the first layer represents the input (i.e., lags), followed by two hidden layers with sigmoidal activation functions, and the fourth layer represents the output (i.e., streamflow). Since the time-consuming trial-and-error method does not always result in the ideal solution. Consequently, metaheuristic algorithms were incorporated into the ANN to choose the optimal hidden neurons' number and learning rate value to avoid under - and overfitting the model [30].

3.4.Model Performance Assessment

This study used three statistical criteria to evaluate the model's efficacy in forecasting streamflow. Firstly, mean absolute error (MAE), secondly, root mean squared error (RMSE), and finally, coefficient of determination (R^2) are the metrics used. To estimate them, use the following formula [31, 32]:

$$MAE = \frac{\sum_{i=1}^{N} |O_i - F_i|}{N} \tag{1}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (O_i - F_i)^2}{N}} \tag{2}$$

$$R^2 = \left[\frac{\sum_{i=1}^{N} (O_i - \bar{O}_i)(F_i - F_i)}{\sqrt{\sum (O_i - \bar{O}_i)^2 \sum (F_i - \bar{F}_i)^2}}\right]^2 \tag{3}$$

Where:

O_i: measure streamflow,

F_i: predicted streamflow,

N: sample size,

$\overline{F_i}$: average of predicted streamflow, and

$\overline{O_i}$: average of measure streamflow.

In addition, this work performed a graphical test during the validation stage to check the PSO-ANN model's ability to mimic the streamflow data set.

4. Results and Discussion

4.1.Development Model Input

Tabachnick and Fidell [33], recommended to normalizing data to decrease the influence of outliers. After that, the outliers that remained following the transformation (if detected) were then rescaled. Then, SSA was employed to get streamflow time series data devoid of noise (which was gained by analyzing the normalized and cleaned data into three components). Figure 3 demonstrates the normalized and cleaned data (first component), the updated data (second component), and two components of noise (representing the third and fourth components). The monthly streamflow data were enhanced by using the pre-processed data, where the correlation coefficient rose dramatically from 0.84 to 0.97 for the first lag. Additionally, the correlation coefficients of the remaining denoise data were 0.91, 0.82, 0.75, and 0.68, respectively.

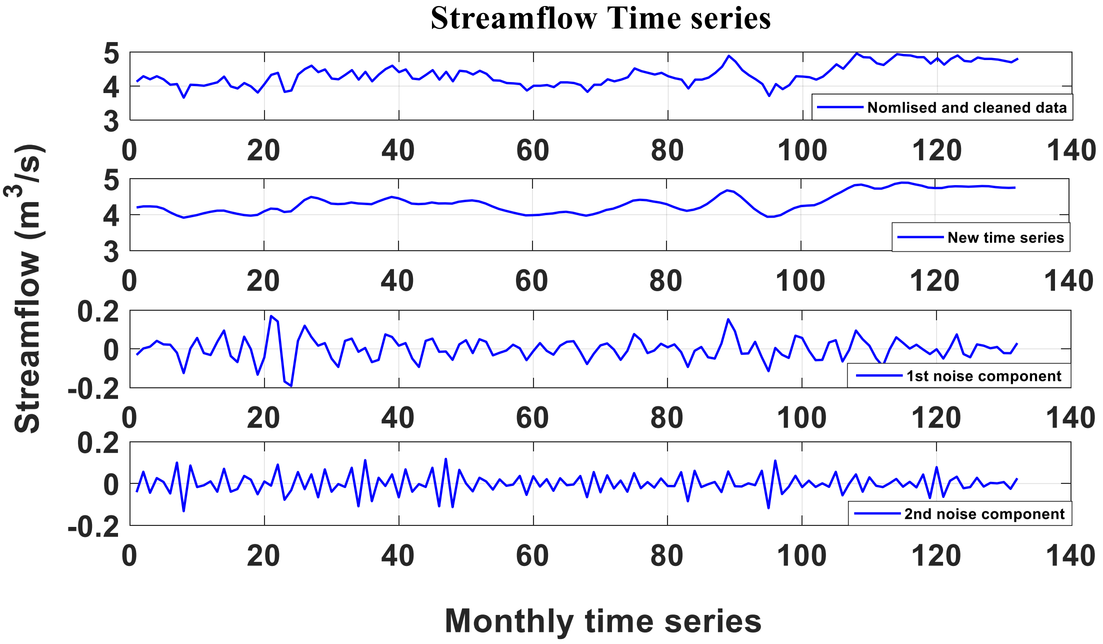

Figure 3 Normalized and cleaned streamflow data (top row) and the three components acquired by SSA (2nd to 4th rows).

Three shapes of box plots (i.e., normalized, cleaned, and denoised streamflow data) are demonstrated in figure 4. From the figure can see that the normalized data contained two outliers. Also, the shape of the data did not alter significantly from the cleaned form. All three shapes' median and upper and lower quartiles were almost identical. However, the denoised time series shape had shorter upper and lower whiskers than the other two (normalized and cleaned time series).

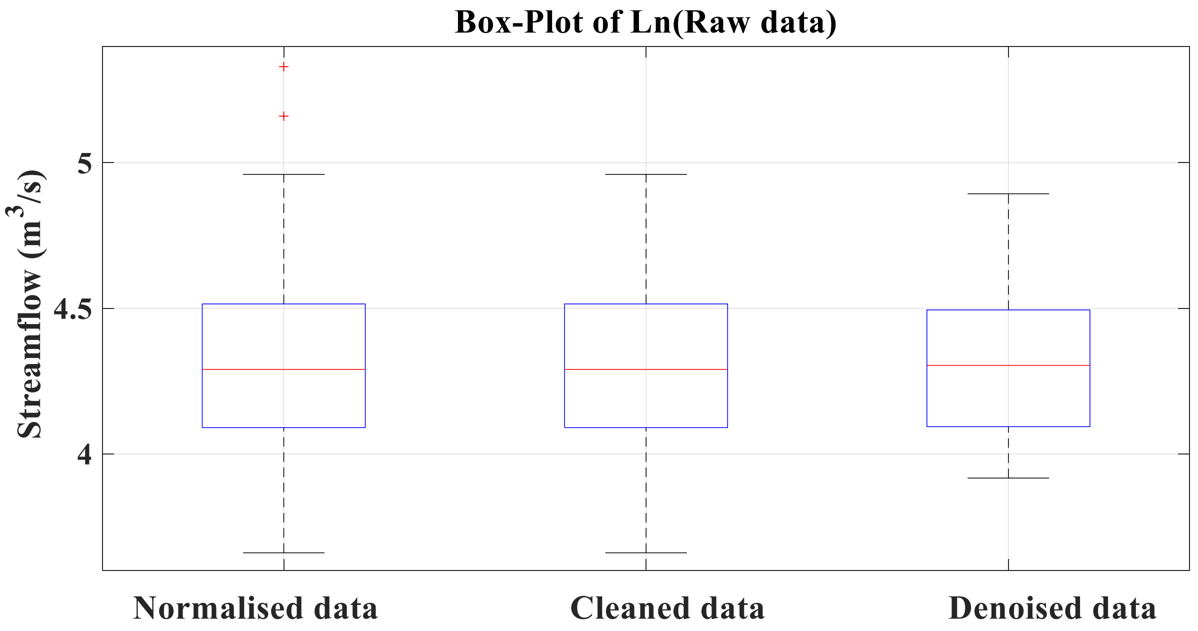

Figure 4 The box plot shows normalized, cleaned, and denoised data distribution.

Following figure 5 of AMI, five lags of monthly streamflow time series (Lagt 1 through Lagt 5) were employed to estimate future streamflow according to the literature [34].

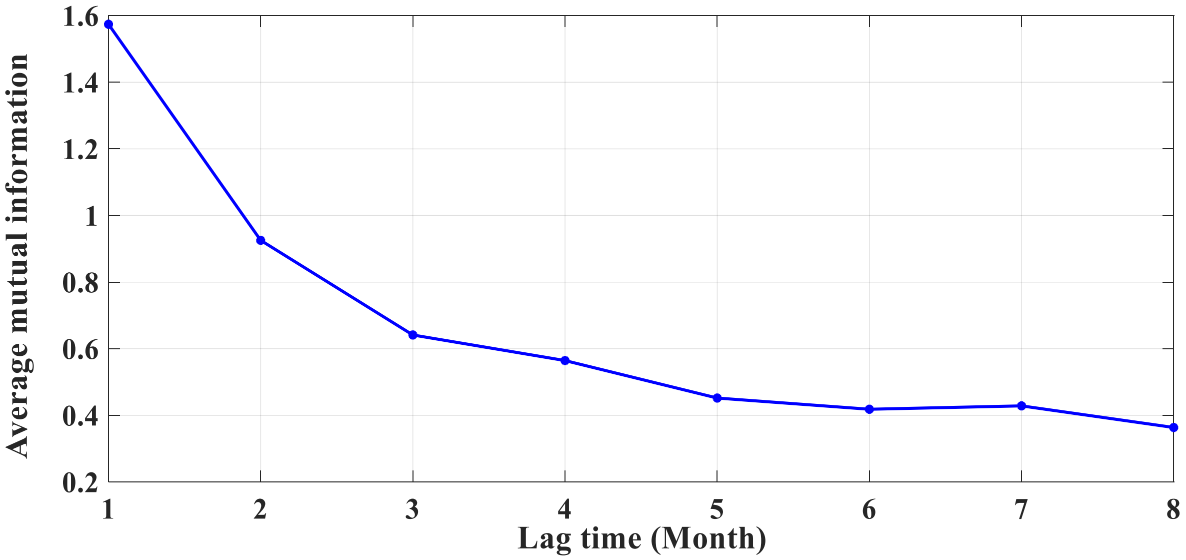

Figure 5 Lag time (month) obtained by AMI

Then, the time series were divided into training (70%), testing (15%), and validation (15%).

4.2.Analysis of the PSO Technique

The ANN model must incorporate a metaheuristic technique to determine the optimum ANN model hyperparameters (N1, N2, and Lr). Consequently, the PSO algorithm has been combined with an ANN model. To determine the least fitness function (MSE), different swarm sizes (i.e., 10, 20, 30, 40, and 50) were utilized five times, each with 200 iterations. The optimum fitness value for each swarm is depicted in figure 6.

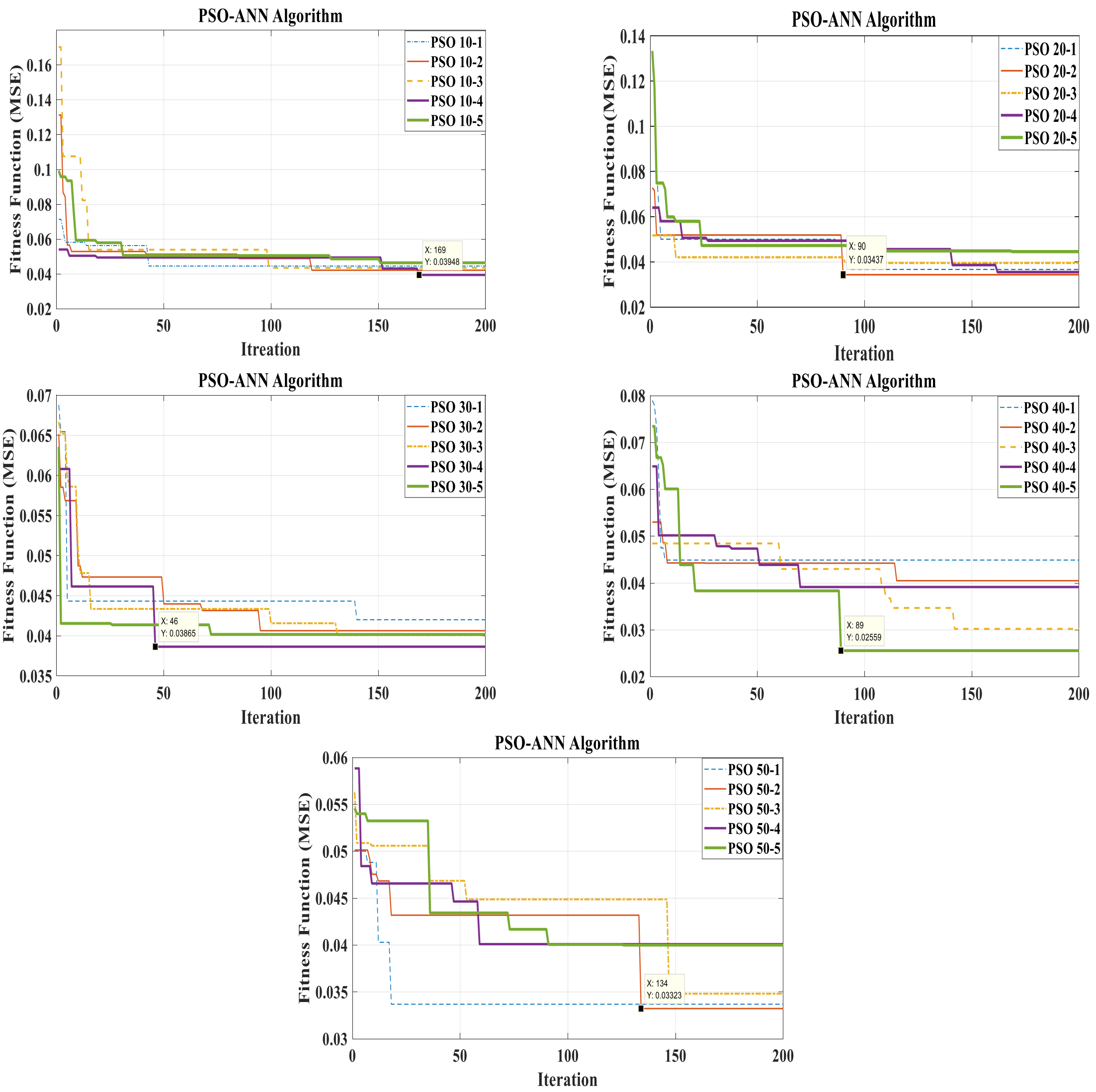

Figure 6 Performance of PSO algorithm.

Figure 7 illustrates that the swarm size (40-5) provided the best result for the suggested model (MSE = 0.02559, after 89 iterations). Streamflow simulation has been enhanced by employing the PSO algorithm's output to improve ANN capabilities. The best swarm size resulted in the following hyperparameter values for the ANN model: Lr = 0.2941; N1 = 4; N2 = 5.

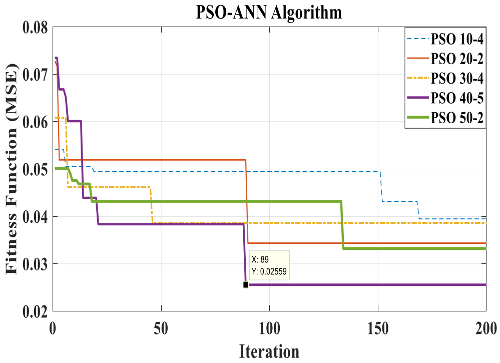

Figure 7 Fitness function versus iteration (PSO).

4.3.Performance Evaluation

After establishing the appropriate N1, N2, and Lr values, the ANN model was developed to mimic streamflow. Several statistical indicators were employed to assess the performance of the created technique. The R^2, RMSE, and MAE can be seen in Table 1. According to Dawson, Abrahart [35], the PSO-ANN model demonstrated accurate streamflow prediction.

Table 1 Criteria for evaluating PSO performance in the validation stage.

Output	Algorithm-Model	MAE	R^2	RMSE
Streamflow	**PSO-ANN**	**0.055**	**0.88**	**0.071**

In addition, a graphical test was utilized during the validation stage to scrutinize the hybrid method's best fit. The observed and predicted streamflow data using PSO-ANN are shown in Figure 8. The PSO-ANN predicted data approximately match the pattern and frequency of the measured data.

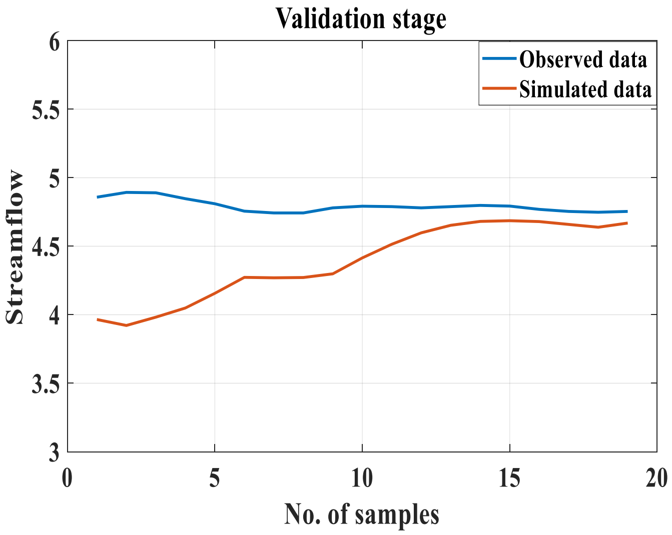

Figure 8 The observed and simulated data at the validation stage.

Overall, the above findings show that:

1. These outcomes demonstrate the potential effectiveness of SSA to denoise data and AMI approaches to selecting explanatory factors scenarios without trespassing the multicollinearity premise.
2. PSO is a dependable algorithm combining the ANN method for monthly streamflow prediction.
3. Multiple statistical tests demonstrated that the suggested technique accurately forecasted monthly streamflow for the medium term.
4. This research demonstrates the importance of future research into new hybrid forecasting techniques at various time - scales.

5. Conclusions

Forecasting streamflow is essential for planning and integrated water resources management. In this study, a novel approach has been developed and analyzed for predicting streamflow based on multiple prior delays. The SSA has been used to remove the noise from raw data, and AMI has been used to select the best prediction lags. In addition, to select the optimal hyperparameters, the PSO technique was used to incorporate the ANN model. Historical data for streamflow were utilized from 2010 to 2020 for the Tigris River, Amara, Iraq. The current study uncovered that pre-processing strategies are essential to enhance raw data and select the best delay scenario. The outcomes show that the suggested approach is a dependable and skillful method for predicting streamflow by yielding $R^2 = 0.88$, with an RMSE equal to 0.071. These results can provide managers and policymakers with scientific, more precise insights about the streamflow, resulting in the enhanced irrigation sector, service, and resource management in the Governorate of Maysan. For future studies, further research and testing utilizing the hybrid technique to emulate the streamflow influenced by climate variables are powerfully encouraged because extreme weather is predicted to increase in the future.

REFERENCES

1. Li, F.-F., et al., *Daily Streamflow Forecasting Based on Flow Pattern Recognition.* Water Resources Management, 2021. **35**(13): p. 4601-4620.
2. Ethaib, S., et al., *Evaluation water scarcity based on GIS estimation and climate-change effects: A case study of Thi-Qar Governorate, Iraq.* Cogent Engineering, 2022. **9**(1): p. 2075301.
3. Yaseen, Z.M., et al., *Implementation of Univariate Paradigm for Streamflow Simulation Using Hybrid Data-Driven Model: Case Study in Tropical Region.* IEEE Access, 2019. **7**: p. 74471-74481.
4. Ibrahim, K.S.M.H., et al., *A review of the hybrid artificial intelligence and optimisation modelling of hydrological streamflow forecasting.* Alexandria Engineering Journal, 2021. **61**(1): p. 279-303.
5. Zhang, Y., et al., *Multi-models for SPI drought forecasting in the north of Haihe River Basin, China.* Stochastic Environmental Research and Risk Assessment, 2017. **31**(10): p. 2471-2481.
6. Mohammadi, B., et al., *Developing Novel Robust Models to Improve the Accuracy of Daily Streamflow Modeling.* Water Resources Management, 2020. **34**(10): p. 3387-3409.
7. Vatanchi, S.M., et al., *A Comparative Study on Forecasting of Long-term Daily Streamflow using ANN, ANFIS, BiLSTM, and CNN-GRU-LSTM.* 2022. **1**.
8. Meshram, S.G., et al., *Streamflow Prediction Based on Artificial Intelligence Techniques.* Iranian Journal of Science and Technology, Transactions of Civil Engineering, 2021. **46**: p. 2393–2403.
9. Mohammed, S.J., et al., *Hybrid Technique to Improve the River Water Level Forecasting Using Artificial Neural Network-Based Marine Predators Algorithm.* Advances in Civil Engineering, 2022. **2022**: p. 1-14.
10. Tiu, E.S.K., et al., *An evaluation of various data pre-processing techniques with machine learning models for water level prediction.* Natural Hazards, 2022. **110**: p. 121-153.
11. Ömer Faruk, D., *A hybrid neural network and ARIMA model for water quality time series prediction.* Engineering Applications of Artificial Intelligence, 2010. **23**(4): p. 586-594.
12. Baydaa Abdul Kareem 1, Salah L. Zubaidi 2 , Hussein Mohammed Ridha 3, Nadhir Al-Ansari 4,* and a.N.S.S. Al-Bdairi, *Applicability of ANN Model and CPSOCGSA Algorithm forMulti-Time Step Ahead River Streamflow Forecasting.* hydrology, 2022. **9**: p. 171.
13. Ouyang, Q. and W. Lu, *Monthly Rainfall Forecasting Using Echo State Networks Coupled with Data Preprocessing Methods.* Water Resources Management, 2017. **32**(2): p. 659-674.
14. Zubaidi, S.L., et al., *Short-Term Urban Water Demand Prediction Considering Weather Factors.* Water Resources Management, 2018. **32**(14): p. 4527-4542.
15. Hayder Oleiwi Shami Al Saidia, N.R.A., *Analysis of Human Development Indicators in the Maysan Governorate.* International Journal of Innovation, Creativity and Change, 2020. **14**(2): p. 1296.
16. Abood, R.H. and R.R. Mahmoud, *Drought Assessment Using Gis And Meteorological Data In Maysan Province /Iraq.* International Journal of Civil Engineering and Technology, 2018. **9**(6): p. 516-524.
17. Edan, M.H., R.M. Maarouf, and J. Hasson, *Predicting the impacts of land use/land cover change on land surface temperature using remote sensing approach in Al Kut, Iraq.* Physics and Chemistry of the Earth, Parts A/B/C, 2021. **123**: p. 103012.
18. Alawsi, M.A., et al., *Tuning ANN Hyperparameters by CPSOCGSA, MPA, and SMA for Short-Term SPI Drought Forecasting.* Atmosphere, 2022. **13**(9): p. 1436.
19. Zubaidi, S.L., et al., *A Method for Predicting Long-Term Municipal Water Demands Under Climate Change.* Water Resources Management, 2020. **34**(3): p. 1265-1279.
20. Golyandina, N. and A. Zhigljavsky, *Singular Spectrum Analysis for Time Series*. 2020, Springer.

21. Zubaidi, S.L., et al., *A Novel approach for predicting monthly water demand by combining singular spectrum analysis with neural networks.* Journal of Hydrology, 2018. **561**: p. 136-145.
22. Khan, M.A.R. and D.S. Poskitt, *Forecasting stochastic processes using singular spectrum analysis: Aspects of the theory and application.* International Journal of Forecasting, 2017. **33**(1): p. 199-213.
23. Zhigljavsky, A., *Singular spectrum analysis for time series: Introduction to this special issue.* Statistics and its Interface, 2010. **3**(3): p. 255-258.
24. Mohammed, S.J., et al., *Application of hybrid machine learning models and data pre-processing to predict water level of watersheds: Recent trends and future perspective.* Cogent Engineering, 2022. **9**(1).
25. Zubaidi, S.L., et al., *Urban Water Demand Prediction for a City That Suffers from Climate Change and Population Growth: Gauteng Province Case Study.* Water, 2020. **12**(7): p. 1885.
26. Zubaidi, S., et al., *A Novel Methodology for Prediction Urban Water Demand by Wavelet Denoising and Adaptive Neuro-Fuzzy Inference System Approach.* Water, 2020. **12**(6): p. 1628.
27. Dash, M., T. Panigrahi, and R. Sharma, *Distributed parameter estimation of IIR system using diffusion particle swarm optimization algorithm.* Journal of King Saud University - Engineering Sciences, 2019. **31**(4): p. 345-354.
28. Alharkan, I., et al., *Tabu search and particle swarm optimization algorithms for two identical parallel machines scheduling problem with a single server.* Journal of King Saud University - Engineering Sciences, 2020. **32**(5): p. 330-338.
29. Poli, R., *Analysis of the Publications on the Applications of Particle Swarm Optimisation.* Journal of Artificial Evolution and Applications, 2008. **2008**: p. 1-10.
30. Alawsi, M.A., et al., *Tuning ANN Hyperparameters by CPSOCGSA, MPA, and SMA for Short-Term SPI Drought Forecasting.* Atmosphere, 2022. **13**(9).
31. Mohammadi, B. and S. Mehdizadeh, *Modeling daily reference evapotranspiration via a novel approach based on support vector regression coupled with whale optimization algorithm.* Agricultural Water Management, 2020. **237**: p. 106145.
32. Mohammadi, B., et al., *Adaptive neuro-fuzzy inference system coupled with shuffled frog leaping algorithm for predicting river streamflow time series.* Hydrological Sciences Journal, 2020. **65**(10): p. 1738-1751.
33. Tabachnick, B.G. and L.S. Fidell, *Using Multivariate Statistics*, ed. S. Edition. 2013: Pearson: Boston, MA, USA.
34. Aldrich, C. and L. Auret, *Unsupervised process monitoring and fault diagnosis with machine learning methods*. Vol. 16. 2013: Springer.
35. Dawson, C.W., R.J. Abrahart, and L.M. See, *HydroTest: A web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts.* Environmental Modelling & Software, 2007. **22**(7): p. 1034-1052.