IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

International Journal of Sustainable Development and Planning
Vol. 16, No. 3, June, 2021, pp. 497-501
Journal homepage: http://iieta.org/journals/ijsdp

Application of Deep Learning Method for Daily Streamflow Time-Series Prediction: A Case Study of the Kowmung River at Cedar Ford, Australia

Sarmad Dashti Latif[1*], Ali Najah Ahmed[2]

[1] Civil Engineering Department, College of Engineering, Komar University of Science and Technology, Sulaimany 46001, Kurdistan Region, Iraq
[2] Institute for Energy Infrastructure (IEI), Universiti Tenaga Nasional (UNITEN), Kajang 43000, Selangor Darul Ehsan, Malaysia

Corresponding Author Email: Sarmad.latif@komar.edu.iq

https://doi.org/10.18280/ijsdp.160310

Received: 12 May 2020
Accepted: 9 December 2020

Keywords:
streamflow prediction, artificial intelligence, deep learning, machine learning, LSTM, random forest, tree boost

ABSTRACT

Sustainable management of water supplies faces a comprehensive challenge due to global climate change. Improving forecasts of streamflow based on erratic precipitation is a significant activity nowadays. In recent years, the techniques of data-driven have been widely used in the hydrological parameter's prediction especially streamflow. In the current research, a deep learning model namely Long Short-Term Memory (LSTM), and two conventional machine learning models namely, Random Forest (RF), and Tree Boost (TB) were used to predict the streamflow of the Kowmung river at Cedar Ford in Australia. Different scenarios proposed to determine the optimal combination of input predictor variables, and the input predictor variables were selected based on the auto-correlation function (ACF). Model output was evaluated using indices of the root mean square error (RMSE), and the Nash and Sutcliffe coefficient (NSE). The findings showed that the LSTM model outperformed RF and TB in predicting the streamflow with RMSE and NSE equal to 102.411, and 0.911 respectively. for the LSTM model. The proposed model could adopt by hydrologists to solve the problems associated with forecasting daily streamflow with high precision. This study may not be generalized because of the geographical condition and the nature of the data for each location.

1. INTRODUCTION

Streamflow is a dynamic process which is not easily predictable. This process is defined by a huge parameter numbers, such as evapotranspiration, temperature, precipitation, land use, and is characterized by a non-linear relationship between the flow and the characteristics of its water body. Models for predicting streamflow can be categorized as physics-based, and data-driven models. Physically-based models are data-intensive and include a wide range of parameters based on rainfall quantity, intensity and distribution, physiography of the watershed, land use, and human activities. However, consistent model performance is not always guaranteed, depending on the area of research and the particular intent. These parameters are not easy to obtain and it is extremely difficult for many watersheds to obtain accurate and adequate data, which results in low model results [1]. Over the last decades, the hydrologists have popularly used the soft computing methods for streamflow modeling. Since Artificial Neural Network (ANN) has ability to model linear and non-linear systems even without making any assumption, the models of ANN were widely used in various water science subjects [2-6].

ANN have been widely used to solve a broad range of hydrological problems including rainfall-runoff modelling [7, 8], hydrological time-series modeling and reservoir operations [9, 10], groundwater modeling [11-14], and regional flood frequency analysis [15]. ANN-based hydrological prediction models can effectively define the input-output relationship in hydrological systems which can address the shortcomings of the traditional parameterized modeling approach. ANNs can also provide reliable outputs for complex rainfall-runoff modeling-using historical data research. Thus, in the past decade, ANNs have become popular and are generally used in streamflow predictions to lessen flood-induced damage. Yuan et al. [16] investigated the accuracy of short-term hybrid memory (LSTM), for which the ant lion optimizer (ALO) algorithm optimized its parameters by predicting the monthly streamflow. The results showed that the historical monthly flux was calculated more accurately when using LSTM-ALO compared to other models.

In computer science, machines demonstrate AI, as opposed to the animals and humans representing natural intelligence. AI technologies have been widely used in recent years to address a great range of the issues of water engineering. These include gene expression programming (GEP), evolutionary polynomial regression (EPR), model tree (MT), adaptive neuro-fuzzy inference system (ANFIS), extreme learning machine and support vector machine. Different investigations have also used AI approaches particularly for river flow forecasting. However, developing a detailed model for forecasting flow is a challenge, as many (nonlinear) variables in the catchment influence rainfall – runoff processes. Using the raw data directly for modeling may not yield permissible results, but applying a pre-processing method can improve model performance [17].

Although many studies have applied ensemble techniques to the hydrological sector, studies on the sensitivity of artificial intelligence (AI) models are still in short supply. The main objective in this paper is to predict streamflow at river Kowmung. At Kowming river, three methods are used to estimate daily streamflow. The study area and hydrological data with the LSTM, RF, and TB methodologies are briefly listed in the section below. Section 3 displays the results of the proposed models and their comparisons. The conclusion of the study is explained in section 4.

2. MATERIALS AND METHODS

2.1 Study area and data

The Kowmung river at Cedar Ford is selected as a study area for this project (Figure 1). The Kowmung river is located in the Hawkesbury Nepean catchment near the Warragamba dam at New South Wales (NSW) in Australia. The majority of the Kowmung River's 80-kilometer stretch lies within Kanangra-Boyd National Park. Blue Mountains National Park lies within the river's lower reaches. For more than 100 years, the river and the catchment have drawn enthusiastic interest from nature lovers but have also been the site of natural resource use and significant mining and forestry proposals. The subcatchment occupies some 76000 hectares, and Kanangra-Boyd National Park is home to just under 75% of the subcatchment. The remaining land is either rural freehold or pine plantations maintained by State Forests NSW. The subcatchment includes the headwaters of the Kowmung river, which provides potable water to Sydney 's major water storage at Lake Burragorang (Warragamba Dam) along with the Kanangra and Jenolan rivers [18]. The study area is shown in Figure 1.

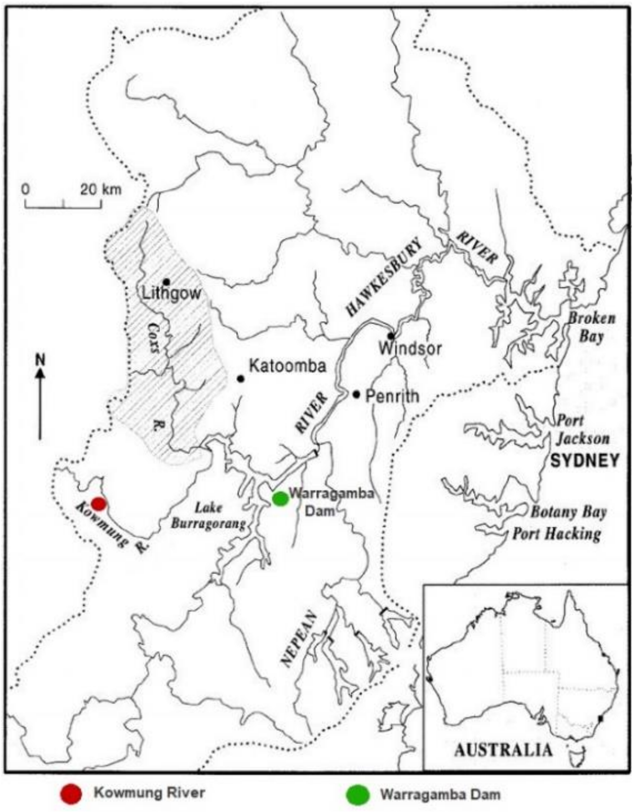

Figure 1. The map of the selected study area [19]

The daily streamflow is collected at Kowmung river at cedar Ford next to Lake Burragorang near Warragamba dam from 1/1/2008 to 1/7/2017 by WaterNSW [20]. Figure 2 shows the daily streamflow of Kowmung river.

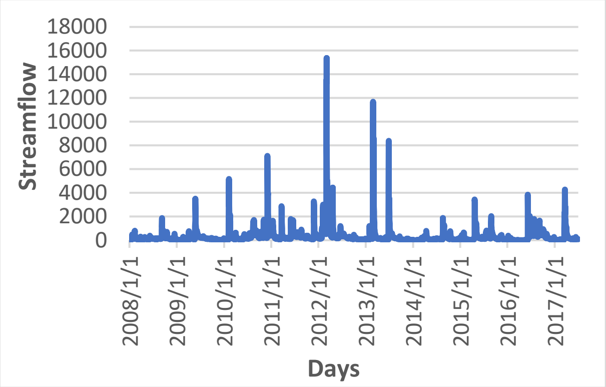

Figure 2. The daily streamflow of Kowmung river from 1/1/2008 to 1/7/2017

2.2 LSTM

In 1997, Hochreiter and Schmidhuber [21] implemented LSTM to solve the gradient blowing up or disappearing problem, which used memory cells and gates to monitor the long-term information that was stored in the network or held away.

$$g_t = \sigma(U_g x_t + W_g h_{t-1} + b_f) \tag{1}$$

$$i_t = \sigma(U_i x_t + W_i h_{t-1} + b_i) \tag{2}$$

$$\tilde{c}_t = \tanh(U_c x_t + W_c h_{t-1} + b_c) \tag{3}$$

$$c_t = g_t * c_{t-1} + i_t * \tilde{c}_t \tag{4}$$

$$o_t = \sigma(U_o x_t + W_o h_{t-1} + b_o) \tag{5}$$

$$h_t = o_t * \tanh\,(c_t) \tag{6}$$

U and W are input weights in various gates: gate input (i_t), gate modulate input ($\tilde{c}_t$), gate forget (g_t), and gate output (o_t). b is a bias function, c_t t is a cell state, h_t is a hidden condition. Both of these controllers decide how much information from the last loop should be obtained, and how much to transfer to the new state.

2.3 Random forest (RF)

RF is a forest created by many decision trees. For each split, the difference is the random subset among the RF and decision tree. This model is developed to solve the issues of classification and regression. For one of the classification modeling classes, a set of forecaster values is used by RF. Alternatively, the target variable is calculated to be the random wood, depending on the regression modeling predictors. The Single Tree Ensemble votes to the most common class for modelling classification. In the regression analysis the results for the target variable are calculated on average, Eq. (7).

$$\text{Random Forest Prediction} = \frac{1}{k}\sum_{k=1}^{k} k^{th} \tag{7}$$

where, k represents the forest trees themselves [22].

2.4 Tree Boost (TB)

The TB algorithm [23] was developed by Jerome Friedman. The model was developed to increase the precision of the decision tree model with the application of the boosting algorithm. Boosting often applies a predictive feature in a series, and adds each test result to boost the accuracy of a sample. A number of decision trees make up TB. The TB algorithm can be described as:

$$\begin{aligned}\mathrm{Target} = \mathrm{E} + \mathrm{C}_1 \text{ x } \mathrm{T}_{r1} + \mathrm{C}_2 \text{ x } \mathrm{T}_{r2}\ (\mathrm{M}) \\ + \cdots \mathrm{C}_n \text{ x } \mathrm{T}_{rn}\ (\mathrm{M})\end{aligned} \tag{8}$$

where, E is the sequence starting value, which represents the target variable's mean value; M is a pseudo-residual asset value matrix, T_{r1} (M), T_{r2} (M), ... T_{rn} (M) are trees equipped with the residual pseudo, and C_1, C_2, ... C_n is the coefficients of the predicted tree node values which are calculated by the algorithm of TB.

2.5 Determination of model inputs

One of the key problems in hydrological modeling is the determinations of the optimal input variables. The Auto-Correlation (ACF) method is used in order select the best input variables for each model. Daily streamflow as a target variable was the cross-correlation function for the Kowmung river, and the input variables were calculated based on different time lags. ACF results for the proposed models are shown in Figure 3. The results of ACF in Figure 3 showed a significant correlation between the daily streamflow data for the Kowmung river for the time lags, (Q_{t-1}, Q_{t-2}, Q_{t-3}, and Q_{t-4}). In this research, four different input's variables combinations were proposed for the models. The input variables' combinations of the four proposed models (Model A, Model B, Model C, and Model D) are summarized in Table 1.

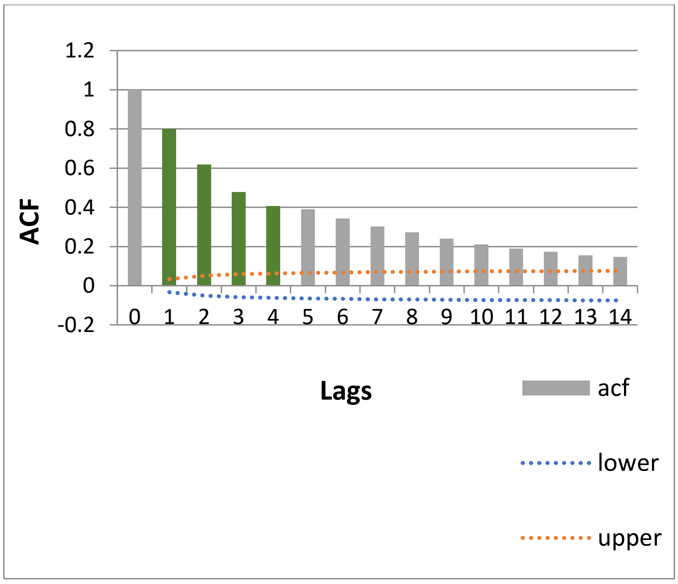

Figure 3. ACF value for the selected models

Table 1. Model combinations

Model	Target Variable	Input Combination
Model A	Q_t	Q_{t-1}
Model B	Q_t	Q_{t-1}, Q_{t-2}
Model C	Q_t	Q_{t-1}, Q_{t-2}, Q_{t-3}
Model D	Q_t	Q_{t-1}, Q_{t-2}, Q_{t-3}, Q_{t-4}

2.6 Performance criteria

RMSE and NSE were used for the performance criteria of this study:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathrm{Qo}-\mathrm{Qp})^2} \tag{9}$$

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}(\mathrm{Qo}-\mathrm{Qp})^2}{\sum_{i=1}^{n}(\mathrm{Qo}-\overline{\mathrm{Qo}})^2} \tag{10}$$

The mean value of the observed streamflow is where *QP* and *Qo* are observed and expected streamflow values, respectively $\overline{\mathrm{Qo}}$. The optimal NSE value is 1. The RMSE index describes the average error range by giving greater weight to large errors.

3. RESULTS AND DISCUSSION

The three models proposed in this study were used to establish the optimal models producing daily streamflow. In this analysis, Table 2 summarized the results of the statistical indices for each model. Table 2 shows that model B which used (Q_{t-1}, and Q_{t-2}) daily streamflow of the Kowmung river as the input variable is the optimal input combination model B compared to model A, model C, and model D for LSTM and TB models. However, Model D is the optimal combination of inputs for the RF model. The Model B was chosen to compare the accuracy of the three models proposed to predict daily streamflow to the Kowmung river. The results show that the LSTM model outperformed RF and TB in predicting the streamflow at Kowmung river with RMSE and NSE equal to 102.411, and 0.911 respectively. The optimal results are highlighted in Table 2. Furthermore, TB outperformed RF model. In Model B, the RMSE is 102.411, NSE is 0.911 for the LSTM, and RMSE is 368.214, NSE= 0.733 for TB model. In Model D, RMSE is 482.123, NSE is 0.542 for the RF model. Figure 4 provides a comparison of the observed and predicted streamflow of the daily time-series, and scatter plot for the proposed models.

Table 2. The statistical indices' results for the proposed models

Models	LSTM		Random Forest		Tree Boost	
	RMSE	NSE	RMSE	NSE	RMSE	NSE
Model A	109.740	0.891	496.978	0.515	374.352	0.724
Model B	**102.411**	**0.911**	492.239	0.522	**368.214**	**0.733**
Model C	112.078	0.873	522.234	0.462	377.615	0.719
Model D	110.507	0.887	**482.123**	**0.542**	376.927	0.720

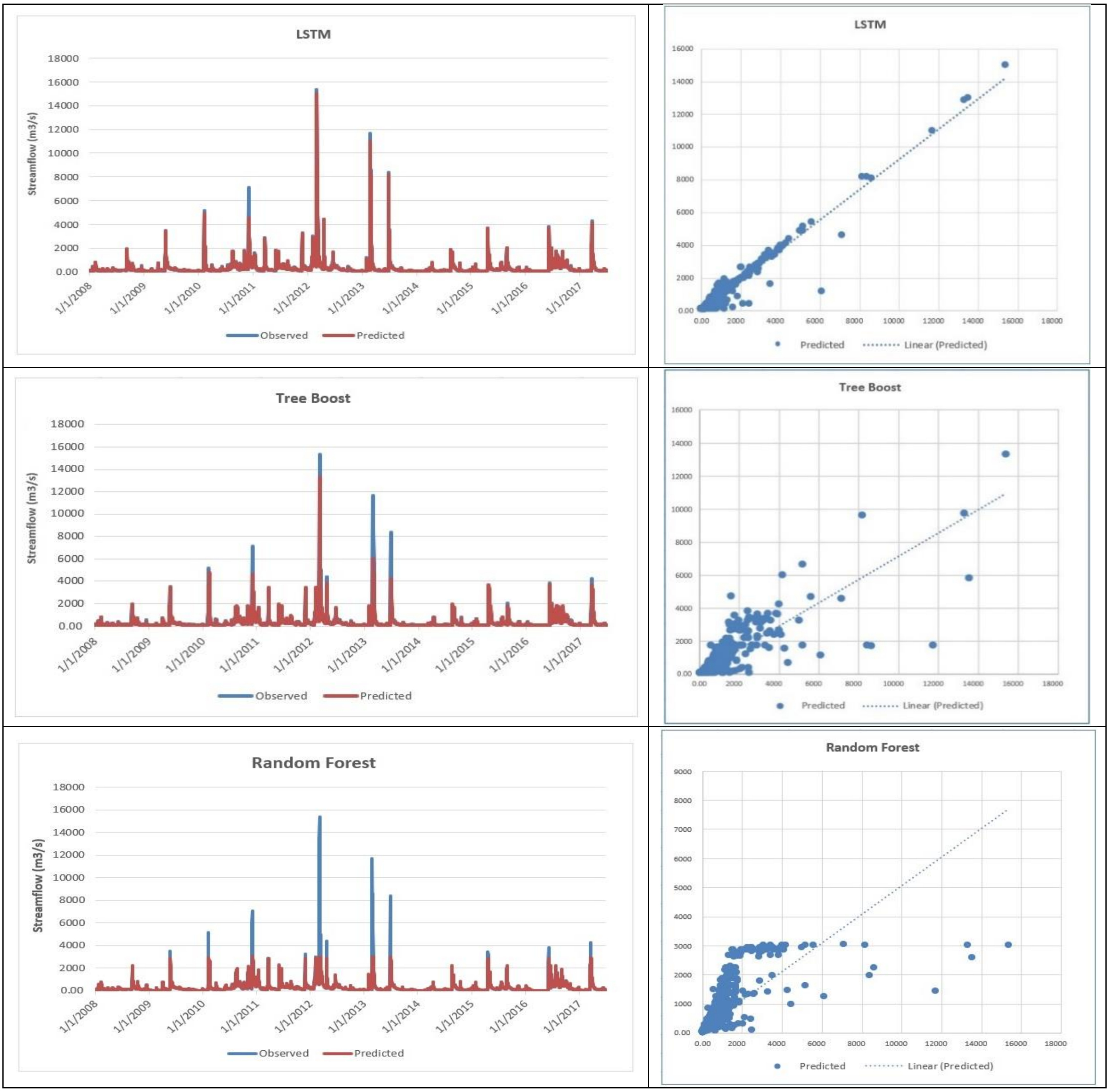

Figure 4. Comparison of observed and predicted daily streamflow, and scatter plot for the LSTM, BT, and RF models

4. CONCLUSION

In this research, three models, LSTM, TB, and RF, were applied to develop models to predict daily streamflow for Kowmung river using historical daily streamflow data from 1/1/2008 to 1/7/2017. For each proposed model, the auto-correlation function (ACF) was used to choose the most accurate predictor variables. The findings showed that the LSTM model outperformed RF and TB in predicting the streamflow with RMSE and NSE equal to 102.411, and 0.911 respectively for the LSTM model. Moreover, TB outperformed RF model. The proposed LSTM model could adopt by hydrologists to solve the problems associated with forecasting daily streamflow with high precision. This study may not be generalized because of the geographical condition and the nature of the data for each location, however, its recommended for future research to apply the LSTM model for predicting hydrological parameters in different locations.

ACKNOWLEDGMENT

The authors thankfully acknowledge Ministry of Higher Education Malaysia for Fundamental Research Grant Scheme (FRGS, Grant No.:FRGS/1/2019/TK01/UNITEN/02/3). The authors also would like to thank WaterNSW for providing the data online.

REFERENCES

[1] Adnan, R.M., Liang, Z., Trajkovic, S., Zounemat-

Kermani, M., Li, B., Kisi, O. (2019). Daily streamflow prediction using optimally pruned extreme learning machine. Journal of Hydrology, 577: 123981. https://doi.org/10.1016/j.jhydrol.2019.123981

[2] Ehteram, M., Ahmed, A.N., Ling, L., Fai, C.M., Latif, S. D., Afan, H.A., Banadkooki, F.B., El-Shafie, A. (2020). Pipeline scour rates prediction-based model utilizing a multilayer perceptron-colliding body algorithm. Water, 12(3): 902. https://doi.org/10.3390/w12030902

[3] Awchi, T.A. (2014). River discharges forecasting in northern Iraq using different ANN techniques. Water Resources Management, 28: 801-814. https://doi.org/10.1007/s11269-014-0516-3

[4] Makwana, J.J.., Tiwari, M.K. (2014). Intermittent streamflow forecasting and extreme event modelling using wavelet based artificial neural networks. Water Resources Management, 28: 4857-4873. https://doi.org/10.1007/s11269-014-0781-1

[5] Wu, C.L., Chau, K.W. (2010). Data-driven models for monthly streamflow time series prediction. Engineering Applications of Artificial Intelligent, 23(8): 1350-1367. https://doi.org/10.1016/j.engappai.2010.04.003

[6] Ahmed, J.A., Sarma, A.K. (2007). Artificial neural network model for synthetic streamflow generation. Water Resources Management, 21: 1015-1029. https://doi.org/10.1007/s11269-006-9070-y

[7] Chang, F.J., Tsai, M.J. (2016). A nonlinear spatio-temporal lumping of radar rainfall for modeling multi-step-ahead inflow forecasts by data-driven techniques. Journal of Hydrology, 535: 256-269. https://doi.org/10.1016/j.jhydrol.2016.01.056

[8] Chokmani, K., Ouarda, T.B.M.J., Hamilton, S., Ghedira, M.H., Gingras, H. (2008). Comparison of ice-affected streamflow estimates computed using artificial neural networks and multiple regression techniques. Journal of Hydrology, 349(3-4): 383-396. https://doi.org/10.1016/j.jhydrol.2007.11.024

[9] Tsai, W.P., Chang, F.J., Chang, L.C., Herricks, E.E. (2015). AI techniques for optimizing multi-objective reservoir operation upon human and riverine ecosystem demands. Journal of Hydrology, 530: 634-644. https://doi.org/10.1016/j.jhydrol.2015.10.024

[10] Yin, X.A., Yang, Z.F., Petts, G.E., Kondolf, G.M. (2014). A reservoir operating method for riverine ecosystem protection, reservoir sedimentation control and water supply. Journal of Hydrology. https://doi.org/10.1016/j.jhydrol.2014.02.037

[11] Gong, Y., Zhang, Y., Lan, S., Wang, H. (2016). A comparative study of artificial neural networks, support vector machines and adaptive neuro fuzzy inference system for forecasting groundwater levels near Lake Okeechobee, Florida. Water Resources Management, 30: 375-391. https://doi.org/10.1007/s11269-015-1167-8

[12] Maiti, S., Tiwari, R.K. (2014). A comparative study of artificial neural networks, Bayesian neural networks and adaptive neuro-fuzzy inference system in groundwater level prediction. Environmental Earth Sciences, 71: 3147-3160. https://doi.org/10.1007/s12665-013-2702-7

[13] Sreekanth, J., Datta, B. (2014). Stochastic and robust multi-objective optimal management of pumping from coastal aquifers under parameter uncertainty. Water Resources Management, 28: 2005-2019. https://doi.org/10.1007/s11269-014-0591-5

[14] Tsai, W.P., Chiang, Y.M., Huang, J.L., Chang, F.J. (2016). Exploring the mechanism of surface and ground water through data-driven techniques with sensitivity analysis for water resources management. Water Resources Management, 30: 4789-4806. https://doi.org/10.1007/s11269-016-1453-0

[15] Shu, C., Ouarda, T.B.M.J. (2007). Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space. Water Resources Research, 43(7). https://doi.org/10.1029/2006WR005142

[16] Yuan, X., Chen, C., Lei, X., Yuan, Y., Muhammad Adnan, R. (2018). Monthly runoff forecasting based on LSTM–ALO model. Stochastic Environment Research and Risk Assessment, 32: 2199-2212. https://doi.org/10.1007/s00477-018-1560-y

[17] Rezaie-Balf, M., Nowbandegani, S.F., Samadi, S.Z., Fallah, H., Alaghmand, S. (2019). An ensemble decomposition-based artificial intelligence approach for daily streamflow prediction. Water, 11(4): 709. https://doi.org/10.3390/w11040709

[18] Parks and Wildlife Division. (2005). Kowmung River Kanangra-Boyd National Park Wild River Assessment. No. June. https://www.environment.nsw.gov.au/research-and-publications/publications-search/kowmung-river-kanangra-boyd-national-park-wild-river-assessment.

[19] Birch, G., Siaka, M., Owens, C. (2001). The source of anthropogenic heavy metals in fluvial sediments of a rural catchment: Coxs River. Australia. Water, Air, & Soil Pollution 126: 13-35. https://doi.org/10.1023/A:1005258123720

[20] Plan, W. Inflow Levels (Rainfall), outflow Levels (Bulk Water Supply to WaterNSW) and Storage for Warragamba Dam Speci Cally, for 1995 to 2017. (2017) https://www.righttoknow.org.au/request/inflow_levels_rainfall_outflow_l.

[21] Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8): 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[22] Al-Juboori, A.M. (2019). Generating monthly stream flow using nearest river data: Assessing different trees models. Water Resources Management, 33(9): 3257-3270. https://doi.org/10.1007/s11269-019-02299-4

[23] Friedman, J.H. (2002). Stochastic gradient boosting. Computational Statistics & Data Analysis, 38(4): 367-378. https://doi.org/10.1016/S0167-9473(01)00065-2

NOMENCLATURE

AI	artificial intelligence
RF	random forest
TB	tree boost
ACF	Auto-correlation function
RMSE	Root mean square error
NSE	Nash and Sutcliffe coefficient
ANN	artificial neural network
ALO	Ant lion optimizer