

PREDICTION OF FLOOD IN KARKHEH BASIN USING DATA-DRIVEN METHODS

S. Kamali ^{1*}, F. Saedi ¹, K. Asghari ¹

¹ Dept. of Civil Engineering, Isfahan University of Technology, Isfahan, Iran - shima.kamali@alumni.iut.ac.ir, f.saedi1992@gmail.com and kasghari@iut.ac.ir

Commission IV, WG IV/3

KEY WORDS: Flood, Data-Driven Methods, Annual Maximum Streamflow, Precipitation, Modelling, Karkheh Basin

ABSTRACT:

Flood causes several threats with outcomes which include peril to human and animal life, damage to property, and adversity to agricultural fields. Hence, flood prediction is highly significant for the mitigating municipal and environmental damage. The aim of this study was assessing the performance of different machine learning methods in predicting flood in Karkheh basin. To aim this, we used Support Vector Machine (SVM), Least Square Support Vector Machine (LSSVM), Feed Forward Back Propagation Neural Network (FFBPNN), and Radial Basis Function Neural Network (RBFNN) to simulate monthly streamflow in the study area. Furthermore, the performance of models was compared in predicting flood. All four models indicated good performance in simulating streamflow. However, LSSVM model had the highest accuracy compared with other models with R^2 and $RMSE$ of 85.89% and 30.02 m³/s during testing periods, respectively. Similarly, LSSVM model performed better in predicting annual maximum streamflow in comparison with other machine learning models.

1. INTRODUCTION

Floods cause serious damage to various infrastructure and socioeconomic systems and make various economic losses (Sieg et al., 2019). The complex behaviour of river flow makes the flood a complex phenomenon. Various factors like soil, land cover, climate, and snowfall can affect the river flow (Faiz et al., 2017). Therefore, non-linear and dynamic nature of flood leads to difficulty in prediction of this phenomenon. On the other hand, it is crucial to predict floods accurately to prepare for the emergency response (Pitt et al., 2008). Physically based and data driven models are the most common types of models for flood prediction. Accurate prediction significantly contributes to water recourse management, policy suggestions and reliable analysis (Xie et al., 2017). A physical model uses mathematical equations to simulate the hydrological components like rainfall/runoff, flood prediction and other hydrological components (Costabile et al., 2015; Jalali et al., 2021). The assumptions involved in the physical models and the complex nature of the flood prediction can sometimes result in inaccurate predictions. (Honert et al., 2011). Furthermore, the development of physically based models often requires in-depth knowledge regarding hydrological parameters, reported to be highly challenging. On the other hand, Machine learning (ML) method is a field of artificial intelligence (AI) for inducing regularities and patterns.

ML models have lower computational cost than physical models. Also, the process of training, validation and testing is faster in ML models (Greydanus et al., 2019). The continuous advancement of ML methods over the last two decades has proved their suitability for flood forecasting (Mosavi et al., 2018). ML methods can numerically formulate the flood nonlinearity, based on historical data without requiring knowledge about the underlying physical processes. They are promising tools as they are quicker to develop with minimal inputs. In addition, recent studies indicate that ML models have been used for flood prediction with greater accuracy than traditional statistical models like autoregressive moving average

(ARMA), multiple linear regression (MLR), and autoregressive integrated moving average (ARIMA) (Xu et al., 2002; Latt et al., 2014).

The study of stream flow prediction is highly significant for the purpose of municipal and environmental damage mitigation. During recent years ML models like ANN and SVM have been used for predicting streamflow and flood analysing. Elsafi (2014) forecasted the River Nile flow at Dongola Station in Sudan using an Artificial Neural Network (ANN) as a modelling tool and validated the accuracy of the model against actual flow. The analysis indicated that the ANN provides a reliable means of detecting the flood hazard in the river Nile. Kourgialas et al. (2015) created a modelling management tool for the simulating extreme flow events under current and future climatic conditions. The prediction-forecasting artificial neural network (ANN) model was applied to accurately and efficiently simulate river flow on an hourly basis. They concluded that ANN is capable of modelling persistent events. Dtissibe et al. (2020) conducted a study for forecasting flood based on an artificial neural network scheme. They used discharge as input-output variables. The designed model has been tested upon intensive experiments and the results showed the effectiveness of their method with a good forecasting capacity. Jajarmizadeh et al. (2015) compared the efficiency of the SVM and SWAT models for predicting the monthly streamflow of the Roodan basin located in Iran as an arid to semi-arid region. This study showed that both SWAT and SVM models possessed a satisfying capability in predicting the monthly streamflow. Yan et al. (2018) developed an urban flood forecast framework combining a numerical model based on MIKE FLOOD with SVM models. The numerical model was the data source for the SVM model, and the SVM model provided fast forecast. Based on the result, a combination of numerical model and SVM model will achieve high solution accuracy and save significant computational time. Adnan et al. (2018) assessed the potential of LSSVM for forecasting of stream flow in poorly gauged catchment. They concluded that the LSSVM models can be

* Corresponding author

used successfully to forecast the stream flows in comparison with fuzzy genetic algorithm (FGA) and M5 model tree (M5T). Previous studies indicated that ML models provide acceptable accuracy for streamflow prediction. This research was planned to investigate the performance of machine learning models in predicting flood in Karkheh basin. We used SVM, LSSVM, FFBPNN, and RBFNN models to simulate monthly streamflow and compared their performance in predicting annual maximum streamflow in the study area.

2. STUDY AREA AND DATA

The Karkheh basin is one of the most important basins in Iran, located in the central and southern regions of the Zagros mountain range and covers an area of 50,000 km² (Figure 1). The Karkheh River is the third largest river in Iran with 900 km long. The Jelogir station is located at the upper reaches of the Karkheh reservoir and has the greatest impact on the reservoir inflow.

The data collected from the Iran Water Resources Management Company (IWRMC) include precipitation at 11 stations and streamflow at 7 stations for the period of 1966-2017.

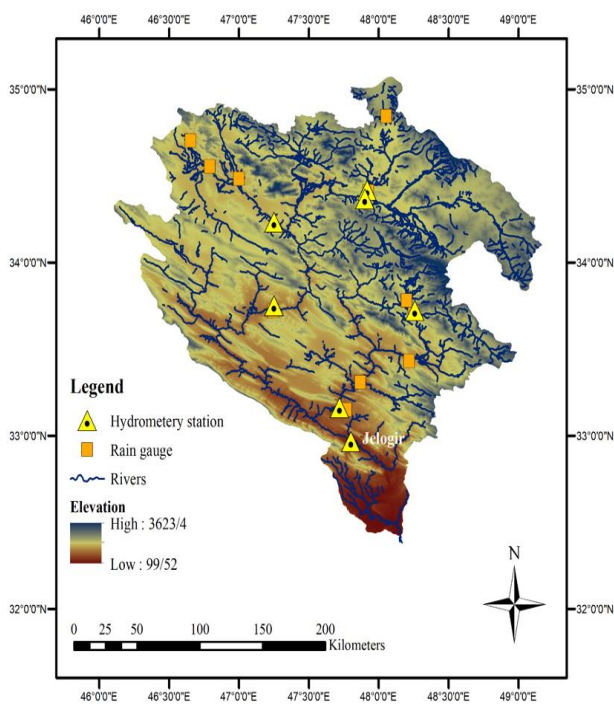


Figure 1. Location of Karkheh basin

3. METHODOLOGY

The aim of this study was investigating and comparing the performance of different machine learning methods in predicting flood in Karkheh basin. We used four data-driven methods including SVM, LSSVM, FFBPNN, and RBFNN models to simulate monthly streamflow and predict flood in the study area. We used R^2 and $RMSE$ to compare the models' performance in simulating flood events. Among all four models, LSSVM performed better in simulating monthly streamflow and predicting annual maximum streamflow.

3.1 Support Vector Machine

In this study, we used ε -SVM as a regression model. Assume N samples of $\{x_i, y_i\}_{i=1}^N$ is the training dataset where x_i is the

input vector and y_i is the corresponding output value. The support vector regression (SVR) function can be expressed by $f(x)$ ($f(x) = w^T \cdot \varphi(x) + b + \varepsilon$), where w (weight vector) and b (constant) are model parameters, φ is a non-linear transfer function and ε represents noise. The w and b are derived by the following optimization procedure:

$$\begin{aligned} & \text{minimize} && \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m (\xi_i + \xi_i^*) \\ & \text{subject to} && \begin{cases} y_i - w^T \cdot \varphi(x_i) - b \leq \varepsilon + \xi_i \\ w^T \cdot \varphi(x_i) + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \end{aligned} \quad (1)$$

where C = positive constant that determine the penalization degree when error occurs
 ξ_i, ξ_i^* = slack variables defining the upper and lower training error over the error tolerance ε

For more details of this method, see Vapnik (1995).

3.2 Least Square Support Vector Machine (LSSVM)

The LSSVM model is a modified version of standard Support Vector Machine (SVM), which uses the least squares loss functions instead of solving quadratic programming problem. For the training sample $\{x_i, y_i\}_{i=1}^N$ with x_i as the input vector and the corresponding output value, y_i , the LSSVM non-linear function is defined by $f(x)$ ($f(x) = w^T \cdot \varphi(x) + b$) where w is a weight vector with the same dimension as the feature space, b represents bias and φ is mapping function that maps the input variables into a higher dimensional feature space. In a regression problem, w and b can be derived from following minimization:

$$\begin{aligned} & \text{minimize} && \frac{1}{2} w^T w + \frac{1}{2} \gamma \sum_{i=1}^N e_i^2 \\ & \text{subject to} && y_i = w^T \cdot \varphi(x_i) + b + e_i \end{aligned} \quad (2)$$

where γ = the regularization parameter
 e_i = the slack variable for x_i

For more details of this method, see Suykens et al. (2002).

3.3 Feed Forward Back Propagation Neural Network (FFBPNN)

Back Propagation Neural Network (BPNN) is the most characteristic learning model for Artificial Neural Network (Wilde 1997). The BPNN process involves error at output layer that back-propagates to input layer via hidden layer in network for obtaining ultimate desired output. Feed Forward BPNN (FFBPNN) is frequently used in hydrological modelling and utilizes BP as training algorithm. The FFBPNN network consists of single input layer, single hidden layer comprising of n neuron numbers, and single output layer. FFBPNN can be specified as follow:

$$y = f\left(\sum_{i=1}^n w_i x_i + b\right) \quad (3)$$

where y = produced output
 f = transfer function

w_i = weight vector
 x_i = input vector
 b = bias

In the present study, a three-layer FFBPNN model skilled using tansigmoid function which helps to define and select number of neurons of hidden layer while linear function is utilized to calculate number of neurons of output layer.

3.4 Radial Basis Function Neural Network (RBFNN)

RBF is contemplated as an authoritative method to interpolate various functions in a multi-dimension space (Broomhead and Lowe 1998). RBFs have three layers; input, hidden with RBF nonlinearity and a linear output layer. In view of complexity in nature of flood process that is generally nonlinear, most appropriate ANNs to model the process must have capability for approximating any continuous function.

RBFs ϕ_1 , ϕ_2 , and ϕ_N , are called hidden functions whereas $\{\phi_i\}_{i=1}^N$ is known as hidden space. In this study, Gaussian RBFs is used which can be represented as follow:

$$\phi(x, \mu) = e^{-\frac{\|x-\mu\|^2}{2d^2}} \quad (4)$$

where μ = center of Gaussian function (mean value of x)
 d = distance (radius) from center of $\phi(x, \mu)$, giving an extent of spreading of Gaussian curvature

3.5 Performance Criteria

To evaluate models' performance, the following performance criteria have been used in this study.

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

$$R^2 = 1 - \frac{\sum_{i=1}^N (Est_i - Obs_i)^2}{\sum_{i=1}^N Est_i^2 - \frac{\sum_{i=1}^N Obs_i^2}{N}} \quad (6)$$

where N = the number of data points
 Obs_i = observed value at time i
 Est_i = estimated values at time i

4. RESULTS AND DISCUSSION

In this study, we used four machine learning methods including SVM, LSSVM, FFBPNN, and RBFNN to simulate streamflow, and compared the performance of these methods in predicting flood in Jelogir station. We considered monthly precipitation and streamflow data at 11 and 7 stations, respectively, with two-month delays as the model inputs, and used the Gamma Test (Agalbjorn et al., 1997) method to select the best input combinations.

4.1 Assessment of Annual Precipitation and Streamflow

The total annual precipitation ranges from 137 mm to 718 mm Figure (2) and the annual average streamflow range from 18 m^3/s to 301 m^3/s over the last 52 years (Figure 3). In addition, both total annual precipitation and streamflow follows reducing trends.

The annual average of streamflow over 52 years of study is 138.25 m^3/s and this number is 430.54 mm for precipitation. On the other hand, Figures 2 and 3 show that the rainfall and streamflow peaks occurred in various years. The amount of peak rainfall and discharge are significantly more than the long-term average of these two variables. Hence, the flood events can occur with considerable possibility, as a result, an accurate model for predicting flood events is necessary in this study area.

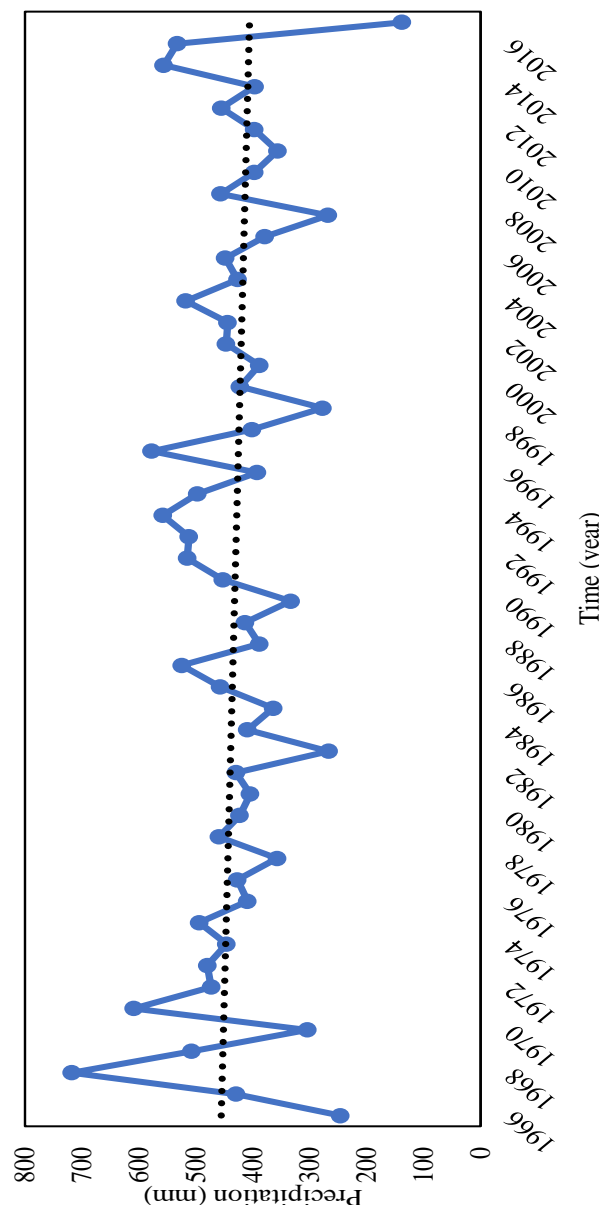


Figure 2. Annual precipitation over Karkheh basin

4.2 Models' Performance in Simulating Monthly Streamflow

In this study, SVM, LSSVM, FFBPNN, and RBFNN were used to simulate monthly streamflow at Jelogir station. We considered 80% of inputs data as training period and 20% of inputs data as testing period. To evaluate models' performance in simulating monthly streamflow, we calculated statistical indices (R^2 and $RMSE$) for both training and testing period (Table 1).

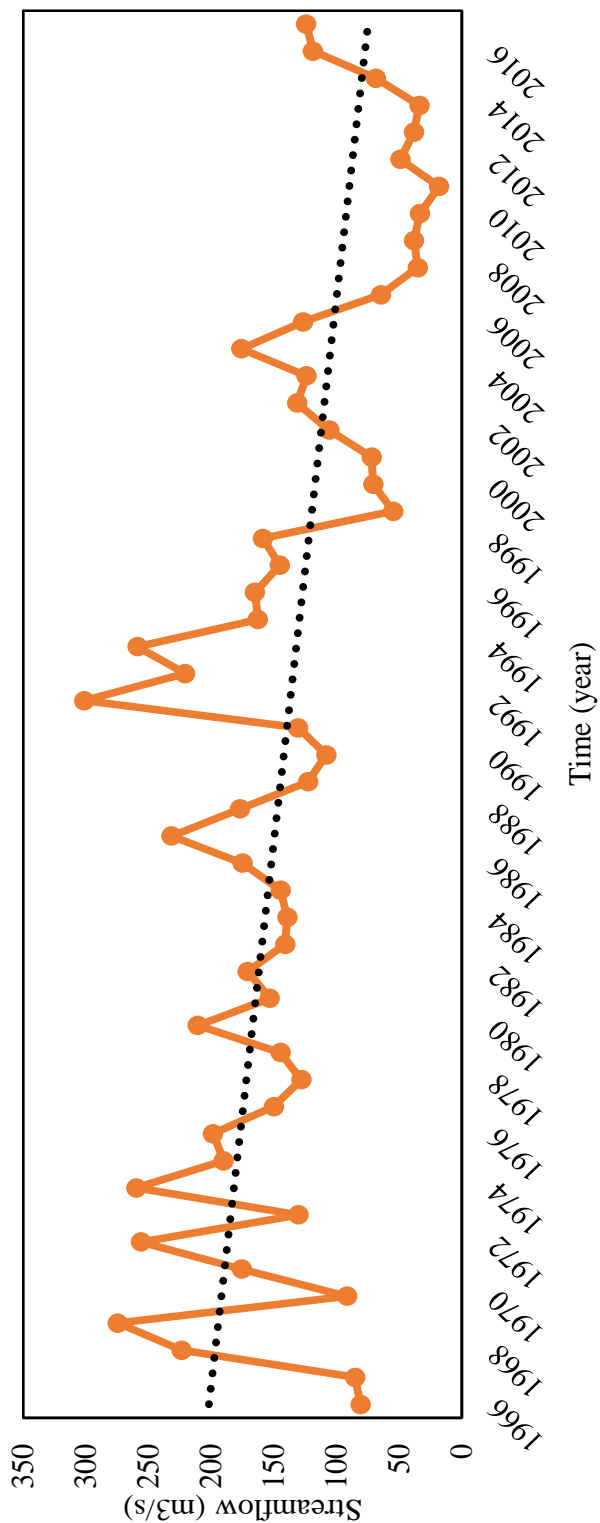


Figure 3. Annual streamflow at Jelogir station

The R^2 values for all models were more than 85% during training and testing periods indicating high correlation between model simulations and observation data. The small $RMSE$ values represent the high accuracy of models in estimating monthly streamflow. However, the highest R^2 values for training and testing periods are 97.76% and 85.89%, respectively. The lowest $RMSE$ values during training and testing periods are 26.03 m^3/s and 30.02 m^3/s , respectively. Hence, LSSVM model

indicates better performance in simulating streamflow during training and testing period in comparison with other three models. For both training and testing periods, FFBPNN shows lesser (or worse) performance among all the machine learning approaches. In general, LSSVM, RBFNN, and SVM accuracy for training and testing periods is similar with slight differences.

| Model | R^2 (%) | | $RMSE$ (m^3/s) | |
|--------|-----------|-------|--------------------|-------|
| | Train | Test | Train | Test |
| SVM | 97.25 | 85.06 | 28.46 | 36.49 |
| LSSVM | 97.76 | 85.89 | 26.03 | 30.02 |
| FFBPNN | 95.97 | 83.83 | 34.54 | 42.57 |
| RBFNN | 97.47 | 85.73 | 27.34 | 30.34 |

Table 1. Models performance in simulating monthly streamflow at Jelogir station during training and testing periods

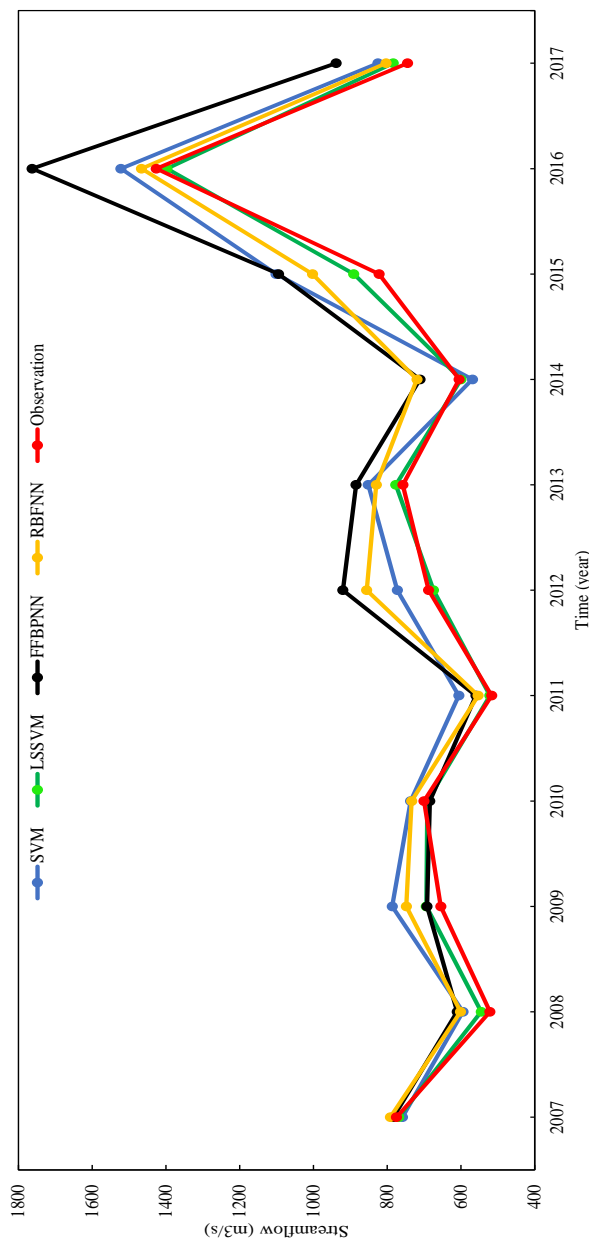


Figure 4. The performance of SVM, LSSVM, FFBPNN, and RBFNN models in predicting annual maximum streamflow at Jelogir station

4.3 Comparison of Models Performance in Predicting Annual Maximum Streamflow

To assess the ability of models in predicting flood, predicted annual maximum streamflow was compared with observation data during testing period (Figure 4). All four models show good performance in predicting the trend of annual maximum streamflow. However, LSSVM model represents better performance in predicting annual maximum streamflow compared with other machine learning methods. According to the model performance results in the previous section, LSSVM model had the highest accuracy in predicting monthly streamflow. On the other hand, similar to the previous section, FFBPNN model has the lowest performance in predicting annual maximum streamflow.

The findings of this study indicate that all mentioned machine learning methods have significant potential in predicting flood events. However, LSSVM model has the best performance in predicting flood in this study area. Furthermore, the presented models could help decision-makers to predict flash floods in different regions.

CONCLUSION

In this study, the performance of different machine learning models in predicting flood was investigated in Karkheh basin. Four machine learning methods including SVM, LSSVM, FFBPNN, and RBFNN were used to simulate streamflow in the study area. The results demonstrated that LSSVM model give more precise outcomes in simulating streamflow compared to other three models. Assessment results also revealed that although all four models showed good performance in predicting the trend of annual maximum streamflow, LSSVM model indicated the highest accuracy in predicting flood. Generally, the models used in this study could be utilized to simulate streamflow and predict flood. The study outcomes also would help decision-makers to predict flood events and mitigate environmental damages in future.

REFERENCES

Adnan, R.M., Yuan, X., Kisi, O., Adnan, M. and Mehmood, A., 2018. Stream flow forecasting of poorly gauged mountainous watershed by least square support vector machine, fuzzy genetic algorithm and M5 model tree using climatic data from nearby station. *Water Resources Management*, 32(14), 4469-4486. doi.org/10.1007/s11269-018-2033-2

Agalbjorn, S., Koncar, N. and Jones, A.J., 1997. A note on the gamma test. *Neural Comput Appl*, 5(3), 131-133.

Broomhead, D.S. and Lowe, D., 1988. Radial basis functions, multivariable functional interpolation and adaptive networks (No. RSRE-MEMO-4148). Royal Signals and Radar Establishment Malvern, United Kingdom.

Costabile, P. and Macchione, F., 2015. Enhancing river model set-up for 2-D dynamic flood modelling. *Environmental Modelling & Software*, 67, 89-107. doi.org/10.1016/j.envsoft.2015.01.009

Dtissibe, F.Y., Ari, A.A.A., Titouna, C., Thiare, O. and Gueroui, A.M., 2020. Flood forecasting based on an artificial neural network scheme. *Natural Hazards*, 104(2), pp.1211-1237. doi.org/10.1007/s11069-020-04211-5

Elsafi, S.H., 2014. Artificial neural networks (ANNs) for flood forecasting at Dongola Station in the River Nile, Sudan. *Alexandria Engineering Journal*, 53(3), 655-662. doi.org/10.1016/j.aej.2014.06.010

Faiz, M.A., Liu, D., Fu, Q., Qamar, M.U., Dong, S., Khan, M.I. and Li, T., 2018. Complexity and trends analysis of hydrometeorological time series for a river streamflow: A case study of Songhua River Basin, China. *River Research and Applications*, 34(2), 101-111. doi.org/10.1002/rra.3236

Greydanus, S., Dzamba, M., & Yosinski, J. (2019). Hamiltonian neural networks. Paper presented at advances in neural information processing systems 32: Annual conference on neural information processing system 2019. Vancouver, Canada, 15353–15363.

Jajarmizadeh, M., Kakaei Lafdani, E., Harun, S. and Ahmadi, A., 2015. Application of SVM and SWAT models for monthly streamflow prediction, a case study in South of Iran. *KSCE Journal of Civil Engineering*, 19(1), 345-357. doi.org/10.1007/s12205-014-0060-y

Jalali, J., Ahmadi, A., Abbaspour, K., 2021. Runoff responses to human activities and climate change in an arid watershed of central Iran. *Hydrological Sciences Journal*, 66(16), 2280-2297. doi.org/10.1080/02626667.2021.1985724

Kourgialas, N.N., Dokou, Z. and Karatzas, G.P., 2015. Statistical analysis and ANN modeling for predicting hydrological extremes under climate change scenarios: The example of a small Mediterranean agro-watershed. *Journal of Environmental Management*, 154, 86-101. doi.org/10.1016/j.jenvman.2015.02.034

Latt, Z.Z. and Wittenberg, H., 2014. Improving flood forecasting in a developing country: a comparative study of stepwise multiple linear regression and artificial neural network. *Water Resources Management*, 28(8), 2109-2128. doi.org/10.1007/s11269-014-0600-8

Mosavi, A., Ozturk, P. and Chau, K.W., 2018. Flood prediction using machine learning models: Literature review. *Water*, 10(11), 1536. doi.org/10.3390/w10111536

Pitt, M., 2008. Learning Lessons From the 2007 Floods. An Independent Review by Sir Michael Pitt. Cabinet Office, London.

Sieg, T., Schinko, T., Vogel, K., Mechler, R., Merz, B. and Kreibich, H., 2019. Integrated assessment of short-term direct and indirect economic flood impacts including uncertainty quantification. *PloS one*, 14(4). doi.org/10.1371/journal.pone.0212932

Suykens, J.A., De Brabanter, J., Lukas, L. and Vandewalle, J., 2002. Weighted least squares support vector machines: robustness and sparse approximation. *Neurocomputing*, 48(1-4), 85-105.

Van den Honert, R.C. and McAneney, J., 2011. The 2011 Brisbane floods: causes, impacts and implications. *Water*, 3(4), 1149-1173. doi.org/10.3390/w3041149

Vapnik, V.N., 1995. The nature of statistical learning. Theory.

Wilde PD, 1997. *Neural network models: theory and projects*, 2nd edn. Springer, London.

Xie, K., Ozbay, K., Zhu, Y. and Yang, H., 2017. Evacuation zone modeling under climate change: A data-driven method. doi.org/10.1061/(ASCE)IS.1943-555X.0000369

Xu, Z.X. and Li, J.Y., 2002. Short-term inflow forecasting using an artificial neural network model. *Hydrological Processes*, 16(12), 2423-2439. doi.org/10.1002/hyp.1013

Yan, J., Jin, J., Chen, F., Yu, G., Yin, H. and Wang, W., 2018. Urban flash flood forecast using support vector machine and numerical simulation. *Journal of Hydroinformatics*, 20(1), 221-231. doi.org/10.2166/hydro.2017.175