Sains Malaysiana 51(8)(2022): 2655-2668
http://doi.org/10.17576/jsm-2022-5108-24

Hydroclimatic Data Prediction using a New Ensemble Group Method of Data Handling Coupled with Artificial Bee Colony Algorithm

(Ramalan Data Hidroklimatik menggunakan Kaedah Pengendalian Data Kumpulan Ensembel Baharu Digandingkan dengan Algoritma Koloni Lebah Buatan)

Basri Badyalina[1,*], Nurkhairany Amyra Mokhtar[1], Nur Amalina Mat Jan[2], Muhammad Fadhil Marsani[3], Mohamad Faizal Ramli[4], Muhammad Majid[4] & Fatin Farazh Ya'acob[4]

[1]*Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Cawangan Johor, Kampus Segamat, 85000 Segamat, Johor Darul Takzim, Malaysia*
[2]*Department of Physical and Mathematical Science, Faculty of Science, Universiti Tunku Abdul Rahman, Kampar Campus, Jalan Universiti, Bandar Barat, 31900 Kampar, Perak Darul Ridzuan, Malaysia*
[3]*School of Mathematical Sciences, Universiti Sains Malaysia, 11800 Minden, Penang, Malaysia*
[4]*Universiti Teknologi MARA, Cawangan Johor, Kampus Segamat, 8500 Segamat, Johor Darul Takzim, Malaysia*

Received: 22 August 2021/Accepted: 22 February 2022

ABSTRACT

Linear regression is widely used in flood quantile study that consists of meteorological and physiographical variables. However, linear regression does not capture the complex nonlinear relationship between predictor and target variables. It is rare to find a hydrological application using the group method of data handling (GMDH) model, artificial bee colony (ABC) algorithm, and ensemble technique, precisely predicting ungauged sites. GMDH model is known to be an effective model in complying with a nonlinear relationship. Therefore, in this paper, we enhance the GMDH model by implementing the ABC algorithm to optimize the parameter of partial description GMDH model with some transfer functions, namely polynomial, radial basis, sigmoid and hyperbolic tangent function. Then, ensemble averaging combines the output from those various transfer functions and becomes the new ensemble GMDH model coupled with the ABC algorithm (EGMDH-ABC) model. The results show that this method significantly improves the prediction performance of the GMDH model. The EGMDH-ABC model satisfies the nonlinearity in data to produce a better estimation. Also, it provides more robust, accurate, and efficient results.

Keywords: ABC algorithm; GEV distribution; GMDH modele; Peninsular Malaysia; ungauged site

ABSTRAK

Regresi linear digunakan secara meluas dalam kajian kuantiti banjir yang terdiri daripada pemboleh ubah meteorologi dan fisiografi. Walau bagaimanapun, regresi linear tidak mengenal pasti hubungan tidak linear yang kompleks antara pemboleh ubah peramal dan sasaran. Sukar untuk menemui aplikasi hidrologi yang menggunakan kaedah kumpulan model pengendalian data (GMDH), algoritma koloni lebah tiruan (ABC) dan teknik penggabungan, khususnya dalam meramalkan kuantil banjir di kawasan tiada data. Model GMDH dikenali sebagai model yang berkesan dalam mengenal pasti hubungan tidak linear. Oleh itu, dalam kajian ini, kami menambah baik model GMDH dengan menerapkan algoritma ABC untuk mengoptimumkan parameter penerangan separa model GMDH dengan beberapa fungsi pemindahan iaitu fungsi polinomial, asas radial, sigmoid dan tangen hiperbolik. Kemudian, penggabungan secara purata digunakan untuk menggabungkan hasil daripada pelbagai fungsi pemindahan tersebut dan membangunkan model baru iaitu EGMDH-ABC. Hasil kajian menunjukkan bahawa kaedah ini meningkatkan prestasi ramalan model GMDH dengan ketara. Model EGMDH-ABC berjaya mengenal pasti ketidaklinearan di dalam data untuk menghasilkan anggaran yang lebih baik. Di samping itu, hasil keputusan yang lebih mantap, tepat dan cekap dapat dihasilkan.

Kata kunci: Algoritma ABC; lembangan tiada data; model GMDH; Semenanjung Malaysia; taburan GEV

Introduction

Data-driven models, among the methodologies available, have caught the scientific community's interest in recent decades due to their adaptability and forecasting accuracy. In water resource management, the hydrological system consists of a lot of intrinsic uncertainty and complexity. Hydrological modelling has become an alternative solution for crucial decision-making tools because hydrological modelling can be used to forecast, manage water resources and obtain a better understanding. Machine learning models such as Group Method of Data Handling (GMDH), a popular data-driven model in recent years, can be implemented for flood quantile prediction at the ungauged site. These methods have advantages such as low cost, high processing speed, and appropriate accuracy. The implementation of data-driven models is gaining popularity. It needs less development and has been shown to provide precise prediction with less information of the behaviors or the process of the hydrological problems (Nariman et al. 2017; Yang et al. 2020).

A flood is a natural disaster that frequently occurs in Peninsular Malaysia. Floods have a substantial negative impact on the environment, national economy, and country infrastructure. Thus, it is crucial to adequately predict the flood on the target site so that the measures to a sustainable implementation plan of water management, flood facilities, and assessing the river activity for the operational decision can be taken. In practice, the length of data at target is an important aspect to produce a satisfactory outcome (Hosking & Wallis 1997). However, there are some of the target sites in Malaysia; the hydrological information is not available or known as the 'ungauged sites' (Mamun et al. 2012). The most common approach to tackling the ungauged site prediction problem is the regionalization method, which transfers the information from the gauged site to the ungauged site. The regionalization method consists of fitting a probability distribution to a flow series and then relating the data-driven models to physical site descriptors (Badyalina & Shabri 2015; Desai & Ouarda 2021). Thus, generalized extreme value (GEV) distribution will be fitted to the target site flow series to obtain the observed flood quantile in this study. The GEV distribution has found numerous hydrological applications for extreme events. Many researchers applied it for flood frequency analysis for streamflow (Guru & Jha 2014). Extreme values are frequently expressed as the maximum value of a particular characteristic for a specified period, such as a year. The GEV distribution can generally describe these maximum values (Badyalina et al. 2021a; Cannon 2010; Mat Jan et al. 2018, 2016a; Wan Zawiah et al. 2009). Besides, the extreme events are more suitable modelled with heavy tails, characterized in the GEV distribution (Otiniano et al. 2019). The tail behavior is strongly significant, as it corresponds to quite different characteristics of extreme value behavior (De Paola et al. 2018; Mat Jan et al. 2016b).

The most common data-its simplicity and low computational model (Desai & Oudriven model used for the regionalization method is multiple linear regression (MLR) due to arda 2021). The drawback of using a linear model is that the model cannot capture the complexity of the relationship between predictor variables and flow characteristics. Sivakumar and Singh (2012) demonstrated that the relationship between these factors is primarily nonlinear. Therefore, the nonlinear data-driven model is proposed to model the nonlinear relationship between predictor variables characteristics and flow characteristics. In this study, the GMDH model has been selected for flood quantile prediction at the ungauged site. Numerous researches in hydrology have been performed utilizing the GMDH model (Adnan et al. 2021; Ahmadi et al. 2019; Maofa et al. 2021). An artificial neural network (ANN) is the common model of nonlinear methods successfully used in flood quantile estimation (Badyalina et al. 2021b; Kordrostami et al. 2020; Shu & Burn 2004; Shu & Ouarda 2007). The capability of the ANN model in the estimating flood quantile for ungauged basins undeniable when studies from Meresa (2019), Jolankai and Koncsos (2018) and Aziz et al. (2017) have proved that the ANN model delivers more consistent accuracy in comparison to the linear regression (LR) model. This is the major feature of the ANN model in order to deal with nonlinear data (Wu et al. 2016). Meanwhile, Khan et al. (2021) applied ANN and MLR to develop a dependent and independent variables model. The models will be used to predict the quantiles of ungauged sites in Pakistan. The research outcome shows that the estimated quantiles using ANN give an accurate and close result to the maximum values of peak flows for all sites. Other than the hydrology area, the GMDH model has been successfully applied in other areas such as mechanical engineering, energy performance, and evapotranspiration rate (Ahmadi et al. 2015; Ashrafzadeh et al. 2020; Kardani et al. 2021). Ivakhnenko (1971) primarily devised the GMDH model for modelling and detection of complex systems. The GMDH model uses quadratic equations to describe the complex relationship between input and output variables.

The number of neurons keeps increasing when the layer is increased. Amiri and Soleimani (2021) stated that each layer of the GMDH model uses polynomial functions to transfer a different subset of potential input combinations of existing features to the desired outcome in each node. Hosseini et al. (2021) stated that the advantage of the GMDH model is its immunity based on the experiment of the power fluctuations detection. The experimental result shows that the GMDH model reacts efficiently and accurately to various power fluctuations and concurrent failures. GMDH has the advantage of autonomously selecting the right model system and several nodes without having to over-train the data (Solanki et al. 2021).

Data-driven models such as GMDH are frequently locked in the local minimum solution and thus unable to locate the global minimum solution (Elbaz et al. 2021). In order to overcome the drawback, utilizing the optimization algorithms is needed. The Artificial Bee Colony (ABC) algorithm is an optimization algorithm proposed by Karaboga and Basturk (2007) that can improve the generalization capability of data-driven models (Le et al. 2019; Lu et al. 2019; Tan et al. 2021). A considerable amount of literature has been published on Artificial Bee Colony (ABC) algorithm. These studies have been proposed to explain how the bumblebee behavior find the near-optimal solutions to the difficult optimization problems. Prior to the work of Tereshko and Lee (2002), there are three main components for the intelligence of honeybee swarms, namely food sources, employed foragers, and unemployed foragers. It has been demonstrated that the exchange of information among bees via waggle dance is the most important occurrence in the formation of collective knowledge. A major advantage of ABC algorithms compared to other established algorithms (for example, evolution strategies, genetic algorithm, differential evolution algorithm, particle swarm optimization) is the term fitness (Karaboga & Akay 2009). In most of the algorithms, the term fitness directly corresponding to the objective function value. In contrast, the ABC optimization employed the fitness to be related to the objective function. From the results obtained in a previous study, it was demonstrated that the performance of the ABC algorithm is better than or similar to that of other algorithms such as genetic algorithm and particle swarm optimization (Karaboga & Akay 2009). However, the ABC uses fewer control parameters.

Earlier research has shown that employing an ensemble multi-model is preferable depending on the selection of a single model (Xiao et al. 2018). The ensemble may highlight the strengths of individual models, which may individually neglect or present system processes in a biased manner. Four ensemble averaging technics from six hydrological models were considered by Broderick et al. (2016) to investigate the performance of considered models and studied methods for improving model applicability in climate impact studies for 37 Irish catchments streamflow, rainfall, and potential evapotranspiration (PET) data. From this research, they found that the ensemble average outperformed most individual ensemble models. Tegegne et al. (2019) improved the reliability ensemble average method to represent spatiotemporal variations in climate model skills at many locations and time steps during assigned weights in climate simulators for climate change impact assessments. The intended reliability ensemble average version provided better weight assignment methods (for each climate simulator) and can support numerous weather stations. The main contribution of this paper is to improve the prediction performance of the GMDH model using the ABC algorithm and ensemble technique. The prediction of flood quantile is used as a case study for the application of the proposed model. Other than that, four different transfer functions will be implemented in the proposed model rather than using a single transfer function. The proposed model is an ensemble group method of handling data with the ABC algorithm model (EGMDH-ABC).

CASE STUDY

The study area selected for this study is located in Peninsular Malaysia. There were 60 hydrometric stations chosen in this study located across Peninsular Malaysia. The peak flow data for each 60-river site were obtained from the Department of Irrigation and Drainage, Malaysia. Prior to Shu and Ouarda (2008) work, the minimum historical data required to produce a meaningful prediction of at-site flood quantile is 15 years. The flood quantile for each site was estimated using three parameters generalized extreme value (GEV) distribution. Based on previous research, flood quantile with 10 years (Q_{10}) and 100 years (Q_{100}) return period to cover the high and low of the distribution. Therefore, both of the flood quantiles will be selected for this study. Physical site characteristics or descriptors comprise catchment area (AR), river slope (RS), longest drainage path (PTH) and site elevation (VT). The meteorological site descriptor used in this study was annual precipitation (AP). The statistics overview of flood quantile, physical site descriptors and meteorological site descriptors are presented in Table 1.

TABLE 1. Statistics overview of flood quantile, physical site descriptors and meteorological site descriptors

Variables	Min	Max	Mean	SD
AE	40 km^2	15600 km^2	1519.18 km^2	2902.69 km^2
RS	0.01%	1.65%	0.38%	0.44%
PTH	4350 m	240000 m	33488.33 m	47082.40 m
VT	5 m	1450 m	99.970 m	264.42 m
AP	723.00 mm	4678.70 mm	2172.93 mm	698.05 mm
q10	18.66 m^3/s	4872.68 m^3/s	636.52 m^3/s	1146.99 m^3/s
q100	35.83 m^3/s	7628.66 m^3/s	906.92 m^3/s	1632.29 m^3/s

HYDROLOGICAL MODELS

MULTIPLE LINEAR REGRESSION

Multiple Linear Regression (MLR) is the most common method to transfer the information from gauged site to an ungauged site. The established functional relationship between site characteristics and flood quantile used in various world sites is shown in Equation (1).

$$Q_i = \varphi_0 x_1^{\varphi_1} x_2^{\varphi_2} \ldots x_5^{\varphi_5} \varepsilon \quad (1)$$

where x is the site descriptors; Q_i is the flood quantile of T-return period; φ is the model parameter; and ε is the multiplicative error term. The power form Equation (1) can be linearized by applying the logarithmic transformation. The linearized Equation (1) is shown on Equation (2). The parameter of Equation (2) can be estimated using the least square method.

$$\log(Q_i) = \phi_0 + \phi_0 x_1 + \phi_2 x_2 + \ldots + \phi_5 x_5 + \log(\varepsilon) \quad (2)$$

ARTIFICIAL NEURAL NETWORK

The most common model applied in flood quantile prediction at the ungauged site is the artificial neural network (ANN) model (Alobaidi et al. 2021; Campos et al. 2021; Desai & Ouarda 2021). Kordrostami et al. (2020) used ANN for streamflow data in New South Wales (NSW) in Australia. Khan et al. (2021) proved that ANN is more accurate than multiple linear regression in estimating the flood quantiles in Khyber Pakhtunkhwa, Pakistan. Fillipova et al. (2020) used ANN in estimating flood frequency quantiles for the contiguous USA river network. Also, the application of ANN may be found for the study of flood frequency in Mumbai for the post-flood management system (Goyal et al. 2021).

The ANN model is a development of mathematical methods with a brain-like architecture. ANN model is a subfield of artificial intelligence in which a computer model of the biological brain is constructed. This comprises interlinked basic processing units (neurons or nodes) associated with weight linkages that interact collectively to provide a signal that solves a particular problem depending on the input signal obtained. The hidden layers and nodes play important roles in the implementation of ANN. In order to determine the suitable value of hidden nodes in the hidden layer, various guidelines were referred. Tang and Fishwick (1993) suggested the value to be x, Wong (1991) recommended the value to be $2x$, and Hecht-Nielsen (1990) suggested the value to be $2x + 1$, where x is the number of inputs. There are two types of activation functions adopted, namely the sigmoid and linear function.

EGMDH-ABC MODEL

The new ensemble group method of data handling coupled with the artificial bee colony algorithm improves the conventional GMDH model. The modification on the EGMDH-ABC model replaces the least square method with the ABC algorithm as parameter estimation of the partial description. Other than that, rather than using only the polynomial transfer function, three other transfer functions are also implemented in the new EGMDH-ABC model. At the end of the process of the EGMDH-ABC model, the output from all transfer functions will be the ensemble to become a new single output.

GROUP METHOD OF DATA HANDLING

Group Method of Data Handling (GMDH) is a type of machine learning introduced by Ivakhnenko (1970). The GMDH utilized in this study is an effective tool for solving problems, including forecasting and data mining. GMDH employs a neural network-based algorithm that explicitly learns the link between input and output variables and builds a model; GMDH can autonomously determine the optimum path to the intended outcome once the relationship is established. The GMDH model algorithm can educate the system to select the crucial independent variables. GMDH is an adaptive and self-organizing model where it can optimize itself according to data input. The Kolmogorov Gabor polynomial is used to define the input-output relationship in the GMDH model. The data will be split into two subsets of data; training and prediction data set. The training data set will be used for parameter estimation of the GMDH model. The GMDH model system only uses the second-order Kolmogorov Gabor polynomial, which is shown in Equation (3).

$$\hat{d}_k = b_0 + b_1 x_i + b_2 x_j + b_3 x_i x_j + b_4 x_i^2 + b_5 x_j^2 \qquad (3)$$

where $\hat{d}$ is the predicted flow, $b_0, b_1, b_2, b_3, b_4, b_5$ is the parameter and x is the site characteristics, i and j represent the vector of inputs at position. Equation (3) is known as the partial description (PD) of the GMDH model as it only consists a few parts of Volterra series. The parameter of Equation (3) is estimated using the least square method. The estimated amount of PD built on the first layer of the GMDH model is calculated based on this formula; $P = w(w\text{-}1)/2$ (Ayoub et al. 2022), where w is the number of input variables. The GMDH model was modelled by substituting each pair of two independent variables into the PD. Therefore, in the first layer, there is P number of PD $(\hat{d}_k)$. Every PD $(\hat{d}_k)$ output will be evaluated, and the best PD $(\hat{d}_k)$ will be chosen as the new input variable for the next layer of GMDH model, while the remainder PD $(\hat{d}_k)$ will be eliminated. The best PD was identified using mean absolute percentage error (MAPE). The MAPE is defined in Equation (4).

$$MAPE = \left(\frac{1}{n} \sum_{i=1}^{n} \left| \frac{Q_i - \hat{d}_i}{Q_i} \right| \right) \times 100\% \qquad (4)$$

This procedure is repeated until the termination criterion has been satisfied. The termination criterion or stopping condition of the GMDH process is when the minimum MAPE on the current layer is the same or higher than the previous layer of the GMDH model. The GMDH process concludes when the smallest MAPE on the current layer is equal to or greater than the preceding layer of the GMDH model.

ABC ALGORITHM

Karaboga and Basturk (2007) developed an Artificial Bee Colony (ABC) algorithm, which focused on a recreation of the pattern of behaviour of honeybee swarms as they forage for food. Generally, in the ABC algorithm, the bee colony is composed of 3 distinct types of bees: employed, onlooker, and scout bees (Karaboga & Akay 2009). Each bee performs a specific function to maximize the amount of nectar stored within the hive. The employed bees were mandated to identify food sources, collect information, and exchange food information with onlooker bees in the hive. When the food supply dries out, the employed bees shift into scout bees. The scout bee's function is to hunt for or explore new food sources immediately. The ABC algorithm's localization of a food supply is a plausible option for optimization. The quantity of nectar of a food supply aligns with the solution's quality (fitness). This algorithm comprises the initiation stage, the employed bee stage, the observer bee stage, and the scout bee stage (Aslan 2019). Following the initialization stages, the algorithm's three main steps are repeated indefinitely until the termination criterion is satisfied. The ABC algorithm generates a randomly distributed BN solution population (food supplies) during the initiation stage. The number of SN solutions is adjusted by the fraction of employed and onlooker bees is set to be similar. The initial food supply X_i (i - 1,2, ..., BN) produced within the restricted range of j^{th} index by Equation (5). Equation (5) denotes as follow (Xiang & An 2013):

$$x_{pq} = x_q^{\min} + rand(0,1)(x_q^{\max} - x_q^{\min}) \qquad (5)$$

where p = 1,2, ..., BN, q = 1,2, ..., D and D is the problem size or the number of optimized parameters. $x_q^{\min}$ and $x_q^{\max}$ are the upper and lower bounds for the dimension q, respectively. Then, each solution is assessed by Badem et al. (2017):

$$fs_p = \begin{cases} \dfrac{1}{(1+f_p)}, & f_p \geq 0 \\ 1 + |f_p|, & f_p \leq 0 \end{cases}$$

where fs_p denotes the fitness value of the solution p and f_p is the cost function for a minimization problem.

Based on probabilistic selection based on the roulette wheel, the onlooker bee will tend to visit a better food supply (Yurtkuran & Emel 2016). Thus, the onlooker bees strive to locate a fresh candidate of food supply situated around the excellent solution. Each employed bee introduces a new food supply (solution) in the vicinity of the previously picked solution during the employed bee stages. As with scout bee stages, if the fitness value of a food supply does not decrease over a specified number of cycles, the food supply will be discarded, and the associated employed bee will transform into a scout bee. The process is repeated until the completion procedure is successfully finished. The termination criteria may be the maximum number of cycles, or the output must be adequate.

EGMDH-ABC MODEL

The EGMDH-ABC model is set up below: *Step 1* Identify the input variables $\{x_1, x_2, ..., x_5\}$ which is the site characteristics and output variables Q_i which is flood quantile. Afterwards, the pooled data is divided into a training and testing data set. In the context of testing in an ungauged site, only a single data is removed to become the testing data to simulate the ungauged location. The remaining data, which is the training data set, will be used to obtain the PD parameter. PD description is described in Equation (3). If necessary, normalization of the original data will be performed.

Step 2 In this step, the type of transfer function in the GMDH model is determined. There are four types of transfer functions used in the EGMDH-ABC model, namely polynomial (PLF), radial basis (RBF), sigmoid (SGF) and hyperbolic tangent function (HTF). The four types of transfer function are shown in Table 2.

TABLE 2. Type of transfer function

Transfer Function	
PLF	$y(plf)_k = d_k$
SGF	$y(s)_k = \dfrac{1}{(1+e^{-d_k})}$
RBF	$y(rbf)_k = e(-d_k^2)$
HTF	$y(htf)_k = \left(\dfrac{2}{1+e^{-2d_k}}\right) - 1$

*where d_k is a partial description that has been described in Equation (3)

Step 3 The development of each transfer function will run separately. The parameter of PD for each transfer function will be estimated using the ABC algorithm. A set of linear equations will be constructed before applying the ABC algorithm. A set of a linear system can be illustrated as follows:

$$Q = Vb$$
$$Q = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix} V = \begin{bmatrix} 1 & x_{1i} & x_{1j} & x_{1i}x_{ij} & x_{1i}^2 & x_{1j}^2 \\ 1 & x_{2i} & x_j & G_{2i}x_{2j} & x_{2i}^2 & x_{2j}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{li} & x_{lj} & x_{li}x_{lj} & x_{li}^2 & x_{lj}^2 \end{bmatrix} b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_5 \end{bmatrix} \quad (6)$$

Table 3 summarises the transfer functions for the MGMDH model.

TABLE 3. t_i transformation according to transfer function

Transfer Function	
PLF	$t_i = Q_i$
SGF	$t_i = \ln\left(\dfrac{Q_i}{1-Q_i}\right)$
RBF	$t_i = \sqrt{-\ln Q_i}$
HTF	$t_i = -\dfrac{1}{2}\ln\left(\dfrac{2}{Q_i+1} - 1\right)$

*where i = 1,2,3, ..., n

Step 4 The parameters for every PD for each transfer function are obtained using the ABC algorithm. The ABC algorithm has been discussed in the previous section. This process is repeated until the realization of the system is achieved. The realization of the system is achieved when the termination criteria are met. The termination criteria for each transfer function model are the same as the GMDH model when the MAPE on the current layer is the same or greater than the previous layer. The transfer function output is chosen based on the output from the preceding layer with the lowest MAPE. Suppose the realization of the system does not achieved. In that case, the output that produces the lowest MAPE will become the new input variable for the next layer, and the process will start from Step 1 until the stopping criteria are met.

Step 5 The last step of the EGMDH-ABC model is to apply the average ensemble concept. The illustration of the ensemble average concept is shown in Equation (7).

$$\hat{Q} = \frac{\hat{d}_{PLF} + \hat{d}_{SGF} + \hat{d}_{RBF} + \hat{d}_{HTF}}{4} \tag{7}$$

$\hat{Q}$ is the predicted output from the EGMDH-ABC model. The output is the average output from all transfers function implement in this study.

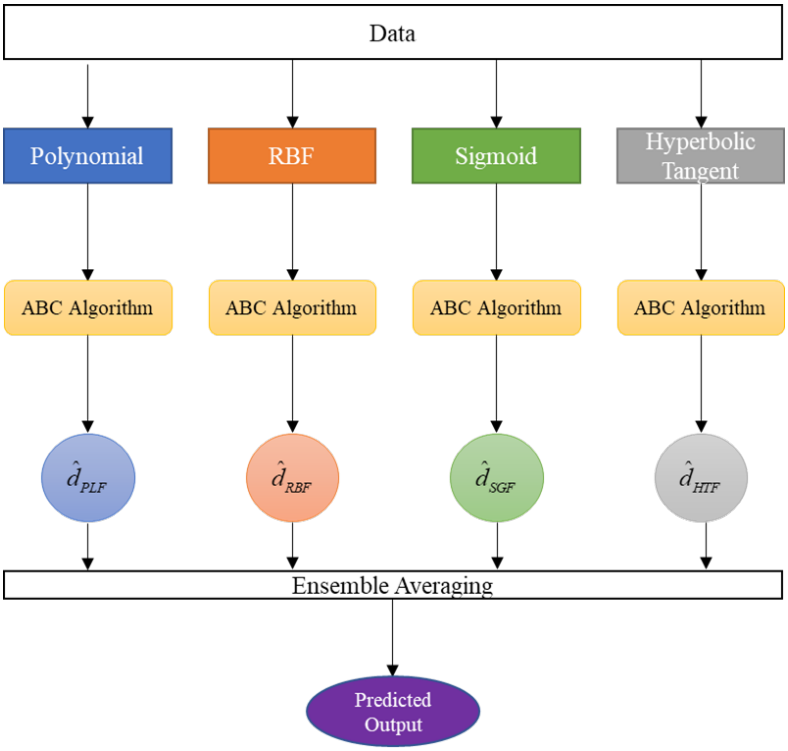

FIGURE 1. Illustration of EGMDH-ABC model

Figure 1 shows the illustration of the EGMDH-ABC model. It shows the initial stage of the EGMDH-ABC model until the predicted output is obtained. Figure 2 shows the flowchart of the EGMDH-ABC model.

EVALUATION METRICS OF THE HYDROLOGIC MODELS

Three statistical error indicators were employed to evaluate the models' performance, namely the Mean Absolute Percentage Error (MAPE) (Lee et al. 2021), Nash-Sutcliffe efficiency (NSE) (Criss & Winston 2008; Yin et al. 2021), and the BIASr (Desai & Ouarda 2021).

$$MAPE = \left(\frac{1}{n} \sum_{i=1}^{n} \left| \frac{Q_i - \hat{Q}_i}{Q_i} \right| \right) \times 100\% \tag{8}$$

$$NASH = 1 - \frac{\sum_{i=1}^{n} (Q_i - \hat{Q}_i)^2}{\sum_{i=1}^{n} (Q_i - \bar{Q}_i)^2} \tag{9}$$

$$BIASr = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Q_i - \hat{Q}_i}{Q_i}\right) \quad (10)$$

where Q_i is at-site; T-year flood quantile of site i ; $\hat{Q}_i$ is the predicted flood quantile of site i; n is the total number of site; and $\overline{Q}_i$ is the mean of flood quantile with T-year return period.

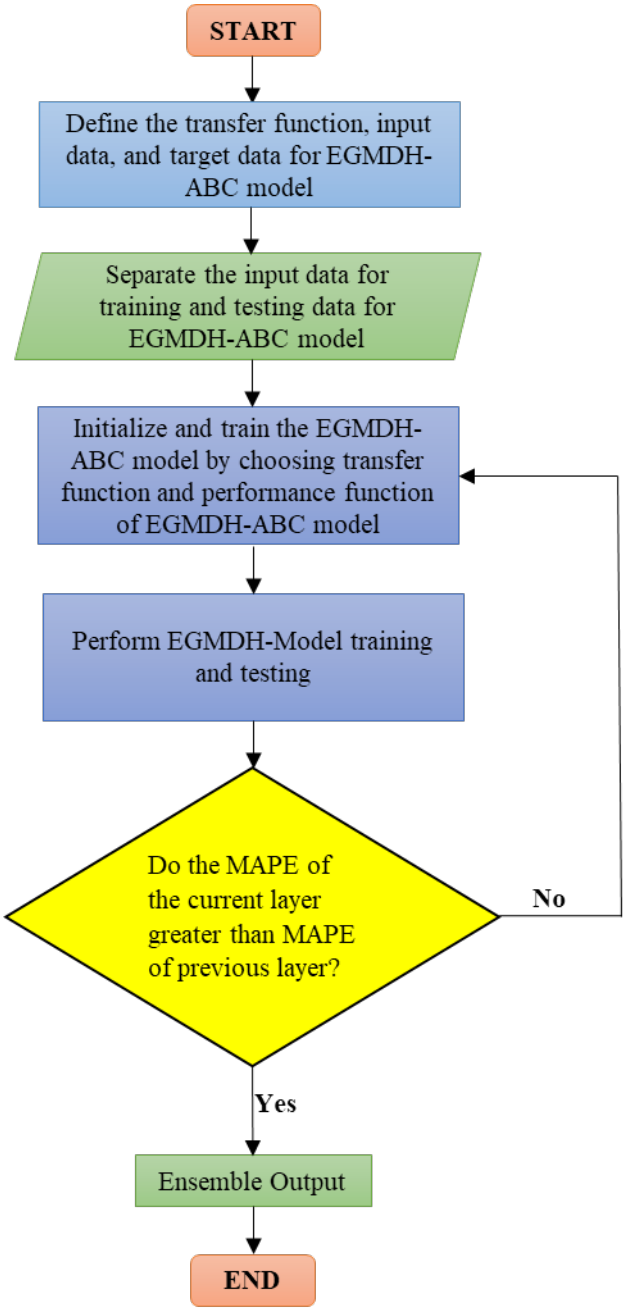

FIGURE 2. Flowchart of EGMDH-ABC model

RESULTS AND DISCUSSION

This study utilized the optimization method and ensemble technique to the standalone GMDH model. The selected optimization method to be implemented in the GMDH model is the ABC algorithm. Meanwhile, the ensemble technique used for this study is ensemble

averaging as illustrated in Figure 1. Other than that, various transfer functions are applied to the proposed EGMDH-ABC model: sigmoid, hyperbolic tangent, radial basis, and polynomial. It is slightly different from the conventional GMDH model, which only uses a single polynomial transfer function. The motivation to use various transfer functions in a single model since each data is unique. Therefore, the advantageous of implementing various transfer function in EGMDH model, it will be able to capture the uniqueness of each data set. The ABC algorithm will replace the least square transfer function to estimate the parameter of the PD. After obtaining the parameter of the PD, the best output from each transfer function will be the ensemble to get the outcome.

In order to assess the proposed model performance, 60 hydrometric stations located in Peninsular Malaysia is used in the present study. Jackknife procedure is used as the model validation technique in this work. This technique involved deleting one site from the data set that was presumed to be ungauged and used the remaining sites in the data set to construct the prediction model's parameter. The procedure was repeated until all of the sites were discarded at least once. As a result, the number of models produced is equal to the total of hydrometric stations studied. Following the previous research (Desai & Ouarda 2021 & Pandey & Nguyen 1999), the return period chosen in this study consists of 10 years and 100 years, incorporating the high and low distributions. Additional analysis was performed to identify the importance of the predictor variables for flood quantile estimation. The result is illustrated in Table 3.

TABLE 3. The relative importance of the predictor's variables

Predictor	Relative importance	
	Q_{10}	Q_{100}
AE	45.52%	44.08%
VT	1.46%	1.71%
PTH	47.56%	47.95%
RS	3.46%	3.45%
AP	2%	2.81%

From Table 3, catchment area (AE) is shown to be by far the most important physio-meteorological variable, followed by longest drainage path (PTH) for both flood quantile with 10 years and 100 years return period. River slope (RS) and annual precipitation (AP) are distant third and fourth, respectively. Elevation of the station (VT) is the least essential variable of all physio-metrological variables. Therefore, based on the Table 3 results, various combinations of predictor variables are implemented in the prediction model. There will be an additional experiment for the ANN model as each combination of predictor variables will be tested using three different number hidden layers. The determination of the hidden is discussed in the ANN methodology section. The best output for each prediction model based on evaluation metrics will be selected for comparison. The results of the applications of the proposed model EGMDH-ABC model and comparison model to the dataset are illustrated in Table 4.

TABLE 4. Evaluation metric results for prediction model

	NASH		BIASr		MAPE	
Model	Q_{10}	Q_{100}	Q_{10}	Q_{100}	Q_{10}	Q_{100}
LR	0.7212	0.7124	0.3327	0.3910	78.14%	80.23%
GMDH	0.8436	0.8108	0.0623	-0.1722	27.97%	26.18%
ANN	0.9079	0.8721	0.0006	0.0210	26.37%	22.98%
EGMDH-ABC	**0.9383**	**0.9270**	-0.0184	-0.0145	**18.56%**	**16.67%**

Table 4 shows the prediction efficiency of the LR model, GMDH model, ANN model, and EGMDH model in terms of MAPE, NASH and, BIASr. In terms of MAPE, a lower value of MAPE indicates excellent prediction performance. The EGMDH-ABC model has the lowest MAPE value. Therefore, it is superior to the LR model, GMDH model, and ANN model. It is linked to the ability of the ABC algorithm, contributes to an optimization of the parameter for PD, and the implementations of various transfer functions in the EGMDH-ABC model. It shows that the optimization method (ABC algorithm) and ensemble technique significantly reduced the relative MAPE of the conventional GMDH model for both return periods with T=10 years and 100 years, respectively.

As for the NASH evaluation metrics, a model with excellent estimation yields a NASH value of one. A model that is considered accurate always has a NASH value greater than 0.8. From Table 4, the NASH value of the EGMDH-ABC model is the highest for flood quantile prediction with T=10 years and T=100 years, respectively. Other than that, the GMDH model, ANN model, and EGMDH-ABC model produced a NASH value greater than 0.8 for the two specific flood quantiles. It is observed that the NASH value decreased when the model is predicting a larger flood quantile. This is due to the smaller (catchment area) site has large return period and as the NASH evaluation metrics are sensitive to outliers and sample size (McCuen et al. 2006; Mokhtar et al. 2021). Therefore, the decrease of NASH value contributed to the smaller site with a large return period value. Other than that, the number of sample size also affects the NASH value. The total sample size is smaller when the data is split into the training and testing data set compared to the original data set. However, as the sample size is the same for both specific quantiles, thus, the significant factor contributing to decreasing the NASH value for the higher return period is the extreme value of the return period for a certain target site.

The BIASr evaluation metric is used to examine whether the model underestimates or overestimates for both specific quantiles. In terms of BIASr, the EGMDH-ABC model underestimate the flood quantiles. For the 10-years return period, the ANN model has the lowest BIASr, while for the 100-years return period, the EGMDH-ABC model has the lowest BIASr. Although the ANN model has the lowest BIASr for a 10-years return period, it is observed that the EGMDH-ABC model still leads the best MAPE and NASH values among all models used in this study. Figures 3 and 4 show the scatter plot of flood quantile prediction using LR, GMDH, ANN, and EGMDH-ABC models, respectively.

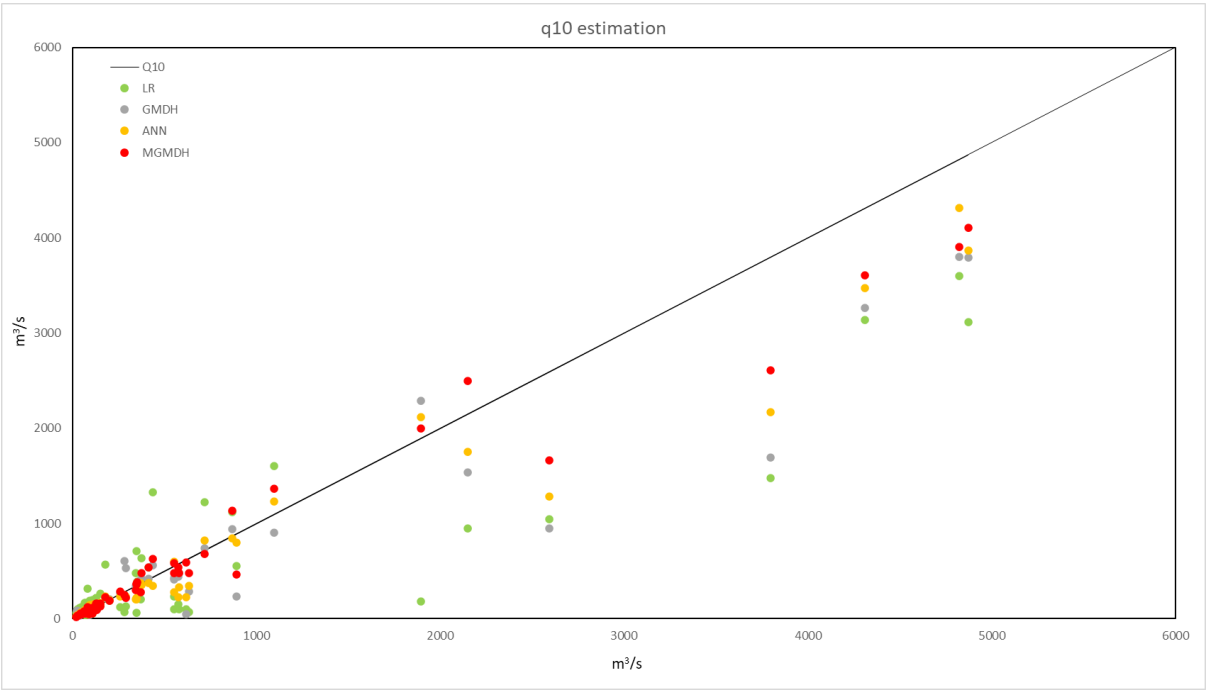

FIGURE 3. q10 estimation for LR model, GMDH model, ANN model, and EGMDH-ABC model

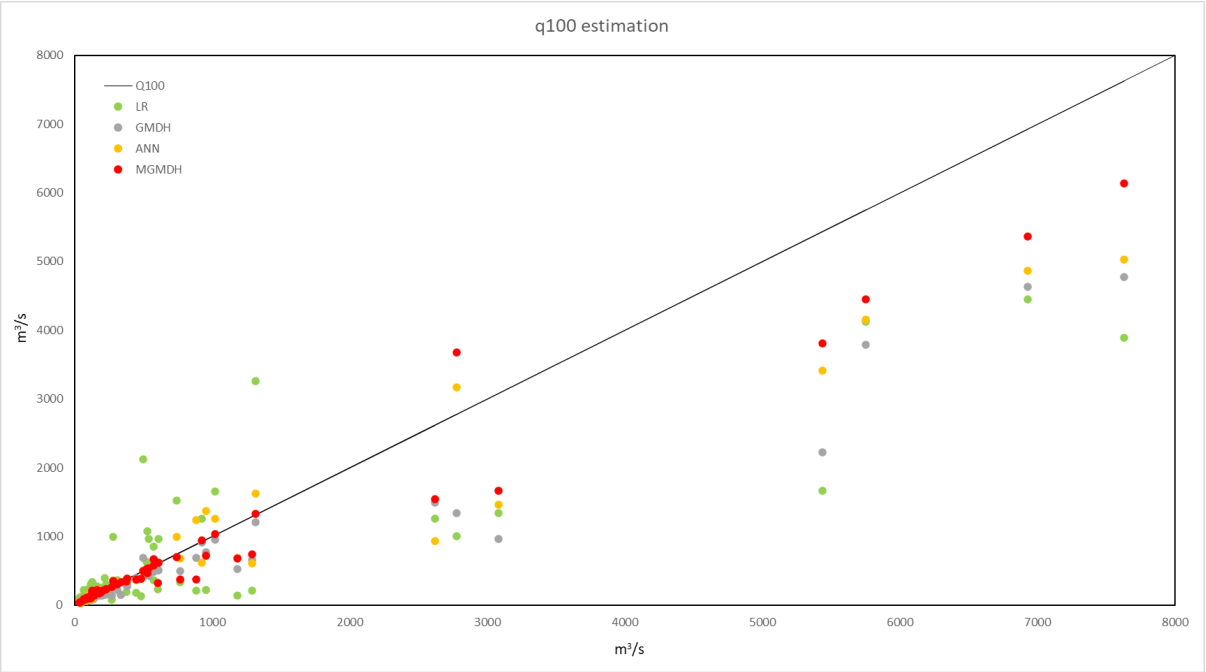

FIGURE 4. q100 estimation for LR model, GMDH model, ANN model, and EGMDH-ABC model

Figures 3 and 4 is the quantile-quantile plot for predicted flood quantile and observed flood quantile. The observation from Figures 3 and 4 show that all the prediction models underestimate flood quantiles with larger flood quantiles values. As expected, when the return period increases, the flood quantile increases simultaneously. This explained that the NASH value decreases when the return period increases because the data variation increased. Although the ANN model shows a great prediction, the drawback of the ANN model is determining the suitable structure for the ANN model. As we need to consider all the possible structures of the ANN model, the computational is very high. Overall, it can be concluded that implementing the ABC algorithm and ensemble technique to the GMDH model significantly improves the performance of the conventional GMDH model for flood quantile prediction at the ungauged site. The proposed EGMDH-ABC model outperformed the LR, GMDH, and ANN models for both return periods.

CONCLUSION

GMDH model, ABC algorithm, and ensemble technique are widely used for image classification, wind speed prediction, stock market prediction, e-banking risk measurement, and house pricing prediction. However, the application of the GMDH model, ABC algorithm, and ensemble technique is rarely found for hydrological fields, especially in prediction at the ungauged site. Usually, for the study of flood quantile prediction, a linear model such as MLR is used to construct the relationship between predictor variables and observed flood quantiles. The predictor's variables consist of meteorological and physiographical variables. The problem of linear relationships usually does not capture the complex nonlinear relationship between the predictor and target variables. GMDH model is among the best data-driven models to capture the nonlinear relationship between predictor and target variables, which has proven to have a good prediction performance in various fields. This study aims to enhance the GMDH model by implementing the ABC algorithm to optimize the parameter of PD. Then various transfer function is applied in the GMDH model. Finally, ensemble averaging is used to combine the output from various transfer functions in the EGMDH-ABC model. The result shows that the enhancement of the GMDH model significantly improves the prediction performance of the GMDH model. It shows that the EGMDH-ABC model can capture the nonlinearity in the data set to produce a better estimation than other models. The EGMDH-ABC model's performance is better compared to the single model GMDH model. The results further indicate that the EGMDH-ABC model provides

more robust, accurate, and efficient results. In future work, investigate the ensemble ANN model with different transfer functions or ensemble the two other models, which are the ANN model and the GMDH model.

REFERENCES

Adnan, R.M., Liang, Z., Parmar, K.S., Soni, K. & Kisi, O. 2021. Modeling monthly streamflow in mountainous basin by MARS, GMDH-NN and DENFIS using hydroclimatic data. *Neural Computing and Applications* 33(7): 2853-2871.

Ahmadi, A., Nasseri, M. & Solomatine, D.P. 2019. Parametric uncertainty assessment of hydrological models: Coupling UNEEC-P and a fuzzy general regression neural network. *Hydrological Sciences Journal* 64(9): 1080-1094.

Ahmadi, M.H., Ahmadi, M-A., Mehrpooya, M. & Rosen, M.A. 2015. Using GMDH neural networks to model the power and torque of a stirling engine. *Sustainability* 7(2): 2243-2255.

Alobaidi, M.H., Ouarda, T.B.M.J., Marpu, P.R. & Chebana, F. 2021. Diversity-driven ANN-based ensemble framework for seasonal low-flow analysis at ungauged sites. *Advances in Water Resources* 147: 103814.

Amiri, M. & Soleimani, S. 2021. ML-based group method of data handling: An improvement on the conventional GMDH. *Complex & Intelligent Systems* 7: 2949-2960.

Ashrafzadeh, A., Kişi, O., Aghelpour, P., Biazar, S.M. & Masouleh, M.A. 2020. Comparative study of time series models, support vector machines, and GMDH in forecasting long-term evapotranspiration rates in northern Iran. *Journal of Irrigation and Drainage Engineering* 146(6): 04020010.

Aslan, S. 2019. A transition control mechanism for artificial bee colony (ABC) algorithm. *Computational Intelligence and Neuroscience* 2019: Article ID. 5012313.

Ayoub, M.A., Elhadi, A., Fatherlhman, D., Saleh, M.O., Alakbari, F.S. & Mohyaldinn, M.E. 2022. A new correlation for accurate prediction of oil formation volume factor at the bubble point pressure using group method of data handling approach. *Journal of Petroleum Science and Engineering* 208: 109410.

Aziz, K., Haque, M.M., Rahman, A., Shamseldin, A.Y. & Shoaib, M. 2017. Flood estimation in ungauged catchments: Application of artificial intelligence based methods for Eastern Australia. *Stochastic Environmental Research and Risk Assessment* 31(6): 1499-1514.

Badem, H., Basturk, A., Caliskan, A. & Yuksel, M.E. 2017. A new efficient training strategy for deep neural networks by hybridization of artificial bee colony and limited–memory BFGS optimization algorithms. *Neurocomputing* 266: 506-526.

Badyalina, B. & Shabri, A. 2015. Flood estimation at ungauged sites using group method of data handling in Peninsular Malaysia. *Jurnal Teknologi* 76(1). https://doi.org/10.11113/jt.v76.2640

Badyalina, B., Mokhtar, N.A., Mat Jan, N.A., Hassim, N.H. & Yusop, H. 2021a. Flood frequency analysis using L-moment for Segamat River. *MATEMATIKA: Malaysian Journal of Industrial and Applied Mathematics* 37(2): 47-62.

Badyalina, B., Shabri, A. & Marsani, M.F. 2021b. Streamflow estimation at ungauged basin using modified group method of data handling. *Sains Malaysiana* 50(9): 2765-2779.

Broderick, C., Matthews, T., Wilby, R.L., Bastola, S. & Murphy, C. 2016. Transferability of hydrological models and ensemble averaging methods between contrasting climatic periods. *Water Resources Research* 52(10): 8343-8373.

Campos, J.A. & Pedrollo, O.C. 2021. A regional ANN-based model to estimate suspended sediment concentrations in ungauged heterogeneous basins. *Hydrological Sciences Journal* 66(7): 1222-1232.

Cannon, A.J. 2010. A flexible nonlinear modelling framework for nonstationary generalized extreme value analysis in hydroclimatology. *Hydrological Processes: An International Journal* 24(6): 673-685.

Criss, R.E. & Winston, W.E. 2008. Do Nash values have value? Discussion and alternate proposals. *Hydrological Processes: An International Journal* 22(14): 2723-2725.

De Paola, F., Giugni, M., Pugliese, F., Annis, A. & Nardi, F. 2018. GEV parameter estimation and stationary vs. non-stationary analysis of extreme rainfall in African test cities. *Hydrology* 5(2): 28.

Desai, S. & Ouarda, T.B.M.J. 2021. Regional hydrological frequency analysis at ungauged sites with random forest regression. *Journal of Hydrology* 594: 125861.

Elbaz, K., Shen, S-L., Zhou, A., Yin, Z-Y. & Lyu, H-M. 2021. Prediction of disc cutter life during shield tunneling with AI via the incorporation of a genetic algorithm into a GMDH-type neural network. *Engineering* 7(2): 238-251.

Fillipova, V., Leedal, D. & Hammond, A. 2020. *Regional Flood Frequency Estimation for the Contiguous USA using Artificial Neural Networks*. EGU General Assembly Conference Abstracts.

Goyal, H.R., Ghanshala, K.K. & Sharma, S. 2021. Post flood management system based on smart IoT devices using AI approach. *Materials Today: Proceedings*.

Guru, N. & Jha, R. 2014. A study on selection of probability distributions for at-site flood frequency analysis in Mahanadi River Basin, India. http://dx.doi.org/10.1201/b17133-241

Hecht-Nielsen, R. 1990. *Neurocomputing.* Boston: Addison-Wesley. pp. 89-93.

Hosking, J.R.M. & Wallis, J.R. 1997. *Regional Frequency Analysis*: *An Approach Based on L-moments.* Cambrige: Cambrige University Press. http://dx.doi.org/10.1017/cbo9780511529443

Hosseini, S.A., Taheri, B., Abyaneh, H.A. & Razavi, F. 2021. Comprehensive power swing detection by current signal modeling and prediction using the GMDH method. *Protection and Control of Modern Power Systems* 6(1): 1-11.

Ivakhnenko, A.G. 1971. Polynomial theory of complex systems. *IEEE Transactions on Systems, Man, and Cybernetics* 4: 364-378.

Ivakhnenko, A.G. 1970. Heuristic self-organization in problems of engineering cybernetics. *Automatica* 6(2): 207-219.

Jolánkai, Z. & Koncsos, L. 2018. Base flow index estimation on gauged and ungauged catchments in Hungary using digital filter, multiple linear regression and artificial neural networks. *Periodica Polytechnica Civil Engineering* 62(2): 363-372.

Karaboga, D. & Akay, B. 2009. A comparative study of artificial bee colony algorithm. *Applied Mathematics and Computation* 214(1): 108-132.

Karaboga, D. & Basturk, B. 2007. A powerful and efficient algorithm for numerical function optimization: Artificial bee colony (ABC) algorithm. *Journal of Global Optimization* 39(3): 459-471.

Kardani, N., Bardhan, A., Kim, D., Samui, P. & Zhou, A. 2021. Modelling the energy performance of residential buildings using advanced computational frameworks based on RVM, GMDH, ANFIS-BBO and ANFIS-IPSO. *Journal of Building Engineering* 35: 102105.

Khan, M.S.R., Hussain, Z. & Ahmad, I. 2021. Regional flood frequency analysis, using l-moments, artificial neural networks and OLS regression, of various sites of Khyber-Pakhtunkhwa, Pakistan. *Applied Ecology and Environmental Research* 19(1): 471-489.

Kordrostami, S., Alim, M.A., Karim, F. & Rahman, A. 2020. Regional flood frequency analysis using an artificial neural network model. *Geosciences* 10(4): 127.

Le, L.T., Nguyen, H., Dou, J. & Zhou, J. 2019. A comparative study of PSO-ANN, GA-ANN, ICA-ANN, and ABC-ANN in estimating the heating load of buildings' energy efficiency for smart city planning. *Applied Sciences* 9(13): 2630.

Lee, W.H., Choi, H.S., Lee, D. & Choi, B. 2021. Stream flow generation for simulating yearly bed change at an ungauged stream in monsoon region. *Water* 13(4): 554.

Lu, R., Hu, H., Xi, M., Gao, H. & Pun, C-M. 2019. An improved artificial bee colony algorithm with fast strategy, and its application. *Computers & Electrical Engineering* 78: 79-88.

Mamun, A.A., Hashim, A. & Amir, Z. 2012. Regional statistical models for the estimation of flood peak values at ungauged catchments: Peninsular Malaysia. *Journal of Hydrologic Engineering* 17(4): 547-553. doi: doi:10.1061/(ASCE) HE.1943-5584.0000464.

Maofa Wang, Mohammad Rezaie-Balf, Sujay Raghavendra Naganna & Zaher Mundher Yaseen. 2021. Sourcing CHIRPS precipitation data for streamflow forecasting using intrinsic time-scale decomposition based machine learning models. *Hydrological Sciences Journal* 66(9): 1437-1456.

Mat Jan, N.A., Shabri, A., Hounkpè, J. & Badyalina, B. 2018. Modelling non-stationary extreme streamflow in Peninsular Malaysia. *International Journal of Water* 12(2): 116-140.

Mat Jan, N.A., Shabri, A., Ismail, S., Badyalina, B., Abadan, S.S. & Yusof, N. 2016a. Three-parameter lognormal distribution: Parametric estimation using L-moment and TL-moment approach. *Jurnal Teknologi* 78: 6-11.

Mat Jan, N.A., Shabri, A. & Badyalina, B. 2016b. Selecting probability distribution for regions of Peninsular Malaysia streamflow. *AIP Conference Proceedings*. 1750: 060014.

McCuen, R.H., Knight, Z. & Cutter, A.G. 2006. Evaluation of the Nash-Sutcliffe Efficiency Index. *Journal of Hydrologic Engineering* 11(6): DOI:10.1061/(ASCE)1084-0699(2006)11:6(597).

Meresa, H. 2019. Modelling of river flow in ungauged catchment using remote sensing data: Application of the empirical (SCS-CN), artificial neural network (ANN) and hydrological model (HEC-HMS). *Modeling Earth Systems and Environment* 5(1): 257-273.

Mokhtar, N.A., Zubairi, Y.Z., Hussin, A.G., Badyalina, B., Ghazali, A.F., Ya'acob, F.F. & Kerk, L.C. 2021. Modelling wind direction data of Langkawi Island during Southwest monsoon in 2019 to 2020 using bivariate linear functional relationship model with von Mises distribution. *Journal of Physics: Conference Series* 1988(1): 012097.

Nariman Valizadeh, Majid Mirzaei, Mohammed Falah Allawi, Haitham Abdulmohsin Afan, Nuruol Syuhadaa Mohd, Aini Hussain, & Ahmed El-Shafie. 2017. Artificial intelligence and geo-statistical models for streamflow forecasting in ungauged stations: State of the art. *Natural Hazards* 86(3): 1377-1392.

Otiniano, C.E.G., De Paiva, B.S. & Neto, D.S.B. 2019. The transmuted GEV distribution: Properties and application. *Communications for Statistical Applications and Methods* 26(3): 239-259.

Pandey, G.R. & Nguyen, V-T-V. 1999. A comparative study of regression based methods in regional flood frequency analysis. *Journal of Hydrology* 225(1-2): 92-101.

Shu, C. & Burn, D.H. 2004. Artificial neural network ensembles and their application in pooled flood frequency analysis. *Water Resources Research* 40(9). https://doi.org/10.1029/2003WR002816

Shu, C. *&* Ouarda, T.B.M.J. 2008. Regional flood frequency analysis at ungauged sites using the adaptive neuro-fuzzy inference system. *J. Hydrol.* 349*(*1-2): 31-43. *doi:*10.1016/j.jhydrol.2007.10.050.

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: doi: 10.1029/2006WR005142.

Sivakumar, B. & Singh, V.P. 2012. Hydrologic system complexity and nonlinear dynamic concepts for a catchment classification framework. *Hydrology and Earth System Sciences* 16(11): 4119-4131.

Solanki, P., Baldaniya, D., Jogani, D., Chaudhary, B., Shah, M. & Kshirsagar, A. 2021. Artificial intelligence: New age of transformation in petroleum upstream. *Petroleum Research* 7(1): 106-114.

Tan, A., Zhou, G. & He, M. 2021. Surface defect identification of Citrus based on KF-2D-Renyi and ABC-SVM. *Multimedia Tools and Applications* 80(6): 9109-9136.

Tang, Z. & Fishwick, P.A. 1993. Feedforward neural nets as models for time series forecasting. *ORSA Journal on Computing* 5(4): 374-385.

Tegegne, G., Kim, Y-O. & Lee, J-K. 2019. Spatiotemporal reliability ensemble averaging of multimodel simulations. *Geophysical Research Letters* 46(21): 12321-12330.

Tereshko, V. & Lee, T. 2002. How information-mapping patterns determine foraging behaviour of a honey bee colony. *Open Systems & Information Dynamics* 9(2): 181-193.

Wan Zawiah Wan Zin, Abdul Aziz Jemain, Kamarulzaman Ibrahim, Jamaludin Suhaila & Mohd Deni Sayang. 2009. A comparative study of extreme rainfall in Peninsular Malaysia: With reference to partial duration and annual extreme series. *Sains Malaysiana* 38(5): 751-760.

Wong, F.S. 1991. Time series forecasting using backpropagation neural networks. *Neurocomputing* 2(4): 147-159.

Wu, J., Wang, Y., Zhang, X. & Chen, Z. 2016. A novel state of health estimation method of Li-ion battery using group method of data handling. *Journal of Power Sources* 327: 457-464.

Xiang, W-L. & An, M-Q. 2013. An efficient and robust artificial bee colony algorithm for numerical optimization. *Computers & Operations Research* 40(5): 1256-1265.

Xiao, Y., Wu, J., Lin, Z. & Zhao, X. 2018. A deep learning-based multi-model ensemble method for cancer prediction. *Computer Methods and Programs in Biomedicine* 153: 1-9.

Yang, S., Yang, D., Chen, J., Santisirisomboon, J., Lu, W. & Zhao, B. 2020. A physical process and machine learning combined hydrological model for daily streamflow simulations of large watersheds with limited observation data. *Journal of Hydrology* 590: 125206.

Yin, H., Guo, Z., Zhang, X., Chen, J. & Zhang, Y. 2021. Runoff predictions in ungauged basins using sequence-to-sequence models. *Journal of Hydrology* 603: 126975.

Yurtkuran, A. & Emel, E. 2016. A discrete artificial bee colony algorithm for single machine scheduling problems. *International Journal of Production Research* 54(22): 6860-6878.

*Corresponding author; email: basribdy@uitm.edu.my