Sains Malaysiana 50(9)(2021): 2765-2779
http://doi.org/10.17576/jsm-2021-5009-22

Streamflow Estimation at Ungauged Basin using Modified Group Method of Data Handling

(Anggaran Aliran Sungai di Lembangan Tiada Data menggunakan Kaedah Kumpulan Terubahsuai Pengendalian Data)

BASRI BADYALINA*, ANI SHABRI & MUHAMMAD FADHIL MARSANI

ABSTRACT

Among the foremost frequent and vital tasks for hydrologist is to deliver a high accuracy estimation on the hydrological variable, which is reliable. It is essential for flood risk evaluation project, hydropower development and for developing efficient water resource management. Presently, the approach of the Group Method of Data Handling (GMDH) has been widely applied in the hydrological modelling sector. Yet, comparatively, the same tool is not vastly used for the hydrological estimation at ungauged basins. In this study, a modified GMDH (MGMDH) model was developed to ameliorate the GMDH model performance on estimating hydrological variable at ungauged sites. The MGMDH model consists of four transfer functions that include polynomial, hyperbolic tangent, sigmoid and radial basis for hydrological estimation at ungauged basins; as well as; it incorporates the Principal Component Analysis (PCA) in the GMDH model. The purpose of PCA is to lessen the complexity of the GMDH model; meanwhile, the implementation of four transfer functions is to enhance the estimation performance of the GMDH model. In evaluating the effectiveness of the proposed model, 70 selected basins were adopted from the locations throughout Peninsular Malaysia. A comparative study on the performance was done between the MGMDH and GMDH model as well as with other extensively used models in the area of flood quantile estimation at ungauged basins known as Linear Regression (LR), Nonlinear Regression (NLR) and Artificial Neural Network (ANN). The results acquired demonstrated that the MGMDH model possessed the best estimation with the highest accuracy comparatively among all models tested. Thus, it can be deduced that MGMDH model is a robust and efficient instrument for flood quantiles estimation at ungauged basins.

Keywords: GMDH; hyperbolic tangent; PCA; radial basis; ungauged basin

ABSTRAK

Antara tugas yang paling kerap dan penting bagi ahli hidrologi ialah memberikan anggaran ketepatan yang tinggi untuk pemboleh ubah hidrologi yang boleh dipercayai. Ini adalah sangat penting untuk projek penilaian risiko banjir, pembangunan tenaga air dan untuk pengurusan sumber air yang cekap. Pada masa ini, pendekatan Kaedah Pengendalian Data (GMDH) telah banyak digunakan dalam sektor pemodelan hidrologi. Namun, secara perbandingan, model tersebut tidak banyak digunakan untuk anggaran pemboleh ubah hidrologi di lembangan yang tiada data. Dalam kajian ini, model GMDH yang diubah suai (MGMDH) dikembangkan untuk memperbaiki prestasi model GMDH dalam menganggar pemboleh ubah hidrologi di lokasi yang tiada data. Model MGMDH terdiri daripada empat fungsi pemindahan yang merangkumi polinomial, hiperbolik tangen, sigmoid dan asas radial untuk anggaran pemboleh ubah hidrologi di lembangan yang tiada data; serta; ia menggabungkan Analisis Komponen Utama (PCA) dalam model GMDH. Tujuan PCA adalah untuk mengurangkan kerumitan model GMDH; Sementara itu, pelaksanaan empat fungsi pemindahan adalah untuk meningkatkan prestasi anggaran model GMDH. Untuk menilai keberkesanan model yang dicadangkan, 70 lembangan dari lokasi di seluruh Semenjung Malaysia telah dipilih. Kajian perbandingan mengenai prestasi dilakukan antara model MGMDH dan GMDH serta model lain yang digunakan secara meluas di kawasan taksiran

kuantitatif banjir di lembangan yang tiada data yang dikenali sebagai Regresi Linear (LR), Regresi Bukan Linear (NLR) dan Rangkaian Neural Buatan (ANN). Hasil yang diperoleh menunjukkan bahawa model MGMDH memiliki anggaran terbaik dengan ketepatan yang tertinggi berbanding semua model yang diuji. Oleh itu, dapat disimpulkan bahawa model MGMDH adalah instrumen yang kuat dan cekap untuk anggaran kuantil banjir di lembangan yang tiada data.

Kata kunci: Asas radial; GMDH; hiperbolik tangen; lembangan tiada data; PCA

Introduction

Data availability of data is crucial for any field that involves planning and decision making, especially in the management of water resources. The availability of hydrological data can provide a proper evaluation of water resource projects such as drainage design, flood control design and low impact development. However, Yang et al. (2019) state that most of the streams around the world are ungauged or partially ungauged. Sivapalan et al. (2003) describe ungauged basins as a hydrological station with unsatisfactory streamflow records. The sparse records are due to expensive, problematic and time-consuming task to build a hydrological station, especially for a stream located in a remote area. Abdullah et al. (2012) have reported that the stream situated in Malaysia is gauged only in a developed area, whereas the stream in the remote area is mostly ungauged. Several extensive majority studies on regionalization methods are by relating basin characteristics and the probability distribution that fit the flow series using a data-driven model (Aziz et al. 2017; Kim et al. 2019; Ojha et al. 2018). A power form equation is the most common approach to build connectivity between basin characteristics and hydrological variables at selected basins (Arsenault et al. 2018; Shu & Ouarda 2008; Tsegaw et al. 2019; Walega et al. 2020). The process that transfers the information from the gauged basin to the ungauged basin is known as the regionalization method. There are a significant number of researches employed linear regression (LR) model in regionalization method (Alobaidi et al. 2015; Arsenault et al. 2018; Shu & Ouarda 2008; Walega et al. 2020).

Pandey and Nguyen (1999) have proposed a different approach to solving the power form equation namely the nonlinear regression (NLR) model. The NLR model particularly useful because it can solve the power form equation directly compare to the linear regression and use the real domain of hydrological variable rather than log-transform the original data. Other than using LR and NLR model, many researchers have utilized the data-driven model for hydrological variable estimation at ungauged basins. A significant advantage of a data-driven model is that it can produce reliable estimation accuracy with a few information about the criteria and the behaviour of the physical and hydrological processes (Rahmati et al. 2019; Yang et al. 2020). Another advantage is that the data-driven model has little development. Hailegeorgis and Alfredsen (2017) have researched on the hydrological variable estimation located at mid-Norway.

The total number of basins involved in their research were 26 basins. The LR model is employed to construct the connection between the hydrological variable and basin characteristics. Then, the LR model is applied to estimate the hydrological variable at ungauged basins. The effectiveness of the ANN model for hydrological variable estimation exemplified in a study by Jolankai and Koncsos (2018), where the ANN model outperforms the LR model. Aziz et al. (2017) have made comparisons between the estimation performance of an artificial neural network (ANN) model and the LR model to estimate the hydrological variable at ungauged basins in Australia. The selected hydrological variable for comparison is flood quantile. The results obtained from his research show that the ANN model produces more reliable accuracy compare to the LR model. Meresa (2019) suggests that the ANN model has a reliable accuracy in estimating flood quantile at the ungauged basins. It is the advantage of the ANN model because of its significant attribute in handling nonlinear data (Wu et al. 2016).

Recently, a considerable literature has grown up around the data-driven model, namely Group Method of Data Handling (GMDH) model. The GMDH model was initially developed by Ivakhnenko (1971) for modelling and recognition of complex networks. The GMDH model successfully used in a wide variety of fields, namely image processing, resource management, health, and chemistry (Mehrabani et al. 2020; Mohebbian et al. 2020; Radaideh & Kozlowski 2020; Tournier et al. 2019). The GMDH model is a highly efficient approach to resolve modelling problems that include numerous numbers of

input variables to one output variable. The main idea of the GMDH model involves building a forward feed network by utilizing the partial description (PD) of the GMDH model.

The PD of the GMDH model is a second-order polynomial (PLY), and the coefficients of the PD were obtained using the least square method (LS). Koopialipoor et al. (2019) use GMDH to estimate the penetration rate performance of a tunnel boring machine. The estimation is vital for the future project due to the safety and economic aspect. Based on the study, GMDH produces better accuracy than the LR model for modelling and estimation. The literature of the GMDH model shows the GMDH model has been implemented in various areas of engineering. Despite that, the GMDH model is neglected as a reliable tool for hydrological estimation at the ungauged basin. However, there are some drawbacks to the GMDH model. Most implementations of the GMDH model solely employ a single transfer function (TF) known as PLY TF, and the GMDH model tends to generate a big network before the realization of the system (Fathi et al. 2020; Rezazadeh Eidgahee et al. 2019; Rostami et al. 2019).

Kondo and Ueeno (2009) have introduced the application of radial basis TF (RBF) and sigmoid (SIG) TF in the GMDH model intended for medical image recognition. The results show that the RBF-GMDH model has recognized and extracted the region of abdominal organs accurately. Nurhaziyatul et al. (2019) have implemented various types of TF in the GMDH model, which include SIG, RBF, and PLY TF. The GMDH model is developed separately for each TF. The findings show that each TF produces a different result and as evidence that shows TF affects the performance of the GMDH model. Other shortcomings of the GMDH model lean to generate an extensive polynomial system, although it has a small input variable. The GMDH model also tends to remove the significant variable while sorting out procedure (Shahabi et al. 2016). To overcome the shortcomings of the GMDH model, most of the researchers combine the GMDH model with the genetic algorithm or data-driven model (Farrokhi et al. 2020; Shaghaghi et al. 2017). Hence, in this study, two modifications will be made on the GMDH model to alleviate its weaknesses. The changes include combining the GMDH model with principal component analysis (PCA) and the application of several TF, such as PLY, RBF, SIG, and hyperbolic tangent (HPT) TF. The motivation employs multiple transfers as every data is unique. Most implementations of the conventional GMDH model implement a single TF, namely PLY TF. Therefore, by applying a different type of TF in the GMDH model, it is expected to enhance the estimation performance of the GMDH model. The reason for the combination of PCA is to decrease the complexity of the GMDH model. PCA is a widely known used for dimensionality reduction for multivariate data analysis (Du 2019). The basic concept of PCA is to reduce the size of the data that consists of a lot of variables and, at the same time keeping the utmost variation from the original data set with fewer principal components. The purpose of PCA can be accomplished by converting original data set into a principal component, which are uncorrelated for the first few principal components that can keep most of the variation in the original data (Jolliffe & Cadima 2016). Noori et al. (2010) have carried out the application of the hybrid model, namely the PCA-ANN model, for weekly solid waste forecasting. The outcomes demonstrate that the ANN model yields better estimation by employing seven principal components as input variables rather than employing the original 13 variables as input variables. Prusty et al. (2017) indicate that the PCA could be applied to decrease the dimensionality of the original by discarding the principal component that has a lower variation of the original data. Therefore, this study aims to develop a robust GMDH model by combining the model with PCA and implementing it with four types of TF, which can enhance the estimation performance of the modified GMDH (MGMDH) model for hydrological variable estimation at ungauged basins.

METHODS

DATA

In this study, there were three types of data collected from seventy selected basins throughout West Malaysia. Previous studies conducted by Shu and Ouarda (2008) have established that to obtain a reliable estimate of at-site flood quantile, the minimum length of historical data is 15 years. Therefore, the range of the streamflow data record for 70 selected basins in this current study is varying from 15 to 50 years by utilizing generalized extreme value (GEV) distribution to estimate the flood quantile. Further information and discussion related to the application of the GEV for an extreme event can be found in Firdaus et al. (2019), Hasfazilah et al. (2015), Wan Zawiah et al. (2020) and Wan Zin et al. (2009). The information gathered in this research include: meteorological data – the mean annual total rainfall (TRF), hydrological data

– the streamflow data that fits the GEV distribution to obtain flood quantile with a return period of 10, 50, and 100 years, and physiographical data – consists of three physiographical variables: mean river slope (MRS), longest drainage path (LPH) elevation (ELT), and basin area (BA). Table 1 provides the statistics summary of hydrological, physiographical and meteorological data.

TABLE 1. Summary statistics of the basin characteristics

Variables	Min	Mean	Max	Std. Dev
BA(km^2)	30	1787.05	19000	3676.28
ELT (m)	4	99.49	1450	249.99
LPH (m)	3800	38457.97	280000	59553.88
MRS (%)	0.01	0.40	2.56	0.50
TRF (mm)	314.30	2099.75	4678.70	717.26

LR MODEL

It is widely known that hydrological variable, such as flood quantile, is strongly related to the meteorological and physiographical characteristics. Based on this context, the empirical equations are established to link the flood quantiles with the meteorological and physiographical features. Thus far, the LR model demonstrated as a reliable estimation model for forecasting the flood quantile at the ungauged basin. The power function is as shown herewith,

$$y = \alpha_0 X_1^{\alpha_1} X_2^{\alpha_2} \cdots X_p^{\alpha_n} \varepsilon_0 \tag{1}$$

where α is the model parameter; p is the number of basins characteristic; and X is the characteristics of the basin. Logarithmic transformation can be used to linearise (1), where the parameters are derived by using the LR model.

NLR MODEL

Employing NLR model, the α or model parameters of (1) can be obtained directly without logarithmic transformation *via* minimizing the objective function in the actual domain. According to Seber and Wild (2003), the objective function (2) and the algorithm of the NLR model are used to narrow the gap between the observed and the estimated flow.

$$\eta(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \left[y_i - F(x_i; \boldsymbol{\alpha}) \right]^2 \tag{2}$$

where y_i and $F(x_i;\boldsymbol{\alpha})$ are the observed and estimated flows, respectively. n is the length of the data. The NLR model enhances the model parameters iteratively starting from an initial value. In this paper, the NLR model with Levenberg-Marquardt method is chosen. Gavin (2016) stated that the Levenberg-Marquardt process is a mixture of both the Gauss-Newton method and the gradient descent method which make the Levenberg-Marquardt method more efficient minimization methods.

ANN MODEL

The ANN model has developed similarly to the human brain architecturally as a universal mathematical model. The current study implements a three-layer feed-forward ANN model for forecasting the flood quantile at ungauged basins. Multilayer perceptron with a single hidden layer feed-forward network is the best-known and vastly used in ANN application (Abbas et al. 2019). Data

are inserted to the network on the first layer, also known as the input layer. Next, the data go through a process at the hidden layer and then, the results are produced at the final or output layer. The hidden layer and nodes play significant parts in the ANN implementation. From the previous research, a single hidden layer was successfully used for the estimation of flood quantile at ungauged basins (Aziz et al. 2017; Samantaray & Ghose 2020). Similar to that, a single hidden layer was employed for the ANN model. In order to determine the suitable value of hidden nodes in the hidden layer, various guidelines were referred to; such as, 'x' suggested by Tang and Fishwick (1993), '$2x$' indicated by Wong (1991), and '$2x+1$' suggested by Hecht-Nielsen (1990), where x is the number of inputs. There are two activation functions adopted, namely the sigmoid and linear function.

PCA

The procedure of the PCA method begins with the data on p variable for n number of data. The necessary steps involved in PCA are stated as follows:

Step 1: Standardize the scale of the data by using the log transform (Jolliffe & Cadima 2016).

Step 2: Find the covariance matrix for the standardized data.

Step 3: Using the covariance matrix from step 2, identify the eigenvalues and eigenvector of the covariance matrix. After eigenvalues of the covariance matrix have been obtained, the eigenvectors can be acquired using Gauss Elimination.

Step 4: Establish the appropriate number of PC. The number of PCs retains for analysis must consist of 90% variation of the original data (Comber et al. 2016).

GMDH MODEL

The primary step related to GMDH model is:

Step 1: Determine the number input variables $X = \{x_1, x_2, ..., x_p\}$ and the output variable y where p is defined as the whole inputs for GMDH model. The normalization of the entire data is conducted if necessary.

Step 2: Then, the whole data set is split into two subsets, namely training and estimation data set. The number of data in estimation data set is only one and the remaining data is used for training. The PD is configured from the training data set in the form of a second-order polynomial. The PD is shown in (3).

$$\hat{w}_k = v_0 + v_1 x_i + v_2 x_j + v_3 x_i x_j + v_4 x_i^2 + v_5 x_j^2 \quad (3)$$

The coefficients of PD are estimated using the least square method. The number of PD constructed are L = p (p -1)/2. Therefore, at the current layer GMDH model contains L estimates of $\hat{y}_k$ where $\hat{y}_k = \hat{w}_k$.

Step 3: Identifying the new best variable for the following layer is a very crucial step in GMDH model. In this step, all the output $\hat{y}_k$ will be screen and only the best $\hat{y}_k$ is selected and the remaining $\hat{y}_k$ will be discarded. The best $\hat{y}_k$ is identified by using mean squared error (MSE) on the training data set. The MSE is defined in (4). The best output $\hat{y}_k$ is chosen as the new input variables for the following layer. The input variable for the subsequent layer will become $\{x_1, x_2, ..., x_p, \hat{y}_k\}$. The number of input variables becomes $p = p + 1$.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{i,k})^2 \quad (4)$$

Step 4: Check the stopping condition. Compare the minimum MSE of the current training layer with the MSE of previous training layer to determine whether the set of equations can be improved. The termination of the process will occur if the minimum MSE on the current training layer is higher or equal on the previous training layer. The process (step 2 & 3) is required to repeat when the minimum MSE on current training layer is lower than the last segment. Otherwise, the process is stopped as the realization of the network has achieved.

MGMDH MODEL

The MGMDH model procedure is set up below:

Step 1: Initially, the input variables $X = \{x_1, x_2, ..., x_p\}$ and the output variable y identify. The overall number of input variables is p. Subsequently, the aggregate data is split into a training and estimation data set. In the case of estimation at the ungauged basin, only one data on estimation set as to simulate ungauged location and remaining data that is on training data set. The MGMDH model is set up, and the parameter of PD is obtained from the training data. The normalization of the original data will be carried out if needed.

Step 2: PCA is implemented to the input variable $X = \{x_1, x_2, ..., x_p\}$ for dimensionality reduction. Total variance explained of more than 90% is required for the number of PCs chosen to become the input for the MGMDH model.

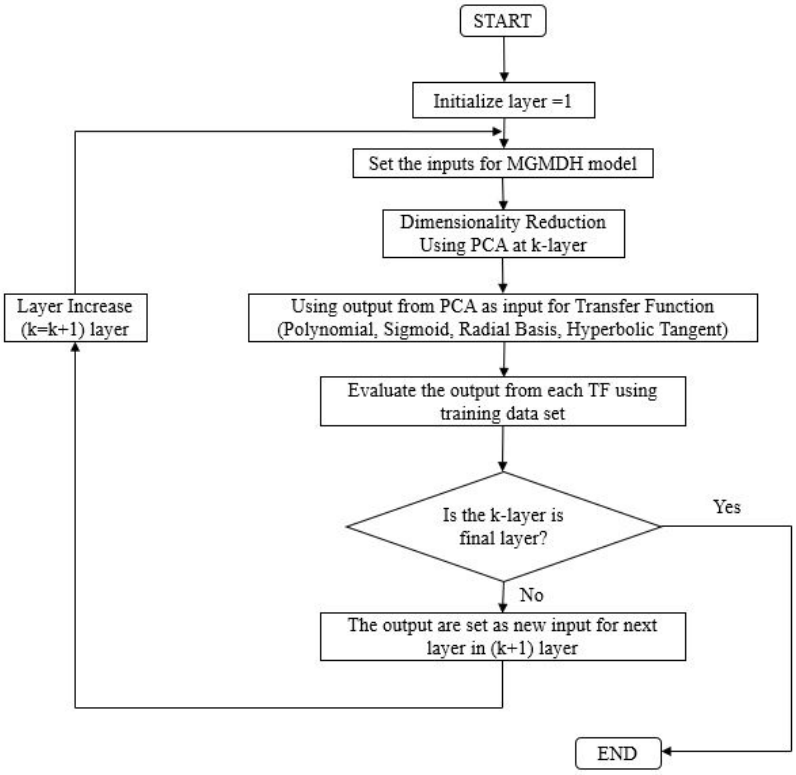

FIGURE 1. Flow chart of MGMDH model

Step 3: After that, the implementation of four TF in MGMDH model into the training set. The four TF are PLY, HPT, SIG, and RBF functions. The type TF employs in MGMDH model is shown in Table 2.

TABLE 2. Type of transfer function

Transfer function	
Polynomial	$y(p)_k = w_k$
Sigmoid	$y(s)_k = 1/(1+\exp(-w_k))$
Radial Basis	$y(rb)_k = \exp(-w_k^2)$
Hyperbolic Tangent	$y(ht)_k = \left(\frac{2}{1+e^{-2w_k}}\right)-1$

*where $\hat{w}_k$ is PD that had been described on (3)

Step 4: In order to estimate the parameter or coefficient of PD for four TF, the LSM is implemented. On a single TF, the amount of PD constructed at the current level is identified by $U = z(z\text{-}1)/2$ where z is the number of PCs preserve. Therefore, for four TF, the number of PD created is $4U$. The set-off linear equations system are:

$$\mathbf{Y} = \mathbf{Gv} \tag{5}$$

Equation 5 is illustrated as follow;

$$\mathbf{Y} = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{bmatrix} \boldsymbol{G} = \begin{bmatrix} 1 & Z_{1i} & Z_{1j} & Z_{1i}Z_{ij} & Z_{1i}^2 & Z_{1j}^2 \\ 1 & Z_{2i} & Z_j & Z_{2i}Z_{2j} & Z_{2i}^2 & Z_{2j}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & Z_{li} & Z_{lj} & Z_{li}Z_{lj} & Z_{li}^2 & Z_{lj}^2 \end{bmatrix} \mathbf{v} = \begin{bmatrix} v_0 \\ v_1 \\ \vdots \\ v_s \end{bmatrix}$$

The transfer functions for MGMDH model are described in Table 3.

TABLE 3. S_i value definition

Transfer function	
Polynomial	$s_i = y_i$
Sigmoid	$s_i = \ln\left(\frac{y_i}{1-y_i}\right)$
Radial Basis	$s_i = \sqrt{-\ln y_i}$
Hyperbolic Tangent	$s_i = -\frac{1}{2}\ln\left(\frac{2}{y_i+1}-1\right)$

The parameters and coefficient of each PD in every layer obtained using the LSM. The matrix form is shown in (6).

$$v_i = (G_i^T G_i)^{-1} G_i Y \tag{6}$$

This step repeatedly implemented for all partial description in the current layer of MGMDH starting from the input layer until the output layer.

Step 5: After completing the previous process, only one output from the PD to be selected to become a new input for the following layer of MGMDH model. The measure to choose the best PD is based on the MSE. Only PD output that produces the minimum MSE chosen as the new input for the next layer. Another variable will be eliminated before proceeding to the next layer. It should be emphasized after the selection of the new input variable, step 1 until step 5 are repeated until the realization of the network is achieved. The term realization of the network is achieved when the stopping criteria are reached. The stopping criteria are reached

when the minimum MSE of the current has increased or no improvement. Then, the process is terminated. The output from the previous layer with the lowest MSE is selected as the output from MGMDH model. Figures 2 and 3 illustrate the architecture of MGMDH.

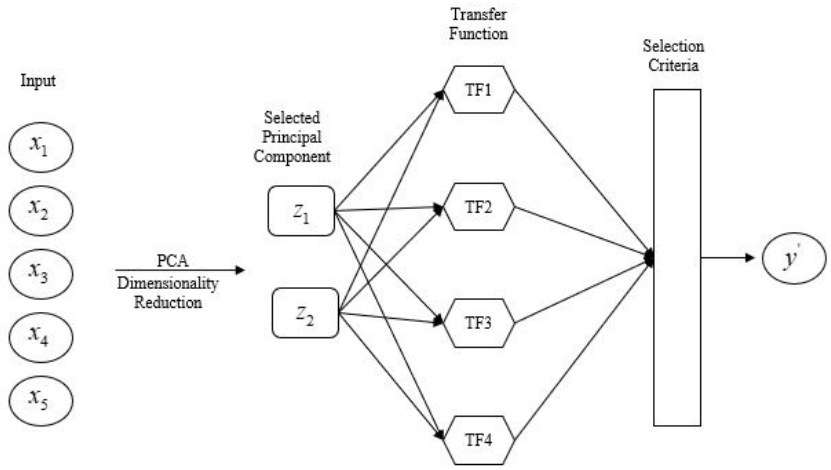

FIGURE 2. First layer of MGMDH model

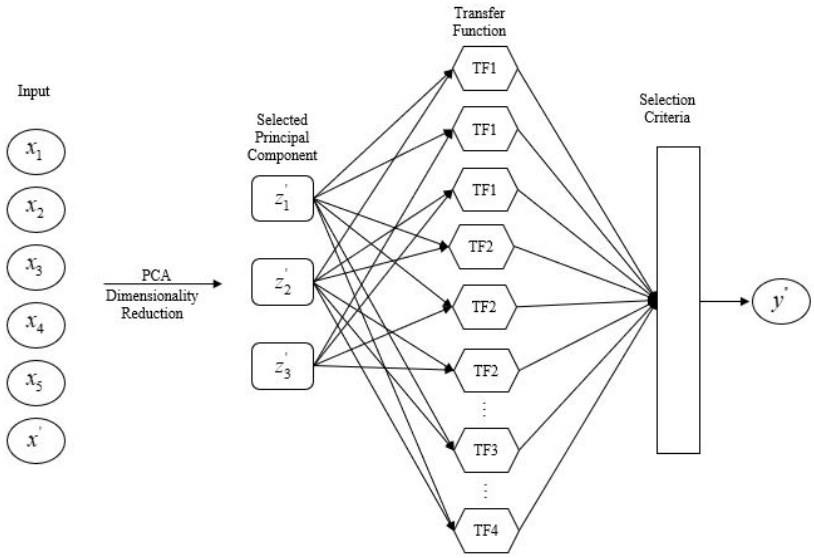

FIGURE 3. Second layer of MGMDH model

EVALUATION OF MODEL ESTIMATION PERFORMANCE

The estimation performance of the MGMDH and comparison model is evaluated using the error indicator, namely, the Mean Absolute Percentage Error (MAPE) and the Nash-Sutcliffe efficiency (NSE). The interpretations of MAPE and NSE are provided in (7) and (8), respectively.

$$NSE = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{7}$$

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

where $\hat{y}_i$ is the estimated flow; y_i is the observed flow; $\bar{y}_i$ is the mean of the observed flow; and n is the number of flow series that have been model. The NSE ranges from $-\infty$ (worst case scenario) to 1 (perfect fit) indicates the effectiveness of a model at estimating the observed values. Meanwhile, a negative efficiency (less than zero) shows the mean value of the observed flow is a more efficient estimator than the estimation model. The MAPE of a model describes how well the model can estimate. MAPE expresses the error in percentage value which makes it easy for the researchers to make the comparison for estimation performance with other models. Typically, the model with the best (lowest) MAPE should be used for estimation.

RESULTS AND DISCUSSION

One of the major drawback in the GMDH model, the complexity of the networks keeps increasing when GMDH layer is increased. In this study, the PCA method is proposed to decrease the size of the network of GMDH model. Therefore, a simulation study conducted to measure how much the PCA method can decrease the structural and computational complexity of GMDH model. The structural complexity is measure by the total number of PD constructed and the computational complexity is measure by the time needed to achieve the realization of the network. The idea behind choosing PCA is to reduce the number of inputs for GMDH model due to the size of GMDH network is dependent on the number of data. The result of the simulation study of structural complexity and computational complexity is illustrated in Figures 4 and 5, respectively.

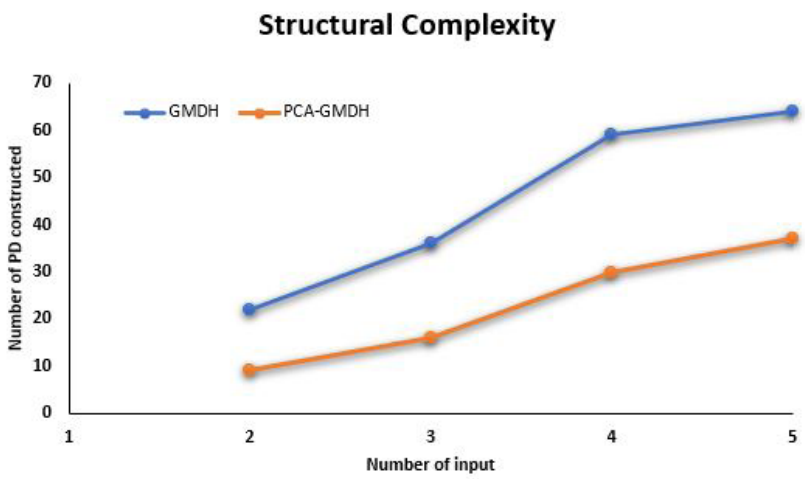

FIGURE 4. Structural complexity

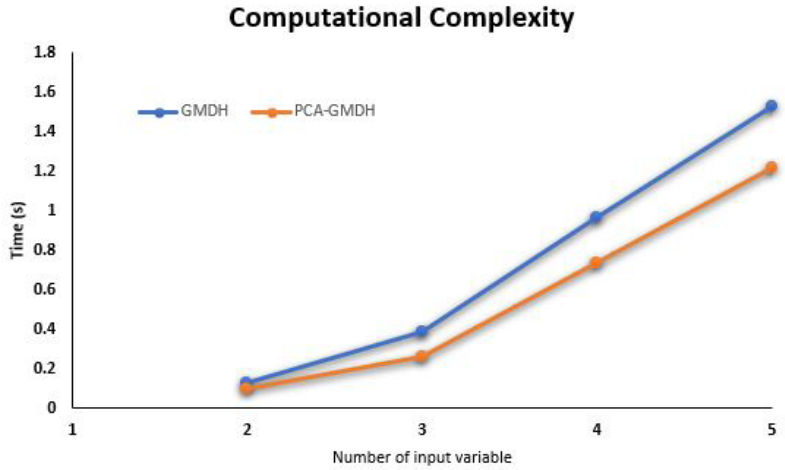

FIGURE 5. Computational complexity

As shown in Figure 4, the number of PD constructed is increasing for both GMDH and MGMDH model when the number of input variable for the model increases. Then on Figure 5, it shows that when the number of PD increases, the computational complexity will also increase. However, with the implementation of PCA method into GMDH model has reduced the number of PD constructed and the computational burden compared to the original GMDH model. On average, the PCA method has reduced the structural complexity of GMDH model by 59% for two input variable, 56% for three input variables, 49% for four input variable and 42% for five input variables, respectively. Other than that, PCA also increased the speed of convergence of the GMDH model. Therefore, based on this simulation study, it can be concluded that PCA has reduced the complexity of GMDH model by an average 52% for structural complexity and 28% for computational complexity.

In assessing the estimation model performance for hydrological estimation at the ungauged basin, jackknife approach is applied to imitate the ungauged basins. This approach involving removing one basin from the data set where it assumes as an ungauged basin, and the remaining basin in the data set was used for model development to obtain the parameter of the estimation model. The procedure is repeated until all the basins were discarded at least once. Therefore, the number of models constructed was equivalent to the number of basins used in this study. The return period adopted in this study is 10 years, 50 years and 100 years as they include both low and high part of the distribution. Selecting input variables for the estimation model is essential as it may affect the model performance in estimating hydrological variable at the target basin. Consequently, a statistical method of the forward selection stepwise regression (SW) was utilized to select the input variables. According to Bowden et al. (2005), the SW was employed to circumvent the whole subset consideration of input variables. The analysis of the SW was done separately for each hydrological variable used in this study. The chosen input variables established on the forward range of stepwise regression are shown in Table 4.

TABLE 4. Best estimator based on stepwise regression

Inputs	Best estimator		
	Q_{10}	Q_{50}	Q_{100}
2	ELT, LPH	ELT, LPH	ELT, LPH
3	ELT, LPH, MRS	ELT, LPH, MRS	ELT, LPH, MRS
4	ELT, LPH, MRS, TRF	ELT, LPH, MRS, TRF	ELT, LPH, MRS, TRF
5	BA, ELT, LPH, MRS, TRF	BA, ELT, LPH, MRS, TRF	BA, ELT, LPH, MRS, TRF

Table 4 presents the best input variables for two, three, four, and five input variables obtain from stepwise regression. The input variables were implemented in the estimation model, then the best input variables that produced the best output for each estimation model based on performance criteria were selected for comparisons. The estimation performance results are presented in Table 5. NSE and MAPE were employed to evaluate the estimation performance of MGMDH and comparison model. Separate models were constructed for Q_{10}, Q_{50} and Q_{100} years. The estimation performance results are presented in Table 5.

TABLE 5. NSE and MAPE for MGMDH and comparison model

	NSE			MAPE		
Model	Q_{10}	Q_{50}	Q_{100}	Q_{10}	Q_{50}	Q_{100}
NLR	0.7568	0.7183	0.6892	69.16	64.85	76.14
ANN	0.8632	0.8424	0.8153	32.82	35.01	28.04
LR	0.7090	0.6819	0.6379	79.83	73.52	80.25
GMDH	0.7868	0.7411	0.8424	43.88	49.71	39.58
MGMDH	**0.8956**	**0.8612**	**0.8337**	**31.29**	**32.63**	**21.67**

As shown in Table 5, the results have demonstrated that the performance of GMDH model was significantly improved when integrated with PCA and the four types of the transfer function. The bold font indicates the best performing model for each error measurement. A model that has a perfect estimation produces NSE value of 1. Typically, an accurate model has a NSE value of greater than 0.8. The results, as shown in Table 5, indicate that only the ANN and MGMDH model produced NSE value more than 0.8 for the estimation of the three specific quantiles. These results suggested that the ANN and MGMDH model can provide a satisfactory forecast. Another important finding was that MGMDH model produce NSE value more closed to 1 compared to the ANN model for the estimation of the three specific quantiles. Based on MAPE, the MGMDH model produced smaller MAPE compared to GMDH model for the estimation of the three particular quantiles, which indicated that MGMDH model has better estimation accuracy than the GMDH model. Therefore, the modification done on GMDH model had improved the GMDH model estimation performance. Based on the performance criteria of MAPE, and NSE, it appeared that MGMDH model yielded better accuracy among other models for the estimation of the three specific quantiles in the ungauged site. Although ANN model produced better accuracy in estimation compared to LR, NLR, and GMDH model, there is a drawback in the ANN model. The disadvantage in the ANN model is to determine the suitable structure for ANN model for specific data. For flood quantile estimation with T=10 years, the best ANN structure is 3-6-1 (input-hidden layer-output), for flood quantile estimation with T=50 years, the best ANN structure is 5-10-1 and for T=100 years 5-5-1. Therefore, an ANN model needs to use different structures for the estimation of the three different specific quantiles. Until now, there is no guideline to identify the suitable structure of the ANN model for particular data. The only way is by trying out one by one of the ANN structure, and the best ANN structure that produces the best estimation is selected.

Therefore, in the real-life problem, it becomes a significant problem to determine which structure is the best for the ANN model for specific data. As a result, the uncertainty in ANN estimation is high. Based on the simulation, PCA had refined the complexity of GMDH model by an average 52% for structural complexity and 28% for computational complexity. MAPE, the MGMDH model produced smaller MAPE compare to GMDH model for the estimation of the three specific quantiles, which indicated that MGMDH model has better estimation accuracy than the GMDH model. Therefore, the modification done on GMDH model enhanced the GMDH model estimation performance. Based on the performance criteria of MAPE and NSE, it appears that MGMDH model yields better accuracy among other models for the estimation of the three specific quantiles in the ungauged site. Hence, it can be concluded that the implementation of PCA and four transfer function able to enhance the performance of GMDH model. The proposed MGMDH model also had better estimation accuracy in comparisons with LR, NLR and ANN model.

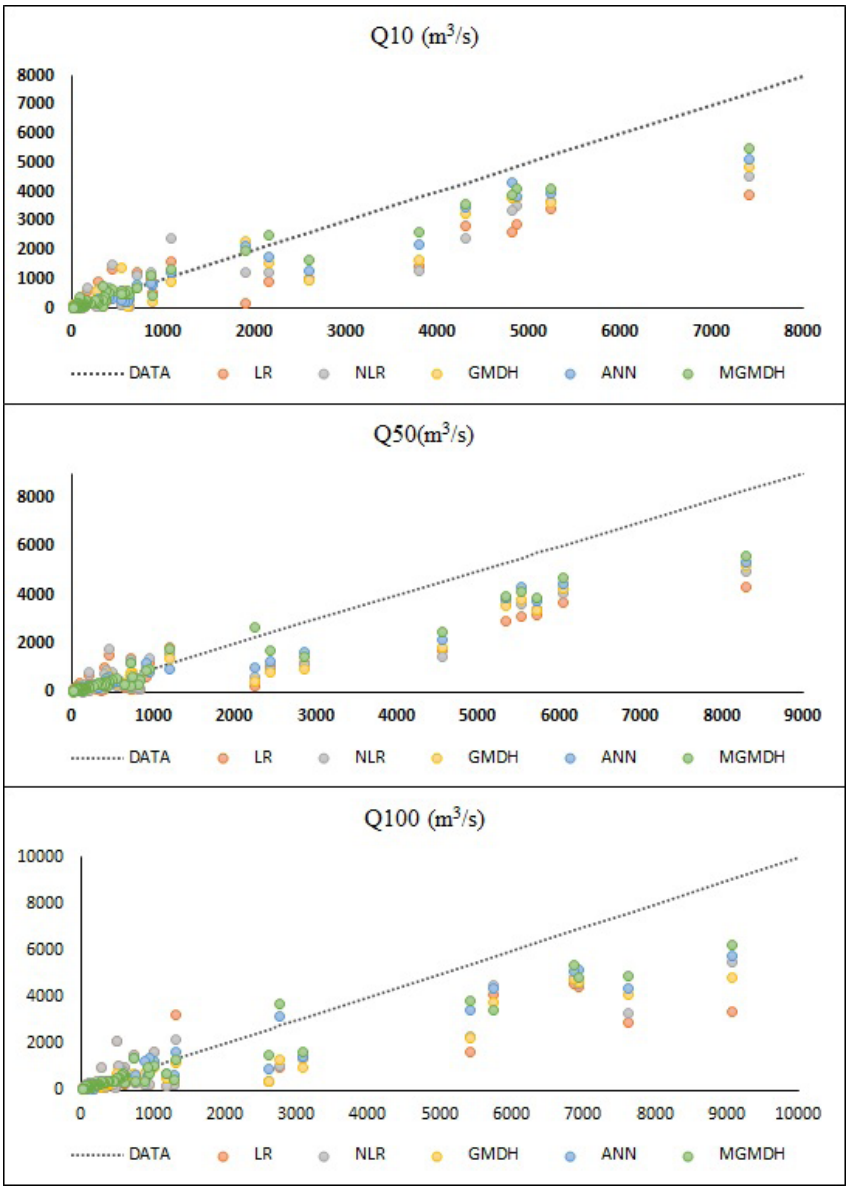

FIGURE 6. Comparison of the observed and estimated flood quantiles using estimation model for Q_{10}, Q_{50} and Q_{100}

Figure 6 is a quantile-quantile plot (QQ-plot) for observed and estimated flood quantile obtained for the prediction model. The 45 degree line represent the observed flood quantile and the point represent the estimated flood quantile obtained from each prediction model respectively. From Figure 6, it can be observed that the estimation error tend to increase with the return period. Then, all models under estimates the flood quantile at basin with higher value for all return period with value more than 4000 m^3/s. The MGMDH model provide a better estimation compare to other model with flood quantiles less than 2000 m^3/s for all return period.

CONCLUSION

This study has achieved its aim of presenting a novel combination of the estimation model that was proposed based on GMDH model for hydrological estimation at the ungauged basins. In enhancing the estimation performance of GMDH model, four types of transfer functions were introduced in the model as well as incorporating PCA to alleviate the complexity of the GMDH model. The four transfer functions include polynomial, hyperbolic tangent, sigmoid, and radial basis functions. The experiments based on real data were conducted for verification of the proposed model. The findings from this study showed that the importance of the proposed MGMDH model in flood quantile estimation at the ungauged basin, in which it has successfully dereased the structural and computational complexity of GMDH model. Besides, the implementation of four transfer functions in MGMDH model has improved the estimation accuracy and efficiency of GMDH model. According to past studies, the ANN model was the best method for flood quantile estimation at the ungauged site. However, from this study, the findings computed that the MGMDH model outperformed LR, NLR, GMDH, and ANN model in the estimation accuracy of flood quantiles at ungauged basin. Therefore, the experimental results indicated that MGMDH is robust and efficient for flood quantile estimation at the ungauged basin. Consequently, serves as a useful tool for application in flood quantile estimation at the ungauged site.

ACKNOWLEDGEMENTS

We would like to thank the Drainage and Irrigation (DID), Malaysia for providing us data used in this study.

REFERENCES

Abbas, A.K., Al-haideri, N.A. & Bashikh, A.A. 2019. Implementing artificial neural networks and support vector machines to predict lost circulation. *Egyptian Journal of Petroleum* 28(4): 339-347.

Abdullah A. Mamun, Alias Hashim & Zalin Amir. 2012. Regional statistical models for the estimation of flood peak values at ungauged catchments: Peninsular Malaysia. *Journal of Hydrologic Engineering* 17(4): 547-553.

Alobaidi, M.H., Marpu, P.R., Ouarda, T.B.M.J. & Chebana, F. 2015. Regional frequency analysis at ungauged sites using a two-stage resampling generalized ensemble framework. *Advances in Water Resources* 84: 103-111.

Arsenault, R., Breton-Dufour, M., Poulin, A., Dallaire, G. & Romero-Lopez, R. 2019. Streamflow prediction in ungauged basins: Analysis of regionalization methods in a hydrologically heterogeneous region of Mexico. *Hydrological Sciences Journal* 64(11): 1297-1311.

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.

Bowden, G.J., Dandy, G.C. & Maier, H.R. 2005. Input determination for neural network models in water resources applications. Part 1 - Background and methodology. *Journal of Hydrology* 301(1-4): 75-92.

Comber, A.J., Harris, P. & Tsutsumida, N. 2016. Improving land cover classification using input variables derived from a geographically weighted principal components analysis. *ISPRS Journal of Photogrammetry and Remote Sensing* 119: 347-360.

Du, T.Y. 2019. Dimensionality reduction techniques for visualizing morphometric data: Comparing principal component analysis to nonlinear methods. *Evolutionary Biology* 46(1): 106-121.

Farrokhi, F., Firoozfar, A. & Maghsoudi, M.S. 2020. Evaluation of liquefaction-induced lateral displacement using a GMDH-type neural network optimized by genetic algorithm. *Arabian Journal of Geosciences* 13(1): 4.

Fathi, S., Eftekhari Yazdi, M. & Adamian, A. 2020. Estimation of contact heat transfer between curvilinear contacts using inverse method and group method of data handling (GMDH)-type neural networks. *Heat and Mass Transfer* 56: 1961-1970.

Firdaus Mohamad Hamzah, Siti Hawa Mohd Yusoff & Othman Jaafar. 2019. L-moment-based frequency analysis of high-flow at Sungai Langat, Kajang, Selangor, Malaysia. *Sains Malaysiana* 48(7): 1357-1366.

Hailegeorgis, T.T. & Alfredsen, K. 2017. Regional flood frequency analysis and prediction in ungauged basins including estimation of major uncertainties for Mid-Norway. *Journal of Hydrology: Regional Studies* 9: 104-126.

Hasfazilah Ahmat, Ahmad Shukri Yahaya & Nor Azam Ramli. 2015. PM_{10} analysis for three industrialized areas using extreme value. *Sains Malaysiana* 44(2): 175-185.

Hecht-Nielsen, R. 1990. *Neurocomputing.* Reading, MA: Addison-Wesley.

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

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.

Jolliffe, I.T. & Cadima, J. 2016. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374(2065): 20150202.

Kim, N.W., Lee, J-Y., Park, D-H. & Kim, T-W. 2019. Evaluation of future flood risk according to RCP scenarios using a regional flood frequency analysis for ungauged watersheds. *Water* 11(5): 992.

Kondo, T. & Ueno, J. 2009. Medical image recognition of abdominal multi-organs by RBF GMDH-type neural network. *International Journal of Innovative Computing Information and Control* 5(1): 225-240.

Koopialipoor, M., Nikouei, S.S., Marto, A., Fahimifar, A., Armaghani, D.J. & Mohamad, E.T. 2019. Predicting tunnel boring machine performance through a new model based on the group method of data handling. *Bulletin of Engineering Geology and the Environment* 78(5): 3799-3813.

Mehrabani, M.N., Golafshani, E.M. & Ravanshadnia, M. 2020. Scoring of tenders in construction projects using group method of data handling. *KSCE Journal of Civil Engineering* 24: 1996-2008.

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.

Mohebbian, M.R., Dinh, A., Wahid, K. & Alam, M.S. 2020. Blind, cuff-less, calibration-free and continuous blood pressure estimation using optimized inductive group method of data handling. *Biomedical Signal Processing and Control* 57: 101682.

Noori, R., Karbassi, A. & Sabahi, M.S. 2010. Evaluation of PCA and gamma test techniques on ANN operation for weekly solid waste prediction. *Journal of Environmental Management* 91(3): 767-771.

Nurhaziyatul A. Yahya, Ruhaidah Samsudin, Ani Shabri & Faisal Saeed. 2019. Combined group method of data handling models using artificial bee colony algorithm in time series forecasting. *Procedia Computer Science* 163: 319-329.

Ojha, R. & Tripathi, S. 2018. Using attributes of ungauged basins to improve regional regression equations for flood estimation: A deep learning approach. *ISH Journal of Hydraulic Engineering* 24(2): 239-248.

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.

Prusty, M.R., Jayanthi, T., Chakraborty, J. & Velusamy, K. 2017. Feasibility of ANFIS towards multiclass event classification in PFBR considering dimensionality reduction using PCA. *Annals of Nuclear Energy* 99: 311-320.

Radaideh, M.I. & Kozlowski, T. 2020. Analyzing nuclear reactor simulation data and uncertainty with the group method of data handling. *Nuclear Engineering and Technology* 52(2): 287-295.

Rahmati, O. & Pourghasemi, H.R. 2017. Identification of critical flood prone areas in data-scarce and ungauged regions: A comparison of three data mining models. *Water Resources Management* 31(5): 1473-1487.

Rezazadeh Eidgahee, D., Haddad, A. & Naderpour, H. 2019. Evaluation of shear strength parameters of granulated waste rubber using artificial neural networks and group method of data handling. *Scientiairanica* 26(6): 3233-3244.

Rostami, A., Hemmati-Sarapardeh, A., Karkevandi-Talkhooncheh, A., Husein, M.M., Shamshirband, S. & Rabczuk, T. 2019. Modeling heat capacity of ionic liquids using group method of data handling: A hybrid and structure-based approach. *International Journal of Heat and Mass Transfer* 129: 7-17.

Samantaray, S. & Ghose, D.K. 2020. Modelling runoff in a River Basin, India: An integration for developing un-gauged catchment. *International Journal of Hydrology Science and Technology* 10(3): 248-266.

Seber, G.A.F. & Wild, C.J. 2003. *Nonlinear Regression.* Hoboken, New Jersey: John Wiley & Sons. p. 63.

Shaghaghi, S., Bonakdari, H., Gholami, A., Ebtehaj, I. & Zeinolabedini, M. 2017. Comparative analysis of GMDH neural network based on genetic algorithm and particle swarm optimization in stable channel design. *Applied Mathematics and Computation* 313: 271-286.

Shahabi, S., Mohammad-Javad, K. & Kermani, M.H. 2016. Hybrid wavelet-GMDH model to forecast significant wave height. *Water Science and Technology: Water Supply* 16(2): 453-459.

Shu, C. & Ouarda, T.B.M.J. 2008. Regional flood frequency analysis at ungauged sites using the adaptive neuro-fuzzy inference system. *Journal of Hydrology* 349(1-2): 31-43.

Sivapalan, M., Takeuchi, K., Franks, S.W., Gupta, V.K., Karambiri, H., Lakshmi, V., Liang, X., Mcdonnell, J.J., Mendiondo, E.M., O'Connell, P.E., Oki, T., Pomeroy, J.W.,

Schertzer, D., Uhlenbrook, S. & Zehe, E. 2003. IAHS decade on predictions in ungauged basins (PUB), 2003-2012: Shaping an exciting future for the hydrological sciences. *Hydrological Sciences Journal* 48(6): 857-880.

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

Tournier, J-D., Smith, R., Raffelt, D., Tabbara, R., Dhollander, T., Pietsch, M., Christiaens, D., Jeurissen, B., Yeh, C-H. & Connelly, A. 2019. MRtrix3: A fast, flexible and open software framework for medical image processing and visualisation. *NeuroImage* 202: 116137.

Tsegaw, A.T., Alfredsen, K., Skaugen, T. & Muthanna, T.M. 2019. Predicting hourly flows at ungauged small rural catchments using a parsimonious hydrological model. *Journal of Hydrology* 573: 855-871.

Wałęga, A., Młyński, D., Wojkowski, J., Radecki-Pawlik, A. & Lepeška, T. 2020. New empirical model using landscape hydric potential method to estimate median peak discharges in mountain ungauged catchments. *Water* 12(4): 983.

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.

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.

Yang, S., Wang, P., Lou, H., Wang, J., Zhao, C. & Gong, T. 2019. Estimating river discharges in ungauged catchments using the slope-area method and unmanned aerial vehicle. *Water* 11(11): 2361.

Wan Zawiah Wan Zin, Abdul Aziz Jemain, Marina Zahari & Kamarulzaman Ibrahim. 2020. Scaling analysis for extreme rainfall events in Peninsular Malaysia. *Sains Malaysiana* 49(10): 2573-2585.

Wan Zin Wan Zawiah, 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.

Basri Badyalina*
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
Cawangan Johor, Kampus Segamat
85000 Segamat, Johor Darul Takzim
Malaysia

Ani Shabri & Muhammad Fadhil Marsani
Department of Mathematics, Faculty of Science
Universiti Teknologi Malaysia
81310 Skudai, Johor Darul Takzim
Malaysia

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

Received: 24 August 2020
Accepted: 17 January 2021