International Journal of Statistics and Probability; Vol. 10, No. 6; November 2021
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Modified Group Method of Data Handling for Flood Quantile Prediction at Ungauged Site

Basri Badyalina[1], Ani Shabri[2], Nurkhairany Amyra Mokhtar[1], Mohamad Faizal Ramli[3], Muhammad Majid [3] & Muhammad Yassar Yusri[1]

[1] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Cawangan Johor, Kampus Segamat, Johor, Malaysia

[2] Department of Mathematics, Faculty of Science, Universiti Teknologi Malaysia 81310 Skudai, Johor Darul Takzim Malaysia

[3] Faculty of Business and Management, Universiti Teknologi MARA, Cawangan Johor, Kampus Segamat, Johor, Malaysia

Correspondence: Basri Badyalina, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Cawangan Johor, Kampus Segamat, Johor, Malaysia

Received: October 1, 2021 Accepted: October 28, 2021 Online Published: October 29, 2021

doi:10.5539/ijsp.v10n6p57 URL: https://doi.org/10.5539/ijsp.v10n6p57

Abstract

Handling flood quantile with little data is essential in managing water resources. In this paper, we propose a potential model called Modified Group Method of Data Handling (MGMDH) to predict the flood quantile at ungauged sites in Malaysia. In this proposed MGMDH model, the principal component analysis (PCA) method is matched to the group method of data handling (GMDH) with various transfer functions. The MGMDH model consists of four transfer functions: polynomial, sigmoid, radial basis function, and hyperbolic tangent sigmoid transfer functions. The prediction performance of MGMDH models is compared to the conventional GMDH model. The appropriateness and effectiveness of the proposed models are demonstrated with a simulation study. Cauchy distribution is used in the simulation study as a disturbance error. The implementation of Cauchy Distribution as an error disturbance in artificial data illustrates the performance of the proposed models if the extreme value or extreme event occurs in the data set. The simulation study may say that the MGMDH model is superior to other comparison models, namely LR, NLR, GMDH and ANN models. Another beauty of this proposed model is that it shows a strong prediction performance when multicollinearity is absent in the data set.

Keywords: ungauged, prediction, simulation, simulation, MAPE

1. Introduction

Prediction in the ungauged station has become a challenging topic in hydrological problems (Grimaldi et al., 2021). Especially in Malaysia, most gauged stations are only located at a strategic location or developing area. Based on Sivapalan et al. (2013), the definition of ungauged is that hydrological data is not available or partially available. Therefore, there are insufficient data to test the hydrological model's actual capability to predict the flood quantile at ungauged stations. Usually, for ungauged problems, the regionalization approach is the most common approach used in ungauged situations. The regionalization approach includes transferring information from gauged stations to the ungauged station (Guo et al., 2021; Desai et al., 2021; Golian et al., 2021).

The regionalization approach has been applied in Peninsular Malaysia to estimate the flood quantile ungauged station in Malaysia (Nazirah et al. 2021; Razaq et al., 2016; Badyalina et al., 2016; Mamun et al., 2012). The regionalization approach used in the case study of Malaysia involved fitting a probability distribution to streamflow at the gauged station and building a relationship with the catchment characteristics. Five distributions are the most commonly used in Peninsular Malaysia, namely generalized extreme value distribution, generalized Pareto distributions, generalized logistics distributions, lognormal distributions and Pearson 3 distributions (Romali et al., 2017; Jan and Shabri, 2016; Jan et al., 2016; Badyalina et al., 2013). Then the parameter of the distribution can be estimated using the L-Moment method or maximum likelihood estimation (Mokhtar et al., 2021; Haddad (2021); Sahu et al., 2021; Mokhtar et al., 2021; Ming et al., 2021; Jan et al., 2018). After obtaining the distribution parameter, the best-fitted distribution will be determined, and flood quantiles will be calculated using the best distribution for each station. Hydrological modelling is employed to model the relationship between catchment characteristics and flood quantile for each gauged station. The

information hydrological information from gauged stations is transferred to the ungauged station using a hydrological model. Although the hydrological data is not available at the ungauged station, the physiographical and metrological data are available. A previous study by Badyalina et al. (2021) stated that the best hydrological model for flood quantile prediction at ungauged stations is the modified group method of data handling (MGDMH) model. The MGMDH model consists of a combination of principal component analysis (PCA) and group method of data handling model (GMDH). Other than that, the MGMDH model employs four different transfer functions in a single model, which makes the model more robust for flood quantile prediction at the ungauged stations. PCA approach gives a significant boost to the GMDH model, which can enhance the performance of the GMDH model.

Usually, for multivariate problems, multicollinearity exists in the data set, making the prediction performance of a particular hydrological model performance poor. Applying PCA in the data set removes the multicollinearity in the data set and is expected to improve the prediction performance of the hydrological model (Barth et al., 2021; Al-Ashkar et al., 2021). The study done by Badyalina et al. (2021) showed that the combination of PCA and various transfer functions outperformed other machine learning models such as the Artificial Neural Network (ANN) model. In that study, the MGMDH model is better than the ANN model, GMDH model, nonlinear regression model and multiple linear regression model for flood quantile prediction at the ungauged site.

This study focuses on simulation studies by generating artificial data to test the accuracy performance of the MGMDH model. The artificial data generated in this study mimic the characteristics of the real data use in the Badyalina et al. (2021) study. The features of the real data pose multicollinearity, and extreme value is present in the data. The simulation study is a continuation of previous to support the findings of Badyalina et al. (2021). The simulation study wants to prove that the MGMDH model can perform very well for flood quantile prediction at ungauged sites with multicollinearity and extreme value present in the data. The simulation study employs five different models: MGMDH model, GMDH model, ANN model, nonlinear regression, and multiple linear regression. Multiple linear regression and the ANN model are the most common model for flood quantile prediction at the ungauged site (Alobaidi et al., 2021; Campos et al., 2021; Desai and Ouarda, 2021) . Other than that, in this study, three conditions will be configured: sample size, multicollinearity level, and outlier percentage. The simulation data is generated using Lawrence and Arthur (2019).

2. Methodology

2.1 Artificial Data Generation

Simulation studies are carried out for the purpose of generating data that are used to evaluate the proposed model MGMDH model. The data generation technique of Lawrence and Arthur (2019) is used in this study. The design of this experiment involves generating data for the following multivariate stochastic model.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \quad (1)$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \quad (2)$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i \quad (3)$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \varepsilon_i \quad (4)$$

Using Eq. 1 – Eq. 5, four types of the dataset can be generated consisting of 2 input variables, three input variables, four input variables and five input variables. In this section, the sample sizes generated are 30, 50, 70 and 100. The data generation was performed using $\beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 1$. The explanatory variables were generated as below,

$$x_{ij} = (1 - \rho^2) z_{ij} + \rho z_{ij} \ , \qquad i = 1, \dots n; j = 1,2,3,4,5 \quad (5)$$

where z_{ij} are independent standard normal random variables, $N(0,1)$. The value of ρ^2 represents the correlation between explanatory variables. The chosen values were $\rho^2 = 0.0$, $\rho^2 = 0.5$ and $\rho^2 = 0.99$, which represent no correlation, medium and high correlation between the explanatory variables. The final factor was the disturbance distribution. Cauchy distribution has been chosen as the disturbance distribution in these simulation studies based on residual fitting of real data. In order to broaden our case study, the normal distribution is also used as disturbance distribution in this study because Cauchy is non-normal distribution. Other than that, the use of normal distribution is to assess the performance of the model without the presence of a larger value in the data. The percentages of outliers used are 5% and 10%. The design of the outlier follows the design suggested by Rana *et al.* (2012) and Midi and Zahari (2012). The inverse CDF of Normal Distribution (P) will be generated with uniform distribution, which is $U(0,1)$. The Cauchy random number is generated from the inverse cumulative distribution function of the Cauchy Distribution (Yao and Liu, 1996). The inverse cumulative distribution function of the Cauchy Distribution is

$$x^{\circ} = tan\left(\pi\left(P - \frac{1}{2}\right)\right) \tag{6}$$

Where P is a random number that uniformly distributed is $U(0,1)$ and x° the random number generated from inverse CDF of Cauchy distributions. The method for artificial data generation technique for ungauged problems has been discussed in Badyalina et al. (2021).

2.2 Modified Group Method of Data Handling

Modified Group Method of Data Handling (MGMDH) model is an improvement of the GMDH model established by Zadeh et al. (2002). PCA method is applied to reduce the complexity of the GMDH model. The shortcoming in the GMDH model is that it tends to produce a complex polynomial network despite having reasonably simple input data for the network. Onwubolu (2009) stated that the GMDH model's complexity increases at each training stage and the selection of a new layer because of the addition of new input variables. Due to the addition of input variables at each layer GMDH model, the number of PD also constructed increases, and the complexity of the GMDH model also increases. PCA method is most commonly employed in the dimensionality reduction of datasets, which can reduce the GMDH model's complexity where the GMDH network's complexity depends on the number of inputs. The input of the GMDH model will be converted into the principal component (PC) using the PCA method, and the least essential PC will be discarded according to rules that have been set in this study. Only the significant PC can become the input variable for the GMDH model. The input variables are treated using a principal component analysis (PCA) method for dimensionality reduction. The PCA method produces principal components equal tu the number of input variables. The number of principal components selected must have more than 90% total variance explained. The selected principal components become the input for the MGMDH model. The number of selected principal components is defined as w, as described below.

$$\begin{aligned} Z_1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p \\ Z_2 &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p \\ Z_3 &= a_{31}x_1 + a_{32}x_2 + \cdots + a_{3p}x_p \\ Z_w &= a_{w1}x_1 + a_{w2}x_2 + \cdots + a_{wp}x_p \\ &\vdots \\ Z_p &= a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pp}x_p \end{aligned} \tag{7}$$

The first principal component is required to have the largest variance. The second component must be orthogonal to the first component while capturing the largest variance within the data set in that direction. More generally, if a set of principal components has more than two, the first few PCs will have most of the variation compared to the last PCs, which have the least variation. Other than reducing the complexity of the GMDH model, the PCA method also removes the multicollinearity in the dataset, which is also one of the problems in the GMDH model established by Zadeh et al. (2002). GMDH model established by Zadeh (2002) used only a single transfer function that is quadratic polynomials as the transfer function. In the proposed MGMDH model, four types of the transfer function are employed in this model, namely polynomial, sigmoid, radial basis, and hyperbolic tangent. The polynomial transfer is the same as with the GMDH model established by Zadeh et al. (2002). Sigmoid and radial basis transfer functions are introduced by Kondo and Euno (2002, 2006). One new addition of the transfer function used in the MGMDH model is the hyperbolic tangent transfer function. There are four types of transfer functions employed in the MGMDH model: polynomial, sigmoid, radial basis, and hyperbolic tangent. Quadratic or second-order polynomials can make a good transfer function for linking the input and output of the GMDH model. According to Kordik (2009), a model with mixed transfer functions usually has better performance than a model that uses a single transfer function because each data set is unique. The MGMDH model can auto-select the suitable transfer function for each data set at the selection process in the MGMDH model. The transfer function is shown in Table 1.

Table 1. Types of transfer function employ on MGMDH model

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

*where w_k is a partial description for the MGMDH model

The partial description (PD) of MGHDH model which comes in the form of a quadratic polynomial which defined by

Eq. 8.

$$\hat{w}_k = v_0 + v_1 x_i + v_2 x_j + v_3 x_i x_j + v_4 x_i^2 + v_5 x_j^2 \quad (8)$$

The coefficients of PD $v_0, v_1, v_2, v_3, v_4 \; and \; v_5$ are estimated using the least square method.

$$v_i = (G_i^T G_i)^{-1} G_i Y \quad (9)$$

where G is a set of the selected principal component and Y is a set of the target variable. Each layer of MGMDH layer, its produce $U = 4(p(p-1)/2))$ number of PD. In order to select new inputs for the next layer of the MGMDH model, the selection criteria employ the mean squared error (MSE) value. In completing the previous process, U possible new input variables for the next layer have been constructed. Then, the identification of the new input proceeds based on the MSE value, where the best variable is selected and the weakest are eliminated. It should be noted that, after determining the new input variable, the entire procedure is repeated until minimum $MSE_r \geq MSE_{r-1}$, where r is the number of the current layer. The process stops when the MSE value for the current layer is greater than that from the previous layer. The MSE is defined by Eq. 10.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{i,k})^2 \quad (10)$$

2.2 Comparison Model

There are three models used for comparison: artificial neural network (ANN), group method of data handling (GMDH), nonlinear regression, and linear regression. The methods have been described in Badyalina et al. (2021).

2.3 Performance Criteria

The performance of each model is evaluated with the following error indices, which is the mean absolute percentage error. The definitions of RMSE and MAPE are provided in Eq. 11.

$$MAPE = \left(\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|\right) \times 100\% \quad (12)$$

where y_i is the observed flows, $\hat{y}_i$ is the predicted flows, $\bar{y}$ is the mean of the observed flows and n is the number of flow series that have been model.

4. Result and Discussion

Historical hydrological record related to streamflow data carries vital information for decision making involved in planning, designing, and managing water projects. A long history of past data can extrapolate future events well and thus produce high accuracy of flood estimation. However, in a country like Malaysia, hydrological data and information are limited. In such cases, the simulation technique serves as a statistical problem-solving tool by using data simulation. The simulation technique generates the data according to the characteristics of the actual data used in the study. Simulation is a tool to evaluate the performance of existing and proposed under configured conditions of simulation data. Based on the actual data used by Badyalina et al. (2021), the (variance inflation factor) VIF value for catchment area is 19.4803, VIF value elevation is 1.4887, VIF value Longest Drainage Path is 20.2307, VIF value for river slope is 1.2618, and VIF value for annual mean maximum total rainfall is 1.0147. Therefore, it can be concluded that there is always a presence of multicollinearity. Multicollinearity is a phenomenon in which two or more predictor variables in a model are highly correlated. The Cauchy Distribution is used to mimic the catchments that have an extreme value of the return period. In this section, the data for simulation is generated using the method discussed in Section 3.8. There are five models implemented in this study: MGMDH, LR, NLR, GMDH and ANN model. The result for simulation studies is shown in Table 2 until Table 7.

Table 2. Values of RMSE and BIAS of LR, NLR, ANN, GMDH and MGMDH for $\rho^2 = 0$ with Cauchy disturbance distribution for sample size $n = 30,50,70,100$

$\rho^2 = 0$; Outliers = 5%					
		MAPE			
Input	Model	$n = 30$	$n = 50$	$n = 70$	$n = 100$
	LR	48.64	34.82	52.54	41.06
	NLR	42.39	43.78	69.23	44.35
2	ANN	26.96	28.04	45.10	25.60
	GMDH	32.78	40.72	61.53	35.98
	MGMDH	38.27	39.08	58.63	33.85
	LR	44.64	61.35	50.99	62.74
	NLR	29.07	64.18	54.32	72.95
3	ANN	41.91	46.24	65.93	54.71
	GMDH	35.39	51.63	44.92	66.57
	MGMDH	**26.47**	55.35	58.21	**48.32**
	LR	25.93	48.19	80.63	41.28
	NLR	42.22	42.47	76.94	39.18
4	ANN	31.90	35.97	72.97	30.57
	GMDH	36.17	38.05	66.56	29.76
	MGMDH	**21.56**	**28.58**	**52.79**	25.59
	LR	43.82	67.02	35.31	48.82
	NLR	52.04	54.97	46.47	34.86
5	ANN	30.23	44.12	41.26	47.11
	GMDH	31.58	60.47	37.52	42.13
	MGMDH	**27.76**	49.71	**29.81**	**27.84**

Table 2 shows the prediction performance when multicollinearity does not present in the simulated data. The result shows that the best performance prediction model is the MGMDH model followed by the ANN model. The MGMDH model shows good performance when the number of input variables increases. ANN model shows god prediction performance when the sample size is small.

Table 3. Values of RMSE and BIAS of LR, NLR, ANN, GMDH and MGMDH for $\rho^2 = 0.5$ with Cauchy disturbance distribution for sample size $n = 30,50,70,100$

$\rho^2 = 0.5$; Outliers = 5%					
		MAPE			
Input	Model	$n = 30$	$n = 50$	$n = 70$	$n = 100$
	LR	48.34	36.47	51.84	59.09
	NLR	36.81	37.16	47.92	53.38
2	ANN	32.16	32.47	38.38	35.82
	GMDH	35.97	41.72	41.44	47.44
	MGMDH	37.89	40.48	45.63	42.51
	LR	51.36	54.03	57.62	62.31
	NLR	61.16	59.77	51.69	55.16
3	ANN	45.80	37.33	41.65	41.26
	GMDH	58.11	42.07	48.85	48.30
	MGMDH	**38.44**	49.73	46.98	**34.97**
	LR	54.82	60.40	56.91	78.01
	NLR	62.81	53.62	61.96	62.46
4	ANN	45.38	40.23	43.39	59.54
	GMDH	49.37	35.65	49.96	48.54
	MGMDH	**41.97**	44.45	**38.54**	**45.24**
	LR	48.38	51.01	55.43	31.70
	NLR	42.83	56.61	51.30	51.38
5	ANN	37.53	43.53	39.09	27.08
	GMDH	33.53	42.59	44.39	41.20
	MGMDH	**28.06**	**35.29**	**36.16**	36.51

Table 3 shows the prediction performance when the level of multicollinearity increase to 0.5. Based on Table 3, the proposed MGMDH model performs well when the number of input variables increases.

Table 4. Values of RMSE and BIAS of LR, NLR, ANN, GMDH and MGMDH for $\rho^2 = 0.99$ with Cauchy disturbance distribution for sample size $n = 30,50,70,100$

$\rho^2 = 0.99$; Outliers = 5%					
		MAPE			
Input	Model	$n = 30$	$n = 50$	$n = 70$	$n = 100$
2	LR	63.80	21.53	38.93	38.93
	NLR	57.88	25.74	46.93	46.93
	ANN	42.16	28.33	37.55	37.55
	GMDH	51.52	32.54	35.69	35.69
	MGMDH	48.10	43.51	42.95	42.95
3	LR	54.01	44.17	42.33	42.33
	NLR	53.99	42.67	52.24	52.24
	ANN	40.05	28.34	37.35	37.35
	GMDH	38.28	31.29	45.96	45.96
	MGMDH	42.19	38.92	49.90	49.90
4	LR	49.59	27.19	50.19	37.32
	NLR	53.91	37.96	64.93	41.71
	ANN	44.98	34.63	58.36	48.47
	GMDH	37.41	39.47	44.53	57.72
	MGMDH	**33.11**	**25.88**	52.01	**31.32**
5	LR	36.14	47.50	47.94	38.35
	NLR	26.83	49.52	59.42	45.11
	ANN	30.55	39.54	42.72	32.84
	GMDH	28.48	43.05	53.51	40.92
	MGMDH	**23.23**	**35.81**	**35.60**	**26.60**

Table 4 shows the prediction performance when the level of multicollinearity increase to 0.99. Based on Table 4, the proposed MGMDH model performs well when the number of input variables increases.

Table 5. Values of RMSE and BIAS of LR, NLR, ANN, GMDH and MGMDH for $\rho = 0$ with Cauchy disturbance distribution for sample size $n = 30,50,70,100$

$\rho^2 = 0$; Outliers = 10%					
		MAPE			
Input	Model	$n = 30$	$n = 50$	$n = 70$	$n = 100$
2	LR	25.68	47.56	49.79	64.37
	NLR	44.90	53.45	46.23	61.57
	ANN	30.78	46.03	43.88	39.38
	GMDH	27.68	62.63	53.64	53.39
	MGMDH	34.65	49.71	55.15	44.72
3	LR	29.21	66.79	55.16	32.15
	NLR	37.97	51.43	49.08	36.09
	ANN	25.10	39.06	47.74	33.93
	GMDH	30.03	43.85	42.34	43.88
	MGMDH	33.84	61.88	**37.53**	**28.99**
4	LR	57.32	48.60	60.31	34.16
	NLR	62.65	41.47	68.49	31.54
	ANN	42.34	33.61	46.29	27.56
	GMDH	53.90	26.67	53.25	29.02
	MGMDH	49.82	29.25	**42.48**	**23.80**
5	LR	38.98	58.69	42.85	58.27
	NLR	43.35	52.62	36.50	61.67
	ANN	31.43	40.20	31.50	66.31
	GMDH	34.02	48.97	28.02	47.17
	MGMDH	**26.11**	**38.82**	**25.36**	**38.46**

Table 5 shows the prediction performance when multicollinearity is set to 0 but the outliers percentage to 10%. Table 5 shows that the proposed MGMDH model performs well when the number of input variables increases.

Table 6. Values of RMSE and BIAS of LR, NLR, ANN, GMDH and MGMDH for $\rho^2 = 0.5$ with Cauchy disturbance distribution for sample size $n = 30,50,70,100$

$\rho^2 = 0.5$; Outliers = 10%					
		MAPE			
Input	Model	$n = 30$	$n = 50$	$n = 70$	$n = 100$
2	LR	49.13	47.61	28.26	48.41
	NLR	58.80	35.26	33.01	55.91
	ANN	42.33	27.25	23.23	39.24
	GMDH	53.70	31.10	25.06	51.71
	MGMDH	46.86	42.13	35.55	44.13
3	LR	75.56	42.34	50.42	40.68
	NLR	53.41	48.82	52.48	48.07
	ANN	48.01	36.69	45.48	29.50
	GMDH	62.04	40.76	48.30	43.84
	MGMDH	**41.77**	45.74	**43.42**	**27.98**
4	LR	63.39	28.61	57.96	38.83
	NLR	53.97	38.22	68.17	45.91
	ANN	37.54	26.60	54.90	33.44
	GMDH	58.39	34.97	49.15	41.45
	MGMDH	43.93	24.88	**42.48**	**29.79**
5	LR	41.96	31.12	40.09	38.81
	NLR	46.56	35.44	46.88	42.45
	ANN	37.97	23.22	45.99	48.51
	GMDH	48.23	27.57	29.82	51.87
	MGMDH	**32.82**	25.39	**26.31**	56.60

Table 6 shows the prediction performance when multicollinearity is increased to 0.5 and the outliers percentage to 10%. Table 6 shows that the proposed MGMDH model performs well when the number of input variables increases. Other than the MGMDH model, the ANN model also shows good prediction performance.

Table 7. Values of RMSE and BIAS of LR, NLR, ANN, GMDH and MGMDH for $\rho^2 = 0.99$ with Cauchy disturbance distribution for sample size $n = 30,50,70,100$

$\rho^2 = 0.99$; Outliers = 10%					
		MAPE			
Input	Model	$n = 30$	$n = 50$	$n = 70$	$n = 100$
2	LR	121.93	38.48	48.91	42.74
	NLR	85.11	34.17	58.67	51.72
	ANN	44.71	30.31	40.73	36.16
	GMDH	68.79	37.10	47.38	45.78
	MGMDH	53.53	41.82	43.95	49.92
3	LR	32.55	54.08	36.32	57.58
	NLR	33.62	50.58	34.14	48.43
	ANN	38.52	49.88	32.50	42.59
	GMDH	27.12	46.68	28.92	50.12
	MGMDH	**24.30**	55.32	**25.66**	46.66
4	LR	42.77	36.19	65.53	45.21
	NLR	45.94	33.61	55.06	53.80
	ANN	38.80	28.72	45.77	48.82
	GMDH	30.49	30.43	48.40	41.77
	MGMDH	35.42	**25.91**	**41.26**	**38.74**
5	LR	43.77	49.10	48.46	38.75
	NLR	47.24	66.26	44.46	44.78
	ANN	35.93	51.35	32.84	33.93
	GMDH	41.26	58.23	36.31	41.66
	MGMDH	**31.53**	**42.49**	39.77	**27.38**

Table 7 shows the prediction performance when multicollinearity is increased to 0.99 and the outlier's percentage to 10%. Table 7 shows that the proposed MGMDH model performs well when the number of input variables increases. The design case study is to investigate the prediction performance with and without the presence of multicollinearity. Other than that, the Cauchy Distribution is used as error disturbance or noise. Cauchy distribution will have a significant tendency to produce extreme value because Cauchy Distribution is a heavy-tailed distribution. Heavy tailed distribution is a highly skewed distribution. It shows that the random number produce by the Cauchy Distribution is very large, which is suitable for the simulation of extreme value. On the other hand, a various number of input variables and sample sizes are used. The sample size chosen in this study, as the previous study done by Badyalina et al. (2021), only used 70 stations in the data set. Therefore, the sample size generated is set between 30 and 100. The number of input variables generated to simulate the catchment characteristics. The winning model is the model that has the lowest MAPE for prediction. The summary of the simulation study is shown in Table 7, which shows the win frequency of each model in each case study design.

Table 8. Model winning frequency based on MAPE and RMSE

	Win Frequency				
Model Simulation Condition	MGMDH	ANN	GMDH	NLR	LR
$\rho^2 = 0$; Outliers = 5%	8	7	1	0	0
$\rho^2 = 0.5$; Outliers = 5%	8	6	2	0	0
$\rho^2 = 0.99$; Outliers = 5%	7	6	2	0	1
$\rho^2 = 0$; Outliers = 10%	8	6	1	0	1
$\rho^2 = 0.5$; Outliers = 10%	9	7	0	0	0
$\rho^2 = 0.99$; Outliers = 10%	8	6	2	0	0

Table 7 shows the win frequency for MGMDH, ANN, GMDH, NLR and LR models based on MAPE and RMSE. From Table 7, it can be summarized that the MGMDH model outperformed other models in all cases. On average, the winning frequency percentages for MGMDH, ANN, GMDH, NLR and LR model are 50 %, 40 %, 8 %, 0 % and 2 %, respectively. In addition, the MGMDH model outperformed the ANN model when using Cauchy Distribution as an error disturbance. Using Cauchy Distribution as noise makes output variables have more extreme values compared to Normal Distribution. Based on real data used by Badyalina et al. (2021), the flood quantile characteristics are highly skewed, indicating that the three specific flood quantile contains extreme values. Another observation is that when the outliers increased from 5% to 10 %, the still result shows that MGMDH is superior to other models in all cases. Based on this observation, the results will be the same by increasing outlier to 20% or more. This simulation study showed that when extreme values were present in the output variable, MGMDH outperformed other models based average winning frequency percentage. On the other hand, the MGMDH model has a higher winning percentage than the GMDH model that is 50% for the MGMDH model and 8% for the GMDH model, respectively. It shows that the implementation of PCA and four type transfer functions have improved the GMDH model's prediction performance. Therefore, based on these results, the MGMDH model has the most efficient and robust prediction performance compared to other models when the data set contains extreme values. For future research, it is suggested to hybrid the GMDH model optimization tools such as the artificial bee colony algorithm.

4. Conclusions

This study explores the potential of the MGMDH model in prediction at ungauged sites. In this study, the combination of principal component analysis (PCA) and group method of data handling (GMDH) with various transfer functions, namely modified group method of data handling (MGMDH), is proposed for prediction at ungauged catchment. MGMDH model consists of four types of transfer function: polynomial, sigmoid, radial basis function, and hyperbolic tangent sigmoid transfer function compared to the conventional GMDH model. In order to demonstrate the appropriateness and effectiveness of the proposed models, a simulation study was done. The simulation study used Cauchy Distribution as a disturbance error for the simulation data. Implementation of Cauchy distribution as error disturbance in artificial data evaluated the model prediction performance if the extreme value or extreme event occurs in the data set. The simulation study showed that the MGMDH model is superior to other comparison models, namely LR, NLR, GMDH and ANN models. Other than that, the MGMDH model shows strong prediction performance when multicollinearity is not present in the data set.

Acknowledgements

The authors gratefully acknowledge the financial support by Universiti Teknologi MARA Johor under grant Bestari 600-UiTMCJ (PJIA. 5/2).

References

Al-Ashkar, I., Al-Suhaibani, N., Abdella, K., Sallam, M., Alotaibi, M., & Seleiman, M. F. (2021). Combining Genetic and Multidimensional Analyses to Identify Interpretive Traits Related to Water Shortage Tolerance as an Indirect Selection Tool for Detecting Genotypes of Drought Tolerance in Wheat Breeding. *Plants*, *10*(5), 931. https://doi.org/10.3390/plants10050931

Alobaidi, M. H., Ouarda, T. B., 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. https://doi.org/10.1016/j.advwatres.2020.103814

Badyalina, B., Mokhtar, N. A., Jan, N. A. M., Hassim, N. H., & Yusop, H. (2021). Flood Frequency Analysis using L-Moment For Segamat River. *MATEMATIKA: Malaysian Journal of Industrial and Applied Mathematics*, 47-62.

Badyalina, B., Mokhtar, N. A., Ramli, M. F., Majid, M., & Yusri, M. Y. (2021). Design of Simulation Studies for Flood Quantile Prediction Problems at Ungauged Site. *Applied Mathematical Sciences*, *15*(3), 137-140. https://doi.org/10.12988/ams.2021.914422

Badyalina, B., & Shabri, A. (2013). Streamflow Forecasting at Ungauged Sites using Multiple Linear Regression. *MATEMATIKA: Malaysian Journal of Industrial and Applied Mathematics*, 67-75.

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., Shabri, A., & Jan, N. (2016). Prediction At Ungauged Site with Topological Kriging And Modified Group Method Of Data Handling. *Journal Of Environmental Hydrology*, *24*, 6.

Badyalina, B., Shabri, A., & Marsani, M. F. (2021). Streamflow Estimation at Ungauged Basin using Modified Group Method of Data Handling. *Sains Malaysiana*, *50*(9), 2765-2779. https://doi.org/10.17576/jsm-2021-5009-22

Badyalina, B., Shabri, A., & Samsudin, R. (2014). Streamflow estimation at ungauged site using wavelet group method of data handling in Peninsular Malaysia, *International Journal of Mathematical Analysis*, *8*(11), 513–24. https://doi.org/10.12988/ijma.2014.4251

Barth, J., Katumullage, D., Yang, C., & Cao, J. (2021). Classification of wines using principal component analysis. *Journal of Wine Economics*, *16*(1), 56-67. https://doi.org/10.1017/jwe.2020.35

Campos, J. A., & Pedrollo, O. C. (2021). A regional ANN-based model to estimate suspended sediment concentrations in ungauged heterogeneous basins. *Hydrological Sciences Journal*(just-accepted). https://doi.org/10.1080/02626667.2021.1918695

Desai, S., & Ouarda, T. B. (2021). Regional hydrological frequency analysis at ungauged sites with random forest regression. *Journal of Hydrology*, *594*, 125861. https://doi.org/10.1016/j.jhydrol.2020.125861

Grimaldi, S., Nardi, F., Piscopia, R., Petroselli, A., & Apollonio, C. (2021). Continuous hydrologic modelling for design simulation in small and ungauged basins: A step forward and some tests for its practical use. *Journal of Hydrology*, *595*, 125664.

Guo, Y., Zhang, Y., Zhang, L., & Wang, Z. (2021). Regionalization of hydrological modeling for predicting streamflow in ungauged catchments: A comprehensive review. *Wiley Interdisciplinary Reviews: Water*, *8*(1), e1487. https://doi.org/10.1002/wat2.1487

Habshah, M., & Marina, Z. (2007). A simulation study on ridge regression estimators in the presence of outliers and multicollinearity. *Journal Teknologi*, *47*, 59-74. https://doi.org/10.11113/jt.v47.261

Haddad, K. (2021). Selection of the best fit probability distributions for temperature data and the use of L-moment ratio diagram method: a case study for NSW in Australia. *Theoretical and applied climatology*, *143*(3), 1261-1284. https://doi.org/10.1007/s00704-020-03455-2

Jan, N. A. M., & Shabri, A. (2017). Estimating distribution parameters of annual maximum streamflows in Johor, Malaysia using TL-moments approach. *Theoretical and applied climatology*, *127*(1-2), 213-227. https://doi.org/10.1007/s00704-015-1623-7

Jan, N. A. M., Shabri, A., & Badyalina, B. (2016). Selecting probability distribution for regions of Peninsular Malaysia streamflow. AIP Conference Proceedings. https://doi.org/10.1063/1.4954619

Jan, N. A. M., Shabri, A., Hounkpè, J., & Badyalina, B. (2018). Modelling non-stationary extreme streamflow in Peninsular Malaysia. *International Journal of Water*, *12*(2), 116-140. https://doi.org/10.1504/IJW.2018.091380Jan, N. A. M., Shabri, A., Ismail, S., Badyalina, B., Abadan, S. S., & Yusof, N. (2016). Three-Parameter Lognormal

Distribution: Parametric Estimation Using L-Moment And Tl-Moment Approach. *Jurnal Teknologi*, *78*(6-11). https://doi.org/10.11113/jt.v78.9202

Kondo, T., & Ueno, J. (2008). Multi-layered GMDH-type neural network self-selecting optimum neural network architecture and its application to 3-dimensional medical image recognition of blood vessels. *International Journal of innovative computing, information and control*, *4*(1), 175-187.

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.

Lawrence, K. D., & Arthur, J. L. (2019). Robust nonlinear regression. In *Robust Regression* (pp. 59-86). Routledge. https://doi.org/10.1201/9780203740538-3

Li, M., Robertson, D. E., Wang, Q. J., Bennett, J. C., & Perraud, J.-M. (2021). Reliable hourly streamflow forecasting with emphasis on ephemeral rivers. *Journal of Hydrology*, *598*, 125739. https://doi.org/10.1016/j.jhydrol.2020.125739

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. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000464

Mokhtar, N. A., Badyalina, B., Chang, K. L., Yaa'cob, F., Ghazali, A., & Shamala, P. (2021). Error-in-Variables Model of Malacca Wind Direction Data with the von Mises Distribution in Southwest Monsoon. *Applied Mathematical Sciences*, *15*(9), 471-479. https://doi.org/10.12988/ams.2021.914521

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. https://doi.org/10.1088/1742-6596/1988/1/012097

Nariman-Zadeh, N., Darvizeh, A., Felezi, M., & Gharababaei, H. (2002). Polynomial modelling of explosive compaction process of metallic powders using GMDH-type neural networks and singular value decomposition. *Modelling and Simulation in Materials Science and Engineering*, *10*(6), 727. https://doi.org/10.1088/0965-0393/10/6/308

Nazirah, A., Sabki, W. W. M., Zulkarnian, H., & Afizah, A. (2021). Simulation of runoff using HEC-HMS for ungauged catchment. AIP Conference Proceedings. https://doi.org/10.1063/5.0051957

Onwubolu, G. C., & Davendra, D. (2009). *Differential evolution: A handbook for global permutation-based combinatorial optimization* (Vol. 175). Springer Science & Business Media. https://doi.org/10.1007/978-3-540-92151-6

Rana, S., Midi, H., & Imon, A. R. (2012). Robust wild bootstrap for stabilizing the variance of parameter estimates in heteroscedastic regression models in the presence of outliers. *Mathematical Problems in Engineering*, *2012*. https://doi.org/10.1155/2012/730328

Razaq, S. A., Ismail, T., Heryansyah, A., Alamgir, M., & Pour, S. H. (2016). Streamflow prediction in ungauged catchments in the east coast of Peninsular Malaysia using multivariate statistical techniques. *Jurnal Teknologi*, *78*(6-12). https://doi.org/10.11113/jt.v78.9231

Romali, N. S., & Yusop, Z. (2017). Frequency analysis of annual maximum flood for Segamat River. MATEC Web of Conferences.

Sahu, R. T., Verma, M. K., & Ahmad, I. (2021). Regional Frequency Analysis Using L-Moment Methodology—A Review. *Recent Trends in Civil Engineering*, 811-832. https://doi.org/10.1007/978-981-15-5195-6_60

Sivapalan, M., Takeuchi, K., Franks, S. W., Gupta, V. K., Karambiri, H., Lakshmi, V., . . . 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. https://doi.org/10.1623/hysj.48.6.857.51421

Yao, X., & Liu, Y. (1996). Fast Evolutionary Programming. *Evolutionary programming*, *3*, 451-460.