Streamflow Estimation at Ungauged Site Using
Wavelet Group Method of Data Handling in
Peninsular Malaysia

Basri Badyalina
Faculty of Science, Universiti Teknologi Malaysia

Ani Shabri
Faculty of Science, Universiti Teknologi Malaysia

Ruhaidah Samsudin
Faculty of Computing, Universiti Teknologi Malaysia

Copyright © 2014 Basri Badyalina et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

This study investigates the ability of wavelet group method of data handling (WGMDH) conjunction model in the estimation of flood quantiles in ungauged sites in Peninsular Malaysia. The conjunction method was obtained by combining two methods, discrete wavelet transform and group method of data handling. Comparison between the WGMDH model, group method of data handling model (GMDH), wavelet regression (WR) model and linear regression (LR) model were done. To assess the effectiveness of this model, 70 catchments in the province of Peninsular Malaysia were used as case studies. The performance of WGMDH model was compared with the conventional LR, GMDH and WR models using various statistical measures such as the mean absolute error, root mean square error and Nash-Sutcliffe coefficient of efficiency. Jackknife procedure was required for the evaluation of the performance of the four approaches. The jackknife procedure was needed to simulate the ungauged sites. The results of the comparison indicate that the WGMDH model was more accurate and perform
better than the traditional LR, GMDH and WR models. Thus, WGMDH model is a promising new method for estimation of flood quantiles in ungauged sites.

**Keywords**: group method of data handling, ungauged, discrete wavelet transform

### 1 Introduction

Flood is one of the most dangerous and recurrent type of natural disasters that occurs in Peninsular Malaysia. Floods are one of the leading causes of death and property damage from natural disasters in Malaysia and it occurs almost every year. This make a reliable estimation of flood quantiles is important for flood risk assessment project (e.g., dams, spillways, road, and culverts), the safe design of the river system, and it give a closed valuation budget of flood protection project.

In order to acquire accurate estimation of flood quantiles, recorded historical time series data of stream flows is required. Long term historical data used for estimation are more reliable compared to short term data and may also reduce risk. However, it often happens that the historical data at-site of interest not always available. Although at-site of interest may have some available data but the data are not enough to describe the catchment flow because of the changes in watershed characteristics such as urbanization (Pandey and Nguyen, 1999). The UK Flood Estimation Handbook (FEH) notes that “many flood estimation problems arise at ungauged sites which there are no flood peak data” (Reed and Robson, 1999). Mamun et al. (2012) stated that river located in Malaysia is gauged only at a strategic location and other river is usually ungauged.

Typically some site characteristics for the ungauged sites are known. Thus, regionalization is carried out to make the estimation of flow statistics at ungauged sites using physiographic characteristics. Regionalization technique includes fitting a probability distribution to series of flow and then linking the relationship to catchment characteristics (Dawson et al., 2006). The variables affecting the flood quantile estimation include catchment characteristics (size, slope, shape and storage characteristics of the catchment), storm characteristics (intensity and duration of rainfall events), geomorphologic characteristics (topology, land use patterns, vegetation and soil types that affect the infiltration) and climatic characteristics (temperature, humidity and wind characteristics) (Jain and Kumar 2007). In relating flood quantile at site of interest to catchment characteristics a power form equations are mostly used (Pandey and Nguyen, 1999; Mamun, 2012).

At ungauged sites, linear regression (LR) model is always reliable to make estimates of flow statistics or flood quantiles (see e.g. Shu & Ouarda, 2008; Pandey & Nguyen, 1999). The performances of LR models in estimating the flood quantiles for ungauged sites have been assessed in Pandey and Nguyen (1999) by applying jackknife procedure in simulating the ungauged sites.

Group Method of Data Handling (GMDH) has shown a significant improvement by doing a combination with other method. Zadeh et al. (2002) combined GMDH model with singular value decomposition and it has shown that
the combined method prediction is better than GMDH itself. Nowadays, wavelet transform analysis has gained its popularity because it can produce an encouraging outcome in multi-resolution analysis, variations, periodicities, and trends in time series. The wavelet transforms has the ability to decompose a signal into different level of decompositions which allows the required information to be extracted from data. Usually the extracted data gained from wavelet transformation become the input to other model. The result shows a significant improvement in predictions ability of the model applied. Thus, the ability wavelet transform has become a major reason in improving the ability of model applied predictions. Kisi (2009) had proposed the combination of the wavelet transform and linear regression since the hybrid model is much easier to interpret for monthly stream flow forecasting.

The aims of the present investigation are: 1) to explore the potential application of wavelet group method of data handling (WGMDH) solutions to the problem of flood estimation in ungauged catchments; 2) to compare WGMDH model estimation performance with conventional method linear regression (LR), GMDH and wavelet regression (WR).

2 Catchment Data Set

2.1 Introduction

The data were obtained from Department of Irrigation and Drainage, Ministry of Natural Resources and Environment, Malaysia. There were seventy gauged stations selected including all the stations located at Peninsular Malaysia. They are located within latitude 1° N-5° N and longitudes of 100° N-104° N. The stations include wide variety of basins region ranging between 16.3 km2 to 19,000 km2. The period of the flow series for different sites varies from 11 -50 years starting from 1959 – 2009. These data were processed in two stages. First, catchment descriptors were extracted for each site. Second, the annual peak flow was used to estimate selected T-year flood events for each catchment. Five distributions used are GEV, GLO, GPA, LN3 and P3. Distributions of GEV, GLO, GPA, LN3 and P3 is a widely variety of applications for estimating values of given data sets. Hence, the best fitted distributions at each stations are used to estimate flood quantile with return period for T=10, T=50, T=50 years.

2.2 Catchment descriptors

The variables selected in this study on the basis of previous study by Seckin (2011) and by Shu and Ouarda (2008). The four physiographical variables are catchment area, elevation, mean river slope and longest drainage path. The meteorological variable is mean annual total rainfall. The summary statistics of these variables are presented in Table 1.
3 Methodology

3.1 Group Method of Data Handling

Group Method of Data Handling (GMDH) model was introduced by Ivakhnenko on 1970 to solve complex non-linear multidimensional that has short data series (Ivakhnenko 1970). The GMDH algorithm that describes the relationship between input and output signal can be represented by Volterra series (Ivakhnenko 1970) in form of:

$$\text{Eq. 1 is known as Kolmogorov-Gabor polynomial. From Eq. 1, } X(x_1, x_2, ..., x_n) \text{ is referring to input variable vector, } n \text{ is the number of input and } V(v_1, v_2, ..., v_n) \text{ is vector of coefficient weight. In the GMDH algorithm, Eq. 1 is called the complete description of the nonlinear system. However, most application only used second order polynomial called partial descriptions (PD) of the nonlinear system that can be expressed by a system of transfer function consisting of only two variables (Srinivasan 2008; Najafzadeh & Barani 2011). The PD is in the form of}

$$\hat{y} = v_0 + v_1 x_1 + v_2 x_2 + v_3 x_1 x_2 + v_4 x_1^2 + v_5 x_2^2$$

$$\text{Eq. 2 as partial description (PD) provides the mathematical relation between the input and output variable. Linear regressions mostly applied in GMDH to obtain the weight coefficients for the models (Ivakhnenko 1970; Zadeh et. al 2002). The data set that consists of input and output are divided into two subset which are the modeling and forecasting based on jackknife procedure. The input variable } X = \{x_1\} \text{ are paired using partial description in Eq. 2 in modeling data set. Then a linear regression applied in Eq. 2 is to obtain the vector of coefficients.}

$$Gv = Y$$

$$\text{where } v \text{ is the vector of coefficient of the partial description in Eq. 2.}

$$v = \{v_0, v_1, v_2, v_3, v_4, v_5\}$$

$$\text{and}

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}^T$$
is the vector of output from training data set.

\[
G = \begin{bmatrix}
1 & x_{i_l} & x_{i_j} & x_{i_l}x_{i_j} & x_{i_l}^2 & x_{i_j}^2 \\
1 & x_{2_l} & x_{2_j} & x_{2_l}x_{2_j} & x_{2_l}^2 & x_{2_j}^2 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & x_{pi} & x_{pj} & x_{pi}x_{pj} & x_{pi}^2 & x_{pj}^2 
\end{bmatrix}
\] (6)

Then, the best estimated coefficients of partial description in Eq. 4 were obtained in the form of:

\[
v = (G^T G)^{-1} G^T Y
\] (7)

Therefore in each layer the total number of PD generated and the RMSE are as follows:

\[
U = ^nC_2 = \frac{n!}{(n-2)!2!}
\] (8)

\[
RMSE_k = \sqrt{\frac{1}{p} \sum_{n=1}^{p} (y_n - \hat{y}_{i,k})^2} \quad k \in [1,2,\ldots,U]
\] (9)

where \( n \) is the number of input in each layer. The vector coefficient of each PD is determined using linear regression then forming the quadratic equation which approximates the output \( \hat{y} \). After completing the process, the algorithm has constructed \( U \) number of new input variable but only one from \( U \) is chosen for the new input of GMDH based on RMSE value. This approach for identification of GMDH-type networks is called as error driven approach (Zadeh et al. 2002). After determining the new input, the whole GMDH process is repeated again. If \( RMSE_k \leq RMSE_{k-1} \), set new input variables and repeat the GMDH process, otherwise if RMSE show an improvement the process is stopped and use the results from the previous minimum value of RMSE.

### 3.2 Discrete Wavelet Transform

The Discrete Wavelet Transform (DWT) involves choosing scales and positions based on powers of two called dyadic scales and translation. The mother wavelet is rescaled or dilated by powers of two and translated by integers. The results are more efficient and are just as accurate (Kisi et al., 2008). The DWT algorithm is able to produce coefficients of fine scales for capturing low information and coefficients of coarse scales for capturing low frequency information. Wavelet function \( \psi(t) \) called the mother wavelet can be defined as
\[
\int_{-\infty}^{\infty} \psi(t)dt = 0
\]
can be obtained through compressing and expanding continuous wavelet transform.

\[
\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right) \quad a \in \mathbb{R}, b \in \mathbb{R}, a \neq 0
\]  

(10)

where \( b \) is location parameter, \( a \) is scaling parameter and \( \mathbb{R} \) is the domain of real numbers. If \( \psi_{a,b}(t) \) satisfies for data \( f(t) \in L^2(\mathbb{R}) \) or finite energy signal (Zhou et al., 2008) The discrete wavelet transform of \( f(t) \) can be written as:

\[
W_{\psi}f(j,k) = a_0^{-j/2} \int_{\mathbb{R}} f(t) \phi(a_0^{-j} t - kb_0) dt
\]  

(11)

The most common and simplest choice for the parameters \( a_0 \) and \( b_0 \) are 2 and 1 time steps respectively. This power of two logarithmic scaling of the time and scale is known as dyadic grid arrangement and it is the simplest and most efficient case for practical purposes (Mallat et al., 1989)

\[
W_{\psi}f(j,k) = 2^{-j/2} \int_{\mathbb{R}} f(t) \phi(2^{-j} t - k) dt
\]  

(12)

where \( j \) and \( k \) are integers representing the set of discrete translations and discrete scaling respectively.

### 3.3 Wavelet Group Method of Data Handling

Wavelet Group Method of Data Handling (WGMDH) was obtained by combining two methods that are DWT and GMDH. The DWT was applied to the input variable where the DWT decomposed each of the variables into a number of components. The DWT decomposed the variable using Mallat DWT algorithm (Mallat 1989). The Mallat algorithm translates each data without losing the information about the element in the original data. The decomposition process is a successive filtering process where the input variable decomposed into an approximation and detail components This allows the GMDH to learn about the characteristics of the data and produce good estimation. The WGMDH model is the improvement of GMDH model by combining the two methods which are the Discrete Wavelet Transform (DWT) model and the Group Method of Data Handling (GMDH). In this study, multivariate input variables were used. DWT decomposed the input variables into several components that are approximate and details. The numbers of detail components are dependent on the number of resolution level implemented for that DWT. The number of resolution levels implemented in this study is 2, 3, 4 and 5. After the original input variables are decompose using DWT, then the effective components of approximate and details are chosen using correlation. Only the effective components after decomposition
Streamflow estimation at ungauged site

using DWT become input for GMDH model. The effective components are chosen based on positive correlation. Only the components had the positive correlation will be selected as effective components. The proposed WGMDH model structure is shown in Fig. 4.1.

The main motivation of using wavelet transform is the easy analysis of the obtained series. Wavelet transform decomposes complex data into a number of component signals and minimized error criterion. The wavelet also functions as a filter and in this study, the wavelet amplifies the noise from raw data. This will make it easier for GMDH to examine the clean signals. According to Cannas et al. (2006), the results showed that networks trained with preprocessed data performed better than networks trained on noisy raw signals. In this study various number of resolution level are implemented because there are no theory to determine suitable resolution number for any given data. The type of wavelet used in this study is Daubechies wavelet that had been carried out on the discrete wavelet transform. Daubechies wavelet approximate the decomposition signals more accurately. The Daubechies wavelet is very good at representing polynomial behavior within the signals.

3.4 Evaluation Criteria

The performance of each model is evaluated with the following error indices which are the mean absolute error (MAE), root mean square error (RMSE) and Nash-Sutcliffe coefficient of efficiency (CE) and correlation coefficient. The definitions of MAE, RMSE and CE are provided in Eq. 13- Eq. 14, respectively.

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} \left| Q_{T,i} - \hat{Q}_{T,i} \right|
\]

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Q_{T,i} - \hat{Q}_{T,i})^2}
\]

\[
CE = 1 - \frac{\sum_{i=1}^{n} (Q_{T,i} - \hat{Q}_{T,i})^2}{\sum_{i=1}^{n} (Q_{T,i} - \bar{Q}_{T,i})^2}
\]

where \( Q_{T,i} \) is the observed flows, \( \hat{Q}_{T,i} \) is the predicted flows, \( \bar{Q}_{T,i} \) is the mean of the observed flows, \( \bar{Q}_{T,i} \) is the mean of the predicted flows and \( n \) is the number of flow series that have been modeled. The coefficient of efficiency (CE) provides an indication of how good a model is at predicting values away from the mean. CE ranges from \(-\infty\) in the worst case to 1 (perfect fit). The efficiency of lower than zero indicates that the mean value of the observed flow would have been a better predictor than the model. variance and the multiplied standard deviations of observed and predicted values.

4 Results and Discussion

In this study there are four types of model used to estimate flood quantile at ungauged site. The models are linear regression (LR), group method of data
handling (GMDH), wavelet regression (WR) and wavelet group method of data handling (WGMDH). To simulate the ungauged site, a jackknife procedure is implemented. In jackknife procedure, one site is removed from data and model parameters are estimated using the data from remaining site. The estimated parameters are in turn used to predict quantile for the site not used in the model development. The process is repeated until all stations are removed at least once. The input variables for all models are catchment area, elevation, longest drainage path, river slope and annual mean total rainfall.

The wavelet transform allows the decomposition of data into different resolution levels which means the required data components can be extracted. Wavelet decomposition partitioned the data into approximate scale and details scale. The number of details components is dependent on resolution levels implemented. In this study, four different levels are implemented that are two resolution level, three resolution levels, four resolutions levels and five resolution levels. Many resolution levels are used because there is no method to determine which resolution levels are suitable for any data. The input variables were decomposed into various Ds at different resolution level by using DWT for estimating $Q_{10}$, $Q_{50}$ and $Q_{100}$. The DWT decomposed the variable using Mallat DWT algorithm (Mallat, 1989). The decomposition process is a successive filtering process where the input variable decomposed into an approximation and detail components. The approximation is the high scale (low frequency) and the details are the low scale (high frequency). Table 2 shows the correlation coefficient between discrete wavelet components and the observed flood quantile for T=10 year and decomposed at five resolution level. The effective components are chosen based on correlation coefficient value. The positive correlation coefficient components are chosen as the effective components and become the input for LR and WGMDH. The effective components for $x_1$ are A4, D1, D2, D3, D4 and D5. The effective component for $x_2$ is only D2. The effective components for $x_3$ are A4, D1, D2, D3, D4 and D5 and for $x_4$ there are no effective components because all of the correlation coefficients negative. Then for $x_5$ the effective components are still D1 and D2. The WGMDH model is a GMDH model that uses effective decompose components obtained using DWT on original data. For the WGMDH model inputs, the input variables are decomposed into a certain number of components (D’s) by Mallat DWT algorithm (Mallat et al., 1989). Each component plays a different role in the original data and the behavior of each component is distinct. The main motivation of using the wavelet decomposition is the easy analysis of the obtained series. Daubechies wavelets were implemented because the Daubechies wavelets were suitable for the detection of a sudden signal change at the time or frequency domain. Daubechies wavelets exhibit good trade-off between parsimony and information richness. The data set used was transformed using logarithmic transformation for WGMDH modeling. The input variables were decomposed at two, three, four and five resolu-
Streamflow estimation at ungauged site

There are various number resolution levels applied in GMDH because there are no method to determine which resolution level is suitable. The WGMDH model uses the same input as WR model. In WGMDH, choosing the effective components is also based on correlation coefficient.

The correlation coefficient between decompose components and observed are shown from Table 2. Table 3 shows the comparison between models in estimating flood quantile for $T=10$ years, $T=50$ years, $T=100$ years and comparison are based on RMSE, MAE and CE. For $T=10$ years, the best estimation for WGMDH is when the input variable decompose at three level resolution and for $T=50$ years, the best estimation for WGMDH is at four level resolution. Meanwhile for $T=100$ years, the best estimation produce is at three level resolution. This result show that different level decomposition do affected the estimation accuracy of model. The result shown in Table 3 indicated that WGMDH model outperformed other three models in estimating flood quantile at the ungauged site for $T=10$ years, $T=50$ years and $T=100$ years.

5 Conclusion

The accuracy of the WGMDH model has been investigated to estimate flood quantile at ungauged site. The WGMDH model was developed by combining two methods; DWT and GMDH. Next, WGMDH model was compared with LR, WR and GMDH models. In addition to this study, performance of WGMDH and WR were compared at various resolution levels of DWT were applied. The result has shown the resolution level or decomposition level does affect the performance of WGMDH model. All four models were evaluated using RMSE, MAE, CE and $r$. Table 6.7 and Table 6.8 show the performance of WR and WGMDH at two, three, four and five resolution levels. On top of that, the WGMDH has significantly increased the accuracy of the GMDH model. In estimating the flood quantile at the ungauged site for $T=10$ years, $T=50$ years and $T=100$ years, WGMDH has outperformed LR, WR and GMDH model. The overall comparison results suggested that WGMDH approach may provide an alternative for LR in estimating flood quantile at ungauged sites.

Acknowledgements

This financial support provided by Universiti Teknologi Malaysia under FRGS Grant (R.J130000.7828.4F399) is gratefully acknowledged.

References


Table 1: Descriptive statistics of hydrologic, physiographical and meteorological variables

<table>
<thead>
<tr>
<th>Variables</th>
<th>Min</th>
<th>Mean</th>
<th>Max</th>
<th>STD</th>
</tr>
</thead>
<tbody>
<tr>
<td>AREA [km²]</td>
<td>30</td>
<td>1787.05</td>
<td>19000</td>
<td>3676.28</td>
</tr>
<tr>
<td>ELV [m]</td>
<td>4</td>
<td>99.49</td>
<td>1450</td>
<td>249.99</td>
</tr>
<tr>
<td>LDP [m]</td>
<td>3800</td>
<td>38457.97</td>
<td>280000</td>
<td>59553.88</td>
</tr>
<tr>
<td>SLP [%]</td>
<td>0.01</td>
<td>0.40</td>
<td>2.56</td>
<td>0.50</td>
</tr>
<tr>
<td>AMR [mm]</td>
<td>314.30</td>
<td>2099.75</td>
<td>4678.70</td>
<td>717.26</td>
</tr>
<tr>
<td>$Q_{10}$ [m³/s]</td>
<td>12.87</td>
<td>716.15</td>
<td>7256.76</td>
<td>1451.10</td>
</tr>
<tr>
<td>$Q_{50}$ [m³/s]</td>
<td>29.54</td>
<td>1043.45</td>
<td>10089.80</td>
<td>2029.01</td>
</tr>
<tr>
<td>$Q_{100}$ [m³/s]</td>
<td>43.82</td>
<td>1194.17</td>
<td>11218.89</td>
<td>2270.77</td>
</tr>
</tbody>
</table>

Table 3: Comparison between models in estimating flood quantile for $T=10$ years, $T=50$ years and $T=100$ years

<table>
<thead>
<tr>
<th>Models</th>
<th>$T=10$ year</th>
<th></th>
<th></th>
<th>$T=50$ year</th>
<th></th>
<th></th>
<th>$T=100$ year</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>MAE</td>
<td>CE</td>
<td>RMSE</td>
<td>MAE</td>
<td>CE</td>
<td>RMSE</td>
<td>MAE</td>
<td>CE</td>
</tr>
<tr>
<td>LR</td>
<td>784.1362</td>
<td>386.5151</td>
<td>0.7000</td>
<td>1197.5551</td>
<td>608.6824</td>
<td>0.6420</td>
<td>1403.8435</td>
<td>719.7299</td>
<td>0.6072</td>
</tr>
<tr>
<td>GMDH</td>
<td>522.4403</td>
<td>168.0147</td>
<td>0.8668</td>
<td>1056.4881</td>
<td>407.5615</td>
<td>0.7214</td>
<td>1321.2575</td>
<td>564.6853</td>
<td>0.6520</td>
</tr>
<tr>
<td>WR</td>
<td>967.6145</td>
<td>442.4139</td>
<td>0.5431</td>
<td>1762.7770</td>
<td>779.0667</td>
<td>0.2243</td>
<td>1700.9936</td>
<td>809.6467</td>
<td>0.4232</td>
</tr>
<tr>
<td>WGMDH</td>
<td>276.8987</td>
<td>105.5519</td>
<td>0.9626</td>
<td>343.0564</td>
<td>137.6642</td>
<td>0.9706</td>
<td>444.0493</td>
<td>154.5692</td>
<td>0.9607</td>
</tr>
</tbody>
</table>

Table 2: The correlation coefficient between discrete wavelet components and the observed flood quantile for $T=10$ year for decomposition at five resolution levels

<table>
<thead>
<tr>
<th>Discrete Wavelet Components</th>
<th>$x_{1}/Q_{10}$</th>
<th>$x_{2}/Q_{10}$</th>
<th>$x_{3}/Q_{10}$</th>
<th>$x_{4}/Q_{10}$</th>
<th>$x_{5}/Q_{10}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>0.4601</td>
<td>-0.0096</td>
<td>0.4234</td>
<td>-0.3468</td>
<td>0.1243</td>
</tr>
<tr>
<td>D2</td>
<td>0.3368</td>
<td>0.1465</td>
<td>0.3195</td>
<td>-0.1028</td>
<td>0.0809</td>
</tr>
<tr>
<td>D3</td>
<td>0.1816</td>
<td>-0.0381</td>
<td>0.1409</td>
<td>-0.1280</td>
<td>-0.0169</td>
</tr>
<tr>
<td>D4</td>
<td>0.4442</td>
<td>-0.1779</td>
<td>0.4098</td>
<td>-0.0902</td>
<td>-0.2675</td>
</tr>
<tr>
<td>D5</td>
<td>0.1105</td>
<td>-0.1388</td>
<td>0.3004</td>
<td>-0.0195</td>
<td>-0.0196</td>
</tr>
<tr>
<td>A4</td>
<td>0.3574</td>
<td>-0.0501</td>
<td>0.3185</td>
<td>-0.3348</td>
<td>-0.3049</td>
</tr>
</tbody>
</table>
Figure 1: The WGMDH model structure

Figure 2: The decomposed discrete wavelet components of catchment area for 2-level decomposition

Received: February 17, 2014