Predictive modelling of the stage–discharge relationship using Gene-Expression Programming

Prashant Birbal, Hazi Azamathulla, Lee Leon, Vikram Kumar and Jerome Hosein

ABSTRACT

Modelling the hydrologic processes is an essential tool for the efficient management of water resource systems. Therefore, researchers are consistently developing and improving various predictive/forecasting techniques to accurately represent a river’s attributes, even though traditional methods are available. This paper presents the Gene-Expression Programming (GEP) modelling technique to accurately model the stage–discharge relationship for the Arouca River in Trinidad and Tobago using only low flow data. The proposed method uses the stage and associated discharge measurements at one cross-section of the Arouca River. These measurements were used to train the GEP model. The results of the GEP model were also compared to the traditional method of the Stage–Discharge Rating Curve (SRC). Four statistical parameters namely the Pearson’s Correlation Coefficient (R), Root Mean Square Error (RMSE), Mean Absolute Relative Error (MARE) and Nash–Sutcliffe Efficiency (NSE) were used to evaluate the performance of the GEP model and the SRC method. Overall, the GEP model performed exceptionally well with an R² of 0.990, RMSE of 0.104, MARE of 0.076 and NSE of 0.957.

Key words | Gene-Expression Programming, low flows, modelling, stage–discharge

HIGHLIGHTS

- The stage–discharge relationship for the Arouca River in Trinidad and Tobago was modelled using GEP via the GeneXPro software, using only low flow data.
- The stage–discharge relationship for the Arouca River in Trinidad and Tobago was modelled with the SRC method using only low flow data.
- The performance GEP and SRC techniques were analysed using previous research and statistical parameters such as R², RSME, MARE and NSE.
INTRODUCTION

The most common natural disasters encountered in Trinidad are flooding and drought. It is on the heels of the devastating impacts of these disasters that one begins to understand that accurate information about the characteristics of rivers is important for flood forecasting and drought preparation. There has also been a growing need for increased water supply, irrigation, energy and the need to mitigate flooding as a result of urbanisation (Torabi et al. 2015). In addition, various hydrologic applications such as water and sediment load estimation, water resource planning, operation and development, hydraulic and hydrologic modelling are important (Guven & Ali 2009). However, collecting data for discharge on a continuous basis is quite time consuming and costly, especially during large flood events. Therefore, an alternative approach to better manage water harvesting projects and mitigate flooding events would be to convert records of water stages into discharges using a stage–discharge relationship.

In order to develop a relationship between stage and discharge for a river, data for the two variables must be collected over a long period of time so that a better representation of the non-linear relationship can be established. The stage-discharge relationship is known as rating curve and it represents a graph of stage versus discharge (Londhe & Gauri 2015). In theory, this relationship can be used to predict/forecast discharge from stage values.

This Stage–Discharge Rating Curve (SRC) transforms the continuous stage data to a continuous record of stream discharge, but it is also used to transform model forecasted flow hydrographs into stage hydrographs (Schmidt & Yen 2001). These relations are typically developed empirically from periodic measurements of stage and discharge. Then the data are plotted versus the concurrent stage to define the rating curve for the stream (Braca 2008). According to Kim et al. (2016), in order to obtain a reliable SRC, numerous discharge data from the lowest to bank-full stage should be observed over a long period of time.

However, according to Petersen-Øverleir (2006), the stage–discharge relationship is affected by hysteresis of unsteady flow. In reality, most river flows are unsteady and therefore the flows under analysis will experience hysteresis. According to Kumar (2011), this phenomenon is more predominant in flat sloped streams.

Additionally, Azamathulla et al. (2011) states that discharge at a section in a river is not a function of stage alone. The article further indicates that the discharge of a river is dependent on several factors such as channel geometry, bed roughness and longitudinal roughness, but quantification of all these factors is impractical. Hence, it becomes more viable to establish an accurate relationship between stage and discharge since it is considered the epitome of all the characteristics of a certain reach of a stream.

Moreover, Zakwan et al. (2017) crucially stated that the reliability of discharge predicted from the rating curve depends on the accuracy of the method used in developing the stage–discharge relationship. It should also be noted that low flow extrapolation from SRCs are not very accurate, in the majority of studies (Braca 2008).

With respect to the limitations of the SRC, Fenton (2001) observed that for small flow and stages, the data points are artificially separated, and small differences physically become large differences with a tendency to attach more importance to these points in the least squares procedure than what is really the case. Additionally, Sivapragasam & Muttil (2005) found that a river's cross-section is very dynamic, and the rating curve will change year to year. It is difficult to extrapolate the best fit on high stage discharges with certainty. Furthermore, the determination of stage corresponding to zero discharge is highly subjective Sivapragasam & Muttil (2005).

Apart from the traditional SRC method, several other artificial intelligent techniques have been widely used in flood forecasting (Parsaie 2016). The more widely used techniques, amongst many others, stem from Support Vector Machine (SVM), Artificial Neural Network (ANN), M5 Model Trees and GEP (Bhattacharya & Solomatine 2005; Ryan & Hibler 2011; Norouzi et al. 2019). This study will focus on the GEP model which has proven to outperform typical regression methods and ANN models (Azamathulla et al. 2011).
Gene-Expression Programming (GEP) is an extension and combination of Genetic Programming (GP) and Genetic Algorithms (GA) used for solving complex real-world problems (Azamathulla et al. 2011). According to Ferreira (2001), in terms of the advantages of GEP, the chromosomes are simple entities: linear, compact, relatively small, and easy to manipulate genetically. Furthermore, Ferreira (2001) also states that the GEP models can produce an expression tree (ET) that illustrates empirical expressions which represents the relationship between the variables. These are two main advantages of the GEP over many of the existing modelling programmes.

Additionally, Guven & Ali (2009) developed the stage–discharge relationships for the Schuylkill River at Berne using GEP. Other recent works by Azamathulla & Aminuddin (2011) on the prediction of longitudinal dispersion coefficients in streams using GP, Azamathulla et al. (2010) on bridge pier scour and Tayfur (2017) on sediment transport confirm the suitability of applying GEP for water resource engineering studies. Additionally, Barzegar et al. (2016) stated that the GEP modelling technique is one of the models that addresses nonlinearity of data and it can accurately model this relationship.

Kasiviswanathan et al. (2016) supported the performance of the GEP model by stating that the advantage of using GEP lies in optimizing the model structure and parameters simultaneously, whereas most of the other data driven models uses predefined model structure for optimizing the model parameters. Another research by Azamathulla et al. (2011) illustrated the development of mathematical models for the estimation of stage–discharge relationships based on the GP and GEP techniques for the Pahang River in Malaysia. The performance of the GEP model outperformed the conventional stage rating curve, regression techniques, ANN and GP.

However, Ryan & Hibler (2011) critiques the GEP by stating that it suffers from highly disruptive genetic operators and added complexity in those operators from the need to be aware of the eventual ET. Dealing with complex relationship with several numerical operators, the ET representation scheme makes the algorithm difficult to implement and easily trapped by the explosively growing tree size during evolution. In addition, Zhong et al. (2017) reviewed the limitations of GEP and concluded that the GEP is lacking in solving complicated problems that humans face today. These include the massive GEP framework for solving problems in a large data environment and for handling multiple tasks simultaneously. Also, rigorous theoretical analysis of GEP is required for providing deeper insights of the GEP search process, such as the proof of convergence, time complexity analysis, convergence speed estimation, and analysis of evolution efficiencies of operators (Zhong et al. 2017).

Nevertheless, this paper aims to accurately represent the stage–discharge relationship for the Arouca River in Trinidad and Tobago using only low flow data. In so doing, a GEP model would be developed to establish the stage–discharge relationship and it will be compared to the traditional SRC method. The performance GEP technique will be analysed using previous research and statistical parameters such as R², RSME, MARE and NSE.

METHODOLOGY

Data description

The data set used in this study was obtained from the Water and Sewage Authority – Water Resource Agency (WASA-WRA) and from personal field readings taken at the location highlighted in Figure 1. The discharges were derived from measurements of the cross-sectional area of the stream and the mean velocity within each section was found and summed together to derive the mean velocity at that cross-section of the Arouca River (WMO 1980). The WRA had taken several readings over a span of 10 years, from 2010 to 2019, some of which were used to train the GEP model. In total, 23 valid measurements were used for both models, due to the vast quantity of unworkable missing data. These measurements/readings were separated as 70% data (16 measurements) for calibration and next 30% data (7 measurements) for validation/testing purpose for all the models. It should also be noted that there was only one input parameter (stage) and respectively one output parameter (discharge) for all 23 measurements that were used in the models.
Stage–Discharge Rating Curve (SRC)

The SRC approach is used to convert the limited records of stage and discharge into a graphical relationship that can be extrapolated to predict the discharges of a river at a given water level (stage). In order to establish the stage–discharge relationship, historical stage–discharge data of the gauging site is used to plot the graph between the observed stage ($S$) and observed discharge ($Q$). Once the stage–discharge relationship is established it is then used to convert the

![Location map of the Arouca River](adapted from Ministry of National Security – Office of Disaster and Preparedness Management of Trinidad and Tobago 2013).
records of water level into discharge, thereby eliminating time consuming, costly and sometimes the impractical exercise of continuous discharge measurement (Muzzammil et al. 2015).

The observations of stage (water level) versus discharge covering the range of stages expected at the river gauging section were plotted on a graph. A best-fit curve was then drawn connecting these plotted points. This line is known as the simple rating curve. After the creation of the simple rating curve, this curve must now be fitted using the power type equation shown below:

\[ Q = C (S + a)^n \]  

where \( Q \) – Discharge in cubic meters per second (m\(^3\)/s\(^{-1}\)), \( S \) – Gauge Height in meters (m), \( a \) – Stage at zero flow (datum correction) in meters (m), \( C \) and \( n \) – Constants.

When Equation (1) is transformed into a logarithmic form (Equation (6)), theoretically a straight-line relationship would result when the zero-flow condition (datum correction) is catered for. There may be instances whereby two or more straight lines may be required to fix the data and therefore the location of break points should be carefully established:

\[ \log Q = n \log (S - a) + \log C \]  

where \( \log \) – Y axis (Discharge), \( n \) – Gradient, \( \log (S - a) \) – X axis (Stage), \( \log C \) – Intercept on Y axis

The datum correction was obtained using the arithmetic method also known as the Johnson method (WMO 1980). This arithmetic method involved the use of Equation (3) below, given certain condition were met. These conditions were as follows.

Condition 1 \( ... \frac{Q_1}{Q_2} = \frac{Q_2}{Q_3} ... \) where \( Q_1 \) is discharge from the lowered range, \( Q_3 \) is discharge from the upper range

Condition 2 \( ... 2Q_2 = Q_1Q_3. \) The corresponding gauge heights for these discharges (\( Q_1, Q_2, Q_3 \)) from the plot are \( S_1, S_2 \) and \( S_3 \) respectively. Therefore, by substitution and reduction of \( Q_3 = c (S_1 - a) \) in condition 1 above, Equation (7) can be generated:

\[ a = \frac{[(S_1 \times S_3) - S_2^2]}{S_1 + S_2 - (2 \times S_2)} \]  

After the datum correction was obtained, the constants ‘\( C \)’ and ‘\( n \)’ were obtained using the least squares method as indicated from Equations (4) and (5) below. It should also be noted that the actual value of ‘\( C \)’ is the log inverse of the value obtained from Equation (4):

\[ \log C = \frac{\left( \sum_{k=1}^{n} Q_k - \sum_{k=1}^{n} S_k \right)}{N} \]  

\[ n = \left( \frac{\left( \sum_{k=1}^{n} Q_k \times S_k \right) - \left( \sum_{k=1}^{n} Q_k \times \sum_{k=1}^{n} S_k \right)}{\left( \sum_{k=1}^{n} Q_k^2 \right) - \left( \sum_{k=1}^{n} Q_k \right)^2} \right) \]  

where \( N \) – Total number of observed stage-discharge data, \( Q_k \) – Estimated discharge, \( S_k \) – Stage.

**Gene-Expression Programming (GEP)**

The initial step to develop the GEP model would be to train the model. Sixteen (16) random data points were selected to train the GEP model and the remaining seven (7) data points were used to validate the model. Based on the available data, a population size of 23 chromosomes was selected. After obtaining the population size, the next step would be to choose the fitness function. The fitness, \( f_i \), of the program, \( i \), is measured by the equation below:

\[ f_i = 1000 \cdot \frac{1}{1 + E_i} \]  

Where \( \ldots E_i = P_{ij} - O_{ij} \)

where \( P_{ij} \) is the value predicted by the individual chromosome i, for fitness case j and \( O_{ij} \) is the observed value for fitness case j. For \( P_{ij} = O_{ij} \) means that \( E_{ij} = 0 \) representing a perfect solution with no error. The advantage of this kind of fitness function is that the system can find the optimal solution by itself and the run will continue until the maximum fitness is achieved (Ferreira 2001).

Then the terminals and functions were chosen. The four basic arithmetic operators (\(+, -, \times, /\)) were used as the
functions, while the terminal selected was Stage (S). Although many more mathematical operators could have been used, the goal was to achieve a relatively simple expression to represent the Discharge (Q).

The next step involved determining the number of genes, head and tail length for each gene in a chromosome. To begin with, a single gene and two head lengths were used. Then, the number of genes and head lengths were increased one at a time during each run while monitoring the training and testing performances of each model. It was observed that more than three genes and a head length greater than eight did not significantly improve the training and testing performance of the GEP models. Thus, the head length, \( h = 7 \), and three genes per chromosome were considered for the GEP model. Since the maximum number of arguments per function is equal to two, giving \( n_{\text{max}} = 2 \), the tail length would be calculated by the following relation:

\[
t = h \left( (n - 1) + 1 \right) = 7 \left( (2 - 1) + 1 \right)
\]

Giving tail length, \( t = 8 \)

It should also be noted that all genetic operators such as mutation, inversion, transposition, Insertion Sequence (IS), Root Insertion Sequence (RIS), gene transposition, recombination or crossover (1-point, 2-point and gene recombination), and specific genetic operators were used. Two one-point mutations with mutation rate of 0.0051 were also used. Lastly, the linking function used to join the sub-expression trees was the addition operator (+). Table 1 below summaries the final parameters of the GEP model.

The GeneXproTools software containing the Gene-Expression Programming model was run for a number of generations and was stopped when there was no improvement in the fitness function value and coefficient of determination. Figure 2 on the following page illustrates the summaries methodology of the GEP modelling technique.

After the model was created, several statistical techniques such as the Correlation Coefficient (R), the Root Mean Square Error (RMSE), the Mean Absolute Relative Error (MARE) and the Nash–Sutcliffe Efficiency (NSE) were all used to analyse the performance of the GEP model in comparison with the SRC method.

### Table 1 | Summarised GEP model parameters

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population size</td>
<td>23</td>
</tr>
<tr>
<td>Set of function</td>
<td>+, −, *, /</td>
</tr>
<tr>
<td>Set of terminals</td>
<td>S</td>
</tr>
<tr>
<td>Random numerical constant (RNC)</td>
<td>5</td>
</tr>
<tr>
<td>RNC type</td>
<td>Floating point</td>
</tr>
<tr>
<td>Range of RNC</td>
<td>[−10, 10]</td>
</tr>
<tr>
<td>Length of head</td>
<td>8</td>
</tr>
<tr>
<td>Number of genes</td>
<td>3</td>
</tr>
<tr>
<td>Linking function</td>
<td>+</td>
</tr>
<tr>
<td>Fitness function</td>
<td>RMSE</td>
</tr>
<tr>
<td>Rate of mutation</td>
<td>0.0051</td>
</tr>
<tr>
<td>Rate of inversion</td>
<td>0.1</td>
</tr>
<tr>
<td>Rate of IS transposition</td>
<td>0.1</td>
</tr>
<tr>
<td>Rate of RIS transposition</td>
<td>0.1</td>
</tr>
<tr>
<td>Rate of Gene transposition</td>
<td>0.1</td>
</tr>
<tr>
<td>Rate of One-point recombination</td>
<td>0.3</td>
</tr>
<tr>
<td>Rate of Two-point recombination</td>
<td>0.3</td>
</tr>
<tr>
<td>Rate of Gene recombination</td>
<td>0.1</td>
</tr>
<tr>
<td>Rate of Dc-specific mutation</td>
<td>0.0051</td>
</tr>
<tr>
<td>Rate of Dc-specific inversion</td>
<td>0.1</td>
</tr>
<tr>
<td>Rate of Dc-specific IS transposition</td>
<td>0.1</td>
</tr>
</tbody>
</table>

### Statistical parameters

The following parameters listed below were used to evaluate and compare the performance of the models.

### Pearson’s Correlation Coefficient (R)

Equation (8) represents the formula used to calculate the correlation coefficient for the relationship between stage and discharge for each modelling technique:

\[
R = \frac{n \sum_{k=1}^{n} (Q_k \times S_k) - \left( \sum_{k=1}^{n} Q_k \right) \left( \sum_{k=1}^{n} S_k \right)}{\left[ \left( \sum_{k=1}^{n} Q_k^2 \right) - \left( \sum_{k=1}^{n} Q_k \right)^2 \right] \times \left[ \left( \sum_{k=1}^{n} S_k^2 \right) - \left( \sum_{k=1}^{n} S_k \right)^2 \right]^{1/2}}
\]

(8)
where; Q – Discharge, S – stage, n – number of data points.

**Root Mean Square Error (RMSE)**

The RMSE measures how much error there is between the modelled value and the observed value:

\[
RMSE = \sqrt{\frac{1}{n} \sum_{k=1}^{n} (Q_{obs,k} - Q_{model,k})^2}
\]  

(9)

where \(Q_{obs,k}\) – Actual (Observed) Discharge, \(Q_{model,k}\) – Predicted Discharge from model, \(n\) – number of data points.

**Mean Absolute Relative Error (MARE)**

MARE represents the average of the mean residual errors to errors produced by a model. The formula below illustrates the method used to calculate the MARE for each modelling technique:

\[
MARE = \frac{1}{n} \times \sum_{k=1}^{n} \left| \frac{Q_{obs,k} - Q_{model,k}}{Q_{obs,k}} \right|
\]  

(10)

where \(Q_{obs,k}\) – Actual (Observed) Discharge, \(Q_{model,k}\) – Predicted Discharge from model, \(n\) – number of data points.

**Nash–Sutcliffe Efficiency (NSE)**

The NSE indicates how well the plot of observed versus modelled data fits. In addition, if the NSE value is less
than ‘0’, this would indicate that the observed mean is a better predictor than the model:

\[ NSE = 1 - \left( \frac{\sum_{k=1}^{n} (Q_{\text{obs},k} - Q_{\text{model,k}})^2}{\sum_{k=1}^{n} (Q_{\text{obs},k} - Q_{\text{mean,obs}})^2} \right) \]  

(11)

where \(Q_{\text{obs},k}\) – Actual (Observed) Discharge, \(Q_{\text{model,k}}\) – Predicted Discharge from model, \(Q_{\text{mean,obs}}\) – Average Actual (Observed) Discharge.

RESULTS AND DISCUSSION

Stage–Discharge Rating Curve (SRC)

A graph of logarithmic discharge (log Q) vs logarithmic stage (log S) was plotted as shown in Figure 3.

Then by using the method of least squares regression, the constants ‘n’ and ‘C’ were calculated. Hence, the best fit curve formula can be represented as shown below:

\[ Q = C (h + a)^n \]

\[ Q = 1.8187 (h + 0.046)^{1.5281} \]

(12)

There were no instances whereby two or more straight lines were required to fix the data and therefore the location of break points were not required. This may be due to the consistently low flows of the rivers. Hence the stage–discharge relationship for the Arouca River can be precisely represented from the formula above.

Gene-Expression Programming (GEP)

The GEP model would have produced an ET to represent the mathematical relationship between the variables as one of its outputs. The ET and its mathematical representation for the Arouca River is shown in Figure 4.

The numeral constants for each gene were found within the GEP model itself. In summary, the mathematical representation of each gene was developed, and the overall equation used to represent the stage–discharge relationship.
for the Arouca River is shown as in Equation (13) below:

\[ Q = S^3 + 0.2853S^2 + S \] (13)

**Statistical analysis**

Figure 5 indicates that both the SRC and the GEP model were relatively close to the actual stage–discharge values. Both the SRC and GEP began by underestimating the actual relationship but they both gradually crept up to the curve. The GEP was observed to be the one sticking closer to the actual values as the values increased, whereas the SRC tended to drift away and further underestimated the relationship. From general inspection of Figure 5, the GEP model illustrated the tendency to better estimate the actual stage–discharge relationship in comparison with the SRC method.

As indicated in Table 2, it can be quite clearly observed that each model produced exceptional results with respect to the correlation coefficient \( R \). Both models had an \( R^2 \) value of over 0.99, with the SRC having a slight advantage over the GEP. A superior model could not be determined based solely on \( R \) as both had a similar value, close to perfection. In order to determine which model performed the better of the two, the remaining statistical parameters had to be investigated.

The statistical parameters that relate to errors were the RSME and the MARE. The SRC produced high RSME and MARE values in relation to the low discharge values used while the GEP produced moderate RSME and MARE values. Hence if a model were to be chosen based
on RMSE and MARE solely, given that the R values between the model were relatively the same, then the GEP model would have been chosen as the superior model.

In relation to the NSE parameter that relates how well the plot of observed versus simulated data fits, the GEP model produced a NSE value very close to ‘1’, with ‘1’ being perfect match, in comparison with a negative NSE for the SRC that indicated that the observed mean is a better predictor than the SRC method. Therefore, the superior model was the GEP model as it produced the best RMSE, MARE and NSE results with a very satisfac-
tory R value.

Finally, to solidify that the GEP was the best modelling technique, two graphs were plotted as shown in Figure 6. Each graph illustrated the actual discharge versus the calculated discharge for the respective model used. The graph which produced the $R^2$ value closest to 1 would be the best performing model as it would indicate that the model produced values closer to the actual (observed) values. The SRC and GEP graphs (Figure 6) produced correlation coefficients ($R^2$) of 0.989 and 0.995 respectively. As evident in the statistical parameters and similarly with Figure 5, the GEP model outperformed the SRC method. Hence the GEP model was the superior model used to represent the stage-discharge relationship of low flows for the Arouca River.

**CONCLUSION**

The GEP model performance was superior to the SRC method in terms of the statistical parameters, scatter plots and previous studies. The GEP model produced the smallest errors, possessed the highest efficiencies, and illustrated an excellent correlation between the variables. The Arouca River also produced a correlation coefficient of 0.995 when the GEP model was used. Furthermore, the past studies indicated that GEP not only has advantages over SRC, but it is a better predictive model than those that were traditionally used. Therefore, this GEP model can be used to better estimate and manage water resources from low flows for the Arouca River.

With respect to limitations, only one cross-section of the entire river was used and many consecutive data points throughout the years were missing. Therefore, is it unknown whether these missing data points were the lower and upper
bounds of the discharge and stage measurements; that may have been due to insufficient flows or floods. For future research, in the events of insufficient or no flows, the time frame in which these flows occur should be recorded to better predict drought events so that improved water conservation techniques can be implemented.

It should also be noted that a river’s cross-section changes through time. Therefore, to improve the accuracy and precision of the stage–discharge relationship the cross-sectional location at which the reading will be taken should be monitored in tandem with the necessary measurements. This can result in aggradation and/or degradation which can affect the true representation of the stage–discharge relationship over time.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

REFERENCES


Kumar, V. 2011 Hysteresis. In: Encyclopedia of Snow, Ice and Glaciers. Encyclopedia of Earth Sciences Series (V. P. Singh,


