Rainfall-Runoff Modelling Using Genetic Programming

[1]Jayawardena, A. W., [2]N. Muttil and [3]T.M.K.G. Fernando

[1] University of Hong Kong, [2] The Hong Kong Polytechnic University, [3]University of Adelaide, Australia
E-Mail: hrecjaw@hkucc.edu.hk

Keywords: *Rainfall-runoff modelling; Data-driven models; Evolutionary algorithms; Genetic programming.*

EXTENDED ABSTRACT

The problem of accurately determining river flows from rainfall, evaporation and other factors, occupies an important place in hydrology. The rainfall-runoff process is believed to be highly non-linear, time varying, spatially distributed and not easily described by simple models. Practitioners in water resources have embraced data-driven modelling approaches enthusiastically, as they are perceived to overcome some of the difficulties associated with physics-based approaches. Such approaches have proved to be an effective and efficient way to model the rainfall-runoff process in situations where enough data on physical characteristics of catchment is not available or when it is essential to predict the flow in the shortest possible time to enable sufficient time for notification and evacuation procedures.

In the recent past, an evolutionary based data-driven modelling approach, genetic programming (GP) has been used for rainfall-runoff modelling. In this study, GP has been applied for predicting the runoff from three catchments – a small steep-sloped catchment in Hong Kong (Hok Tau catchment) and two relatively bigger catchments located in the southern part of China (Shanqiao and Shuntian catchments).

For the runoff predictions in Hok Tau catchment, the performance of the data-driven technique was not very satisfactory. This catchment, being a very steep-sloped catchment, has high peak discharge magnitudes with steep rising and recession limbs, which the GP models are unable to capture. This catchment being a small one with an area of about 5 km^2 has a time of concentration of about 30-45 minutes, but the time interval of the available data is one day, which seems to be another reason for GP's inability to capture the complex rainfall to runoff transformation on this catchment. Using a dataset of smaller time interval, the data-driven model should perform better.

A key advantage of GP as compared to traditional modelling approaches is that it does not assume any *a priori* functional form of the solution. For instance, in a typical regression method, the model structure is specified in advance (which is in general difficult to do) and the model coefficients are determined. For neural networks, the time consuming task of initially defining the network structure has to be undertaken and then the coefficients (weights) are found by the learning algorithm. On the other hand, in GP, the building blocks (the input and target variables and the function set) are defined initially, and the learning method subsequently finds both the optimal structure of the model and its coefficients.

Moreover, since GP evolves an equation or formula relating the input and output variables, a major advantage of the GP approach is its automatic ability to select input variables that contribute beneficially to the model and disregard those that do not. GP can thus reduce substantially the dimensionality of the input variables.

In GP, as in any data-driven prediction model, the selection of appropriate model inputs is extremely important. This is especially so when lagged input variables are also used. Inclusion of irrelevant inputs leads to poor model accuracy and creation of complex models, which are more difficult to interpret as compared to simpler ones. Thus, for the remaining two catchments, an attempt is made to use the evolutionary search capabilities of GP for selecting the significant input variables. These variables, indicated as significant by GP are then used as inputs for the actual predictions.

In contrast to the not so satisfactory performance by the GP models for predicting the runoff from Hok Tau catchment, their performance for the other two catchments is quite satisfactory, as the GP models are able to capture the peaks quite well and the goodness-of-fit measures are also acceptable. These results indicate that GP can be used as a viable alternative for rainfall-runoff modelling, and the analytical form of the evolved equations facilitate easy interpretation. In this study, the GP evolved models are used for selection of significant variables influencing the rainfall to runoff transformation.

1. INTRODUCTION

The use of rainfall-runoff models in the decision making process of water resources planning and management has become increasingly indispensable (Liong et al. 2002). Such models are used, for example, in the design and operation of hydraulic structures, for discharge forecasting and for evaluating possible changes taken place over the catchments due to urbanization. The development of rainfall-runoff models has gone through substantial changes since Sherman pioneered the unit hydrograph theory in 1932. The transformation of rainfall into runoff is a complex, non-linear, time and spatial varying process (Singh 1988). Accordingly, various models ranging from linear to non-linear, lumped to distributed have been developed to describe the transformation of rainfall hyetograph to discharge hydrograph.

Black box models like artificial neural networks (ANNs) have been proposed as efficient tools for modeling in hydrology. ANNs are supposed to possess the capability to reproduce the unknown relationship existing between a set of input and output variables (Chakraborty et al 1992; Jayawardena and Fernando 1998; Zhang and Govindaraju 2000).

In the recent past, an Evolutionary Algorithm (EA) based model, Genetic Programming (GP) has also been used to emulate the rainfall-runoff process (Liong et al 2002; Whigham and Crapper 2001; Savic et al. 1999) and has been shown to be a viable alternative to traditional rainfall-runoff models. GP has the advantage of providing inherent functional input-output relationships as compared to traditional black box models, which can offer some possible interpretations to the underlying process.

For a small, steep-sloped catchment in Hong Kong, it was found that the data-driven model poorly represented the rainfall-runoff process in general, and the prediction of peak discharge, in particular. To demonstrate the potential of GP as a viable data-driven rainfall-runoff model, it is successfully applied to two other catchments located in southern China. A brief overview of GP is first outlined in the following section, before providing the details of the analysis.

2. GENETIC PROGRAMMING

GP is a automatic programming technique for evolving computer programs to solve, or approximately solve, problems (Koza 1992). In engineering applications, GP is frequently applied to model structure identification problems. In such applications, GP is used to infer the underlying structure of either a natural or experimental process in order to model the process numerically.

GP is a member of the Evolutionary Algorithm (EA) family. EAs are based upon Darwin's natural selection theory of evolution where a population is progressively improved by selectively discarding the not-so-fit population and breeding new children from better populations. EAs work by defining a goal in the form of a quality criterion and then use this goal to measure and compare solution candidates in a stepwise refinement of a set of data structures and return an optimal or near-optimal solution after a number of generations. Evolutionary Strategies (ES), Genetic Algorithms (GA) and Evolutionary Programs (EP) are three early variations of evolutionary algorithms whereas GP is the most recent variant of EAs. These techniques have become extremely popular due to their success at searching complex non-linear spaces and their robustness in practical applications.

2.1. Basic Principles of Genetic Programming

The basic search strategy behind GP is a genetic algorithm (GA) (Goldberg 1989). GP differs from the traditional GA in that it typically operates on parse trees instead of bit strings. A parse tree is built up from a terminal set (the variables in the problem) and a function set (the basic operators used to form the function). An example of such a parse tree can be found in Figure 1. The *tree size* of this expression is 7, where *tree size* is the maximum *node depth* of a tree and *node depth* is the minimum number of nodes that must be traversed to get from the *root node* of the tree (see Figure 1) to the selected node.

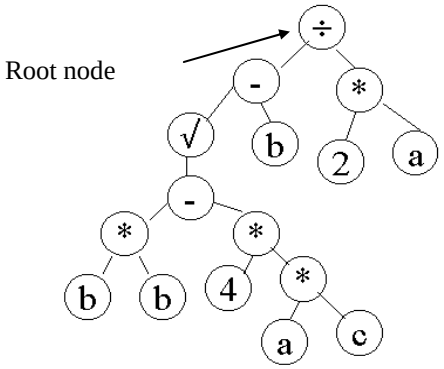

Figure 1. GP parse tree representing $\{\sqrt{(b^2-4ac)}-b\}/2a$

As a GA, GP proceeds by initially generating a population of random parse trees, calculate their

fitness - a measure of how well they solve the given problem - and subsequently selects the better parse trees for reproduction and variation to form a new population. This process of selection, reproduction and variation iterates until some stopping criterion is satisfied.

In the next sub-section, we discuss some advantages of GP, because of which we gave preference to GP, rather than traditional data-driven techniques such as ANN.

2.2. Unique Aspects of Genetic Programming

Traditional approaches like ANNs do have many attractive features, but they suffer from some limitations. The difficulty in choosing the optimal network architecture and time-consuming effort involved thereof is one key issue. In regression also, the model structure is decided in advance and the model coefficients are determined by the regression method. On the other hand, what makes GP unique is that it does not assume any functional form of the solution. GP can optimize both the structure of the model and its parameters.

Since GP evolves an equation relating the output and input variables, it has the advantage of providing inherent functional relationship explicitly over techniques like ANN. This gives the GP approach the automatic ability to select input variables that contribute beneficially to the model and to disregard those that do not. In the next section, GP is applied to predict runoff from the three catchments, two of which are located in southern China. GP evolved equations are also analyzed to select significant input variables.

3. MODELLING AND APPLICATION

In this study, continuous data from three catchments, one located in Hong Kong and other two in mainland China are considered. The first catchment is the Hok Tau catchment located in Hong Kong (Figure 2). The catchment covers an area of 5.22 km^2. It has steep slopes, displaying the sort of characteristics that are common for watersheds occurring in headwaters of streams or rivers. Five years (1993-1997) of daily rainfall and runoff data are used in the present study, out of which 4 years data (1993-1996) is used for calibration or training and 1 year data (1997) are considered as the validation or testing set.

The second catchment used is the Shanqiao catchment (Figure 2), which is an experimental catchment located in the western part of Pearl River (Zhujiang) Delta in southern China with a drainage area of 131 km^2. It drains into Tanjiang, a tributary of the Pearl River. The altitude within the basin varies between 8 and 637 m, the lowest being at the Shanqiao hydrological station. The available hydrological observations are from the published yearly hydrological reports. The daily rainfall and the mean daily discharges from 1985 to 1988 are used in this study, out of which, three years data from 1985 to 1987 are used to train GP models and 1988 data are used for testing.

The third catchment used is the Shuntian sub-catchment (Figure 2), which is located at the middle of the East River (Dongjiang) basin and has a drainage area of 1357 km^2. The East River is one of the three main tributaries of the Pearl River, the fourth largest river in China (in terms of drainage area). For Shuntian also, four years of daily rainfall and mean daily discharge data from 1975 to 1978 are taken, the first three years are used for training and the last year for testing.

3.1. Objective Function and Model Performance Criterion

The objective function used for the GP runs is the root mean squared error (RMSE).

The performance of the predictions is evaluated by two goodness-of-fit measures. They are the root mean square error (RMSE) and the coefficient of efficiency (E). Although the RMSE can indicate the relative performance of different models for the same lengths of calibration periods and validation periods, it cannot really indicate the performance of the models for different calibration record lengths. The E model efficiency criterion is a better choice in such a situation. However, the RMSE can give a quantitative indication of the model error in terms of a dimensioned quantity. Hence, the RMSE and the E are both presented.

3.2. The GP Model

In this study, GP is used to develop relationship between the future runoff at the catchment outlet, and rainfall and runoff data available up to the current time t. Mathematically the relationship may be expressed as:

$$Q_{t+\delta\Delta t} = f(R_t, R_{t-\Delta t, \ldots} R_{t-\omega\Delta t}, Q_t, Q_{t-\Delta t}, \ldots Q_{t-\omega\Delta t}) \quad (1)$$

where Q is the runoff (m^3/s), R is the rainfall intensity (mm/day), δ (with $\delta = 1, 2, \ldots$) refers to how far into the future the runoff prediction is desired, ω (with $\omega = 1, 2, \ldots$) implies how far back the recorded data in the time series are affecting the runoff prediction while Δt stands for time step discretization (i.e. time interval).

Figure 2. Location of the Hok Tau, Shanqiao and Shuntian catchments

In this study runoff forecasting for the three catchments are conducted for 1-day lead-time prediction. The GP is trained with the input data set containing variables as shown in Eq. (1). A value of ω (in Eq. (1)) is set to 1 for the Hok Tau catchment and to 4 for the other two catchments of Shanqiao and Shuntian. Thus, R_t, $R_{t-1\Delta t}$, Q_t, $Q_{t-1\Delta t}$ are taken as terminal set for Hok Tau and R_t, $R_{t-1\Delta t}$, $R_{t-2\Delta t}$, $R_{t-3\Delta t}$, $R_{t-4\Delta t}$, Q_t, $Q_{t-1\Delta t}$, $Q_{t-2\Delta t}$, $Q_{t-3\Delta t}$, $Q_{t-4\Delta t}$ for the other two catchments and $Q_{t+1\Delta t}$ ($\delta = 1$) is the target for all the three catchments. Since the number of time lags for the inputs affecting the 1-day lead-time runoff prediction would be higher for the larger catchments, a higher value of ω is used for Shanqiao and Shuntian catchments.

The GP software used in this study is GPKernel developed by DHI Water and Environment. For each run, GPKernel is run for 60 minutes on a Genuine Intel Pentium 4 PC with 1021 MB RAM.

3.3. Input Variable Selection using GP

The selection of significant input variables is extremely important, especially since lagged input variables are being used and how many lagged variables to use are to be found out. Since there are only 4 input variables for Hok Tau, input variable selection is not done for this catchment. GP equations were evolved to develop relationship between the 1-day ahead discharge and the 10 input variables for Shanqiao and Shuntian catchments. The GPKernel parameters used for the runs are presented in Table 1. The parameter "Maximum tree size" was restricted to 20, because with this limitation on size, the evolved equation contained only 4 to 6 variables. Thus, we are allowing the evolutionary process to select only about 4 to 6 variables from the total of 10 variables that are used as input.

For the GP runs, a function set consisting of the basic math operators (+, -, *, /) is used. Limiting the size of the GP equation and using a simple function set leads to parsimonious models, which are easy to interpret. Fifty GP equations were evolved using 50 GP runs with different initial seeds. Now, since GP has the ability to select input variables that contribute beneficially to the model and to disregard those that do not, it is expected that the GP evolved equations would contain the most significant of the 10 input

variables. The number of times each of the 10 input variables is selected in all the 50 evolved equations is presented in Table 2. The variables indicated as significant by GP for 1-day lead-time predictions are shaded in this table. The significance of past rainfall is consistent with the cause-effect relationship between past rainfall and future runoff. These significant variables are used as input for the GP predictions presented in the next sub-section.

Table 1. Values of GP Control Parameters

Parameter	Value
Maximum initial tree size	45
Maximum tree size	20
Tournament size	3
Crossover rate	1
Mutation rate	0.05
Population Size	1000
Elitism used	Yes

Table 2. Number of Input Variable Selections in 50 GP Runs

Input variables	Number of selections	
	Shanqiao	Shuntian
R_{t-4}	47	8
R_{t-3}	0	1
R_{t-2}	8	3
R_{t-1}	34	8
R_t	110	80
Q_{t-4}	4	11
Q_{t-3}	33	70
Q_{t-2}	1	10
Q_{t-1}	37	9
Q_t	24	132

3.4. GP Predictions and Results

The goodness-of-fit measures for the GP runs for the three catchments are presented in Table 3. The hydrographs are presented in Figure 3, 4 and 5 for the 3 catchments respectively. In the hydrographs, portions of data belonging to training and testing periods have been magnified for clarity. The hydrographs in Figure 3 for the Hok Tau catchment indicate that for both training and testing, the GP model consistently underpredicts the peak discharges, as can be seen in the magnified portions of hydrograph. For the Shanqiao catchment, the GP model trains well and is able to capture most of the peaks, but it does not perform as well on testing, with few underpredictions. This relatively better performance in training is also observed in the goodness-of-fit measures. For the Shuntian catchment, few phase errors are observed in the training hydrograph, but in testing, the model seems to capture the peaks without phase errors, as seen in magnified portion of the hydrograph.

3.5. Discussion on Results

For the Hok Tau catchment, since the data used has high peak discharge magnitudes and steep rise and recession limbs, it is observed that the GP model is unable to evolve an equation that trains well on the calibration data and the performance on testing data is also not encouraging. Moreover, the time interval of the data set is one day, i.e., only daily values of the data is available. The time of concentration of the Hok Tau catchment is about 45 minutes and it can be concluded that using daily data values may not be able to capture the complex rainfall to runoff transformation and data set of much smaller time interval (about 15-30 minutes) would be necessary. Although the performance of GP as runoff prediction tool was not encouraging for the Hok Tau catchment due to the above mentioned reasons, the performance of GP for the other two catchments from southern China were quite satisfactory.

Table 3. Goodness-of-fit measures for the catchments

	RMSE	E
Hok Tau catchment		
Training	0.61	0.52
Testing	1.08	0.16
Shanqiao catchment		
Training	2.72	0.82
Testing	3.03	0.70
Shuntian catchment		
Training	41.55	0.75
Testing	31.74	0.77

Other than the input variable selection, further work is underway to interpret the GP models. To demonstrate the simple and parsimonious nature of GP models, one of them evolved for the prediction of runoff from Shanqiao catchment is presented below in Eq. (2):

$$Q_{(t+1)} = 0.5\ Q_t + R_t\ [R_{t-4} + R_t - Q_t + 0.25 * (Q_{t-1}/Q_t) + Q_{t-3}] \quad (2)$$

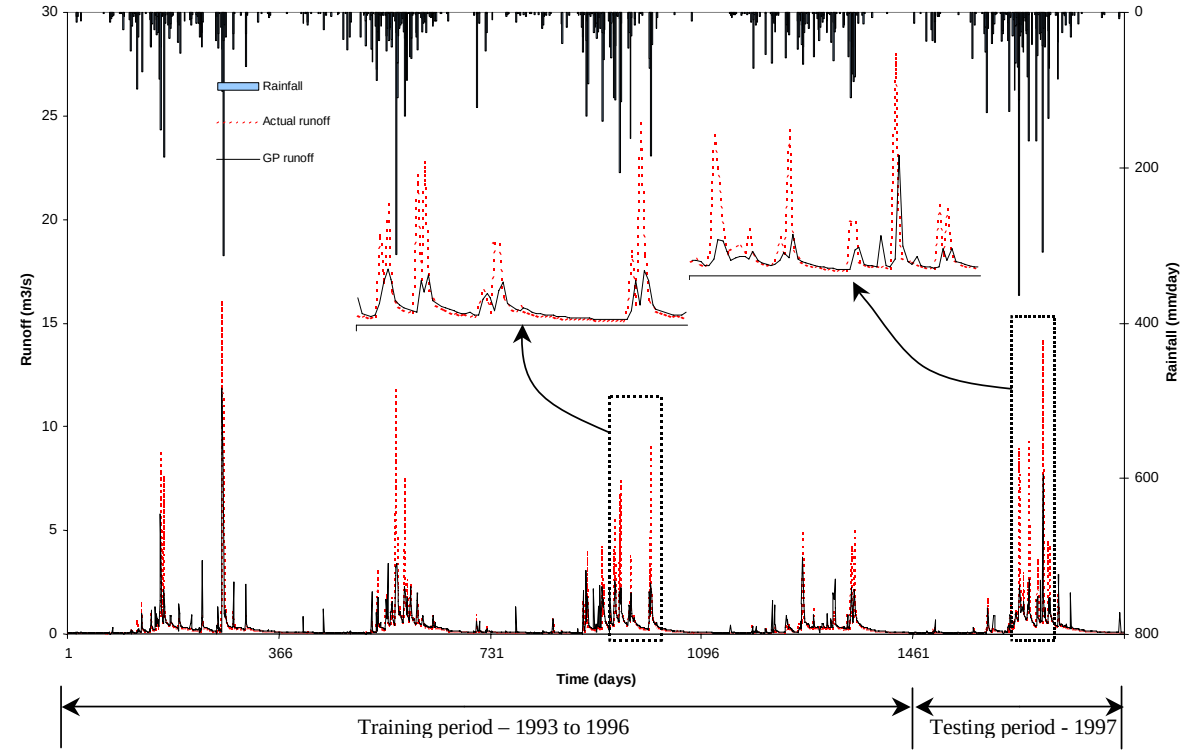

Figure 3. Hydrograph comparing actual and GP simulated runoff for Hok Tau catchment (Day 1 corresponds to 1st January 1993)

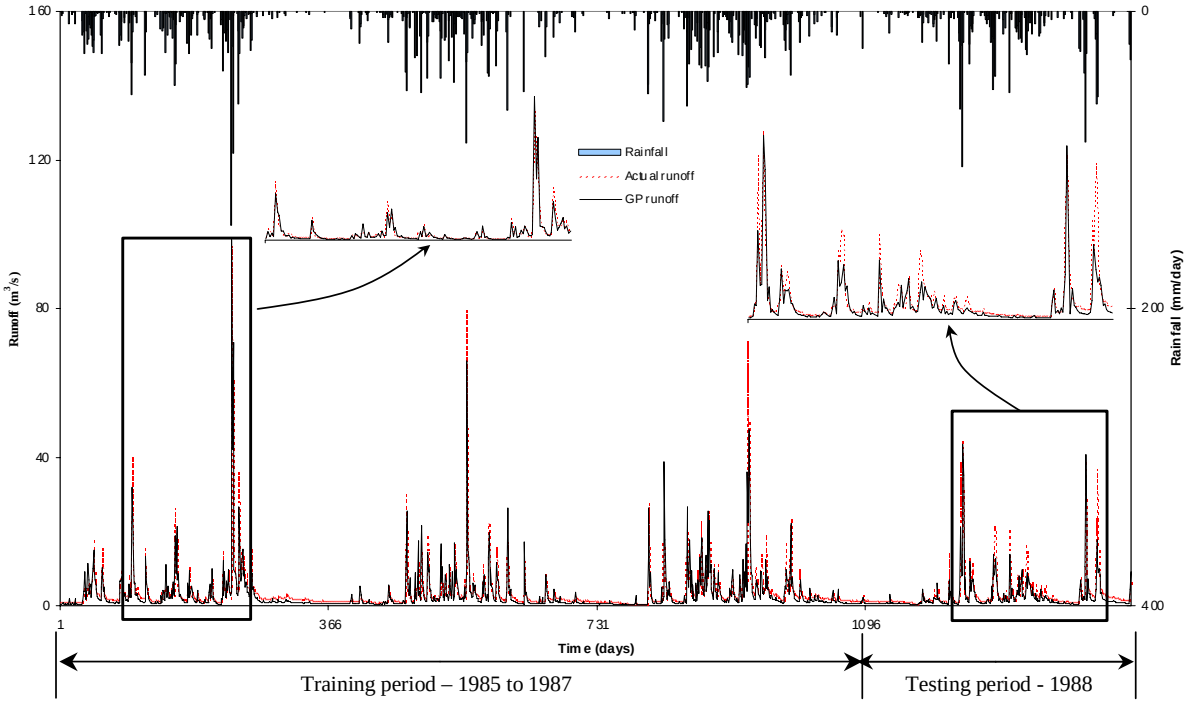

Figure 4. Hydrograph comparing actual and GP simulated runoff for Shanqiao catchment (Day 1 corresponds to 1st January 1985)

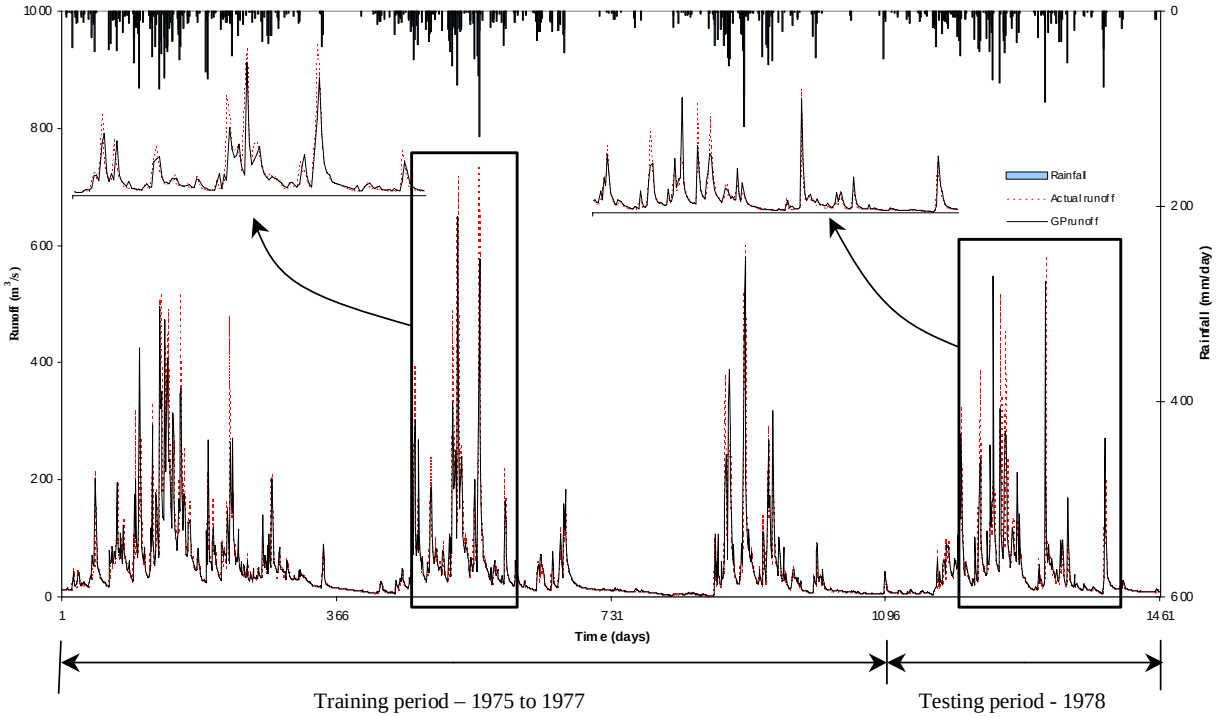

Figure 5. Hydrograph comparing actual and GP simulated runoff for Shuntian catchment (Day 1 corresponds to 1st January 1975)

4. CONCLUSIONS

In this study, the suitability of using the data-driven GP for modeling the rainfall-runoff process in three catchments was studied. It was found that for the Hok Tau catchment, GP could not represent the rainfall-runoff transformation process in general, and the peak discharge, in particular. The rainfall-runoff process is complicated in this catchment because of the steep slopes, resulting in high peak discharge values and pronounced rising and recession hydrograph limbs. Moreover the time interval of the available dataset was not suitable. Since the usefulness of GP as a rainfall-runoff model could not be shown on the above steep-sloped catchment, it was successfully employed for runoff prediction on two other catchments located in southern China. It was shown that simple and small GP models can be evolved, facilitating easy interpretation; in this study, they were used to select significant input variables for prediction.

5. REFERENCES

Chakraborty, K., Mehrotra, K., Mohan, C. K., and Ranka, S. (1992), Forecasting the behaviour of the multivariate time series using neural networks, *Neural Networks*, 5: 961-970.

Goldberg, D.E. (1989), Genetic Algorithms for Search, Optimization and Machine Learning. Addison-Wesley Publishing Co., Reading, Mass.

Jayawardena, A.W. and Fernando, D. A. K. (1998), Use of radial basis function type artificial neural networks for runoff simulation, *Computer-Aided Civil and Infrastructure Engineering* 13(2): 91-99.

Koza, J. (1992), Genetic Programming: On the Programming of Computers by Natural Selection. MIT Press, Cambridge, MA.

Liong, S. Y., Gautam, T. R., Khu, S. T., Babovic, V., and Muttil, N. (2002), Genetic Programming: A New Paradigm in Rainfall-Runoff Modelling, *Journal of American Water Resources Association* 38(3): 705-718.

Savic, D. A., Walters, G. A. and Davidson G. W. (1999), A genetic programming approach to rainfall-runoff modeling, *Water Resources Management* 13: 219-231.

Singh, V.P. (1988), Hydrologic Systems. Prentice Hall, Eaglewood Cliffs, NJ, USA.

Whigham, P. A. and Crapper, P. F. (2001), Modelling Rainfall-Runoff Relationships using Genetic Programming, *Mathematical and Computer Modelling* 33: 707-721.

Zhang, B., and Govindaraju, S. (2000), Prediction of watershed runoff using Bayesian concepts and modular neural networks, *Water Resources Research* 36 (3): 753-762.