Water Resour Manage (2017) 31:3975–3992
DOI 10.1007/s11269-017-1719-1

A Genetic Programming Approach to System Identification of Rainfall-Runoff Models

Jayashree Chadalawada[1] · **Vojtech Havlicek**[2] · **Vladan Babovic**[1]

Received: 5 December 2016 / Accepted: 16 May 2017 /
Published online: 4 July 2017

Abstract Advancements in data acquisition, storage and retrieval are progressing at an extraordinary rate, whereas the same in the field of knowledge extraction from data is yet to be accomplished. The challenges associated with hydrological datasets, including complexity, non-linearity and multicollinearity, motivate the use of machine learning to build hydrological models. Increasing global climate change and urbanization call for better understanding of altered rainfall-runoff processes. There is a requirement that models are intelligible estimates of underlying physics, coupling explanatory and predictive components, maintaining parsimony and accuracy. Genetic Programming, an evolutionary computation technique has been used for short-term prediction and forecast in the field of hydrology. Advancing data science in hydrology can be achieved by tapping the full potential of GP in defining an evolutionary flexible modelling framework that balances prior information, simulation accuracy and strategy for future uncertainty. As a preliminary step, GP is used in conjunction with a conceptual rainfall-runoff model to solve model configuration problem. Two datasets belonging to a tropical catchment of Singapore and a temperate catchment of South Island, New Zealand with contrasting characteristics are analyzed in this study. The results indicate that proposed approach successfully combines the merits of evolutionary algorithm and conceptual knowledge in the generation of optimal model structure and associated parameters to capture runoff dynamics of catchments.

Keywords Automatic model generation · Conceptual modelling · Genetic programming · Rainfall-Runoff models · System identification

✉ Jayashree Chadalawada
jayashree@u.nus.edu

[1] Department of Civil and Environmental Engineering, National University of Singapore, Block E1, #08-25, No 1 Engineering Drive 2, Singapore 117576, Singapore

[2] Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Kamycka 129, Prague 16521, Czech Republic

1 Introduction

1.1 Conceptual Rainfall-Runoff Models

Mathematical modelling of hydrological systems includes time series analysis and stochastic modelling, where the emphasis is on reproducing the characteristics of time series of a hydrological variable of interest. On one extreme, hydrological models can be purely emprical, black box models, such as Artificial Neural Networks, that match the input and output variables of the catchment system without modelling the internal structure of underlying physical processes. On the other extreme are deterministic models involving complex systems of equations based on physical laws and theoretical concepts governing the hydrological processes (Refsgaard and Abbott 1996). Between the two extremes one finds conceptual models that represent structure based on simple mathematical elements, such as linear or nonlinear reservoirs and channels that model processes within the basin in an approximate way (Charizopoulos and Psilovikos 2016). The controlling ideas of the conceptual models are based on basic requirements such as the water balance. For example, the mass balance of soil moisture storage is one of the building blocks of most conceptual water balance models. Such models use a limited number of parameters compared to deterministic models which also require high spatial and temporal resolution input data. Hydrological conceptual models have been successfully used in the past for various applications, namely, rainfall-runoff modelling (Franchini and Pacciani 1991), estimation of groundwater flow (Arnold et al. 1993), climate studies (Füssel 2007), etc. In general, conceptual models can either be discrete models, described by difference equations or continuous models, formulated in terms of ordinary differential equations (spatially invariant) or partial differential equations (spatially variant). The lumped form of such conceptual models assumes even distribution of input-output data over the catchment surface. Customary approach to the development of rainfall-runoff relationships consists of two distinct steps:

- Determination of the volume of runoff as a result of rainfall during a given time period: Partitioning of rainfall among evapotranspiration, infiltration and runoff.
- Distribution of volume of runoff in time: Accounting for travel time and attenuation of the runoff wave due to storage and other effects (Flood routing).

Sugawara Tank model (Sugawara 1979) is used as an example of conceptual rainfall runoff model in this study. It is a simple lumped model that represents catchment as an assemblage of interconnected storages through which water flows, from rainfall (input) to streamflow (output) at the outlet. The soil moisture storage is simulated by a series of tanks arranged one below the other as shown in Fig. 1. Two types of water in Tank model are confined water, namely soil moisture and free water that drains downwards and sideways. The rainfall is assumed to enter the uppermost tank. Each tank has one outlet in the bottom and one or two outlets on the side at some distance above the bottom. Water that leaves any tank through the bottom enters the next lower tank, except for the lowermost tank. The downflow from the lowermost tank forms a negligible part of total outflow. Water leaving any tank through a side outlet is referred to as sideflow and becomes input to the channel system. The number of tanks, the positions of outlets and coefficients of outflow are the parameters of Tank model.

The configurations shown in Fig. 1 are typical representations of rainfall-runoff processes of small basins in humid regions. More complex arrangements that consist of multiple series of many reservoirs are more suitable for large basins with strong seasonality,

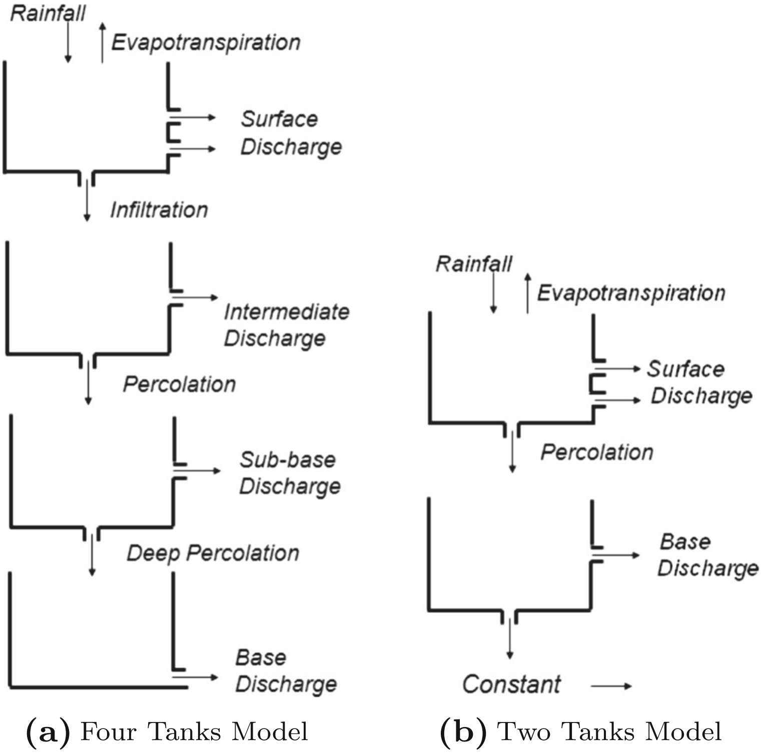


Fig. 1 Sugawara Tank model (Sugawara 1979)

deep and permeable soil column. The meteorological variables, say, precipitation (rainfall/snowfall) and potential evapotranspiration are inputs to Tank model. The basic form of equation of sideflows used in this study is,

$$Q_t = \begin{cases} A(S_t - H), & \text{if } S_t > H \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where, Q_t denotes sideflow, S_t the storage, H the height of the outlet above the bottom of the tank, A the discharge coefficient. The basic form of equation of downflows is,

$$I_t = B(S_t) \tag{2}$$

where, I_t represents downflow, B the infiltration coefficient. There are numerous evidences of the successful implementation of Tank model in hydrological studies at various places (Basri 2013) which in turn requires the selection of appropriate model configuration. The traditional practice of extending the models' application to different geographical areas is to presume a model structure based on prior experience and determine the associated parameters using manual trial and error or automated calibration algorithms. This work presents an evolutionary data driven approach to select an optimal model configuration from model space consisting of multiple model structure hypotheses and parameter sets, in the absence of prior knowledge of interactions amongst observed data and catchment characteristics.

1.2 System Identification in Hydrology using Genetic Programming

System identification is a methodology for building mathematical models of dynamic physical systems that require measurements of system's inputs and output responses (Winkler et al. 2012). Genetic Programming (GP) (Koza 1992; Babovic 1996) evolves grey box models as opposed to black box models where no definite model form is returned and white box models that are purely phenomenological, based on first principles. GP is an approach that copy evolutionary mechanisms for finding functional relationships between input-output variables of the system and generating mathematically meaningful solutions (Babovic and Keijzer 2000). The terminal set of GP consists of independent variables (causative variables of hydrological process of training dataset), and numerical constants (model parameters). The GP's function set can include basic arithmetic operators, trigonometric functions, boolean operators, logical expressions and other user defined domain specific functions. The performance of GP depends on selection of primitive set (functions, terminals) and fitness metrics that measure the predictive capability of the evolved model. The first step in GP implementation is the random creation of initial population, providing satisfactory coverage of the search space. The population for the next generation involves selecting better individuals (represented as syntax trees), focusing on worthwhile regions of the search space. Crossover and mutation are the two main genetic operators used in the transformation of best individuals into new generation individuals. Elitism is introduced so that few (best-so-far) individuals in the mating pool are directly included in the next generation. The evolutionary process continues over successive generations until the termination criterion is met, which is usually set as the maximum number of generations. As GP is a randomized search process, several independent GP runs are carried out for a given list of settings. Performance evaluation of all GP runs is carried out followed by the selection of the model that best reproduces the observed response. A comprehensive list of applications of Genetic Programming in various fields of hydrological modelling can be found in Wang et al. (2009) and Oyebode and Adeyemo (2014), which includes, rainfall-runoff modelling (Khu et al. 2001; Whigham and Crapper 2001; Liong et al. 2002; Dorado et al. 2003; Meshgi et al. 2015), streamflow forecasting (Londhe and Charhate 2010; Whigham and Crapper 2001), sediment transport modelling (Babovic 2000), daily prediction of algal blooms (Muttil and Lee 2005), weather prediction (Bautu and Bautu 2006), deep percolation model (Selle and Muttil 2011), groundwater modelling (Fallah-Mehdipour et al. 2014) etc.

In all previous studies, GP has been successfully used as a symbolic regression tool that uses the given inputs and evolves optimal models in the form of mathematical formulae that offer a good compromise between accuracy and complexity. Adding hydrological concepts into GP framework to evolve physically interpretable models for advancing system identification in the field of hydrology is gaining interest and momentum. For example, in Havlicek et al. (2013), a combined approach of hydrological concepts and GP automatically determines the input (rainfall) memory through the process of optimization, resulting in faster convergence and improved runoff forecasts. In this study, GP is equipped with generic components of Sugawara Tank model (lumped conceptual rainfall-runoff model) to evolve the optimal model configuration for the given data using single and combined statistical criteria for evaluating the fit between observed and simulated values. The resultant GP Tank model configurations vary in terms of number of reservoir units and connections, number and type of outflows, governing functions and parameters, thereby accounting for diversity of climate and geomorphology of catchments. Thus, this approach can be regarded as a fully unsupervised solution to model configuration problem. The purpose of choosing evolutionary data driven technique is its potential to evolve novel model development strategies which will be exploited in the further work. The authors believe that this work will promote widespread use of GP in the process of scientific discovery in the field of hydrology.

2 Materials and Methods

2.1 Description of Datasets

2.1.1 Kent-Ridge Catchment Dataset

A monitoring programme was established to collect dense hydrological data in the Kent-Ridge catchment, National University Singapore (Deng et al. 2013; Meshgi et al. 2015). This tropical catchment contains main landuse types of Singapore that includes impervious surfaces (i.e. roofs, roads, paved car parks), grasses on steep slopes, mixed grasses and trees and natural vegetation. The elevation varies from 14.04 m to 75.84 m above sea level and the topography is characterized by steep slopes. The pattern of rainfall varies over the year due to two monsoons: Northeast (mid November to early March) and southwest monsoon (mid June to September). Rainfall is an everyday phenomenon in this tropical catchment even during the non-monsoon period. Due to its geographical location and maritime exposure, Singapore's climate is characterized by uniform temperature and pressure, high humidity and abundant rainfall. The mean annual precipitation is 2340 mm. The average temperature is between 25 °C and 31 °C. Thunderstorms occur on 40% of all days. Relative humidity is in the range of 70%–80% and mean annual wind velocity is 15 km/h. The rain gauge installed on one of the roof tops of Kent-Ridge catchment recorded precipitation data at one minute interval with an accuracy of 0.2 mm. Water level measurements recorded at 5 monitoring locations at the same temporal resolution are converted into discharge using appropriate stage-discharge relationships. The high resolution data (one minute resolution) collected over a period of 9 months from September 2011 to May 2012 is used for this study. The data includes time series of catchment averaged rainfall intensity (P in mm) and Discharge at the catchment outlet (Q in mm) Potential evapotranspiration (E in mm) computed using Penman-Monteith equation (Monteith 1965).

2.1.2 Maimai Catchment Dataset

Maimai catchment of South island, New zealand was established as hydrological experimental site in late 1974 and is one of the most researched catchments (McGlynn et al. 2002). The climate is mainly humid with mean monthly relative humidity of 87%, microthermal with mean annual temperature of 1.1 °C and adequate rainfall in all seasons with 2450 mm as mean annual rainfall. Maimai is a forested catchment with short, steep slopes and topographical variation between 100 and 150 m. The soils are within 10% saturation through most of the year with poorly permeable sub soils. The hourly data of precipitation (P in mm) recorded with a rain gauge located within the catchment, potential evaporation (E in mm) estimated as described in Rowe et al. (1994) and discharge (Q in mm), recorded over a period of two years from January 1985 to December 1987 are used for this study.

Table 1 shows the geographical location and area of the two datasets used in this study. A split-sample scheme is used for the selection of calibration and validation intervals so that

Table 1 Location and area catchments used in this study

Catchments	Area(Km^2)	Latitude	Longitude
Kentridge	0.085	1°17’45.3”N	103°46’28.0”E
Maimai	0.038	42°05’S	171°48’E

they represent different volumes of runoff formations (high, medium or low flows). Half of each type of events is considered for training and the remaining for validation, All GP runs follow same scheme of partitioning.

2.2 Genetic Programming Framework for Automatic Model Generation

This section presents the GP algorithm developed in R (Team R Core 2014), a free software environment for automatic conceptual model generation, which is a refinement of a standard syntax tree GP method SORD. SORD has already been used in improving rainfall-runoff forecasts (Havlicek et al. 2013) and estimating runoff at ungauged catchments by regionalization (Hermanovsky et al. 2017).

Prior knowledge of the physical system is not required but useful in the definition of potential components of the function set. In Table 2, R2T and R4T are functions representing Tank models with fixed structure consisting of a series of two and four reservoir units respectively as illustrated in Fig. 1. TANK is a function with varying argument size representing the flexible Tank model. TANK function can represent Tank model configurations with one or more series of reservoirs (depends on the tree depth), variable number of reservoirs up to a maximum of 4 in each series, variable number of outflows from each reservoir capped at 3 (2 sideflows and 1 downflows) and associated parameters. TANK function has restricted/bounded arguments which can only be constants within a user specified range, representing model parameters representing heights of outlets of reservoirs (range: 0-100) and discharge coefficients (range: 0-10). The terminal set of GP consists of reservoir inputs (independent variables) and random constants. The GP individual is represented as a linearised syntax tree array. The total number of columns of the GP individual array depends on the maximum arity value of the functions used. The first column can be a function/terminal. In case of function, other columns are pointers to arguments of that particular function or are empty in case of terminal. The total number of rows depends on the tree depth. The evaluation of the GP individual is processed and returned as a symbolic expression. The model configuration is evolved by GP using objective functions (single or combined) that measure the deviation between model and system responses. The fitness functions used here are Madsen metric (Madsen 2000) which is a combination of four numerical performance statistics (3) and (4) and Nash Sutcliffe Efficiency (NSE) measure (7). The performance of GP results are evaluated visually by comparing observed and simulated hydrographs and by calculating Kling-Gupta Efficiency (KGE) (Gupta et al. 2009) (8) in addition to Madsen and NSE metrics.

$$Volume\ Error = \left|\frac{\sum_{i=1}^{n_j}(Q_{obs,i} - Q_{sim,i}(\theta))}{n_j}\right|,\ j = all \tag{3}$$

$$Overall,\ High,\ Low\ RMSE = \sqrt{\frac{\sum_{i=1}^{n_j}(Q_{obs,i} - Q_{sim,i}(\theta))^2}{n_j}},\ j = all,\ h,\ l \tag{4}$$

Table 2 Genetic Programming function set used in this study

Functions	Arity	Remarks
+, −, *, /	2	Basic Algebraic Functions
R2T	7	Two Tanks Model function (Fig. 1)
R4T	13	Four Tanks Model function (Fig. 1)
TANK	Variable arity	Flexible Tank model function

where, n_{all} represents the total length of data, n_h and n_l are the number of time steps corresponding to high and low flows respectively, $Q_{obs,i}$ and $Q_{sim,i}$ denote observed and simulated flows respectively, θ denotes the set of model parameters restricted to parameter space Θ, defined by upper and lower limits on each parameter.

$$Madsen = \sqrt{(F_1 + A_1)^2 + (F_2 + A_2)^2 + (F_3 + A_3)^2 + (F_4 + A_4)^2} \quad (5)$$

$$A_i = Max[F_{j,min},\ j = 1\ to\ 4] - F_{i,min},\ i = 1\ to\ 4 \quad (6)$$

where, F_1 represents volume error, F_2, F_3, F_4 denote overall, high, low RMSE respectively, A_1 to A_4 are transformation constants such that all $F_i + A_i$, i = 1 to 4 have the same distance to the ideal point (0). The minimum values of F_i ($F_{i,min}$) are estimated from initial GP population.

$$NSE = 1 - \frac{\sum_{i=1}^{n}(Q_{obs,i} - Q_{sim,i}(\theta))^2}{\sum_{i=1}^{n}(Q_{obs,i} - \overline{Q_{obs,i}})^2},\ NS0 = 1 - NSE \quad (7)$$

where, $\overline{Q_{obs,i}}$ denotes mean value of observed discharges,

$$KGE = \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2},\ KG0 = 1 - KGE \quad (8)$$

where, r represents Pearson correlation coefficient, α is equal to the ratio of standard deviation of simulated discharges to the standard deviation of observed discharges and β equals the ratio of mean of simulated discharges to the mean of observed discharges. The output of the proposed framework consists of symbolic expressions of evolved model configurations. The simplification of resultant expressions is carried out using Yet Another Computer Algebra System (YACAS) (Pinkus and Winitzki 2002). The constant parameters in GP evolved model configurations can be optionally fine tuned using another calibration algorithm, for example, Differential Evolution (Storn and Price 1995). Complexity is used along with performance on training and testing datasets as criteria for model selection. Wide variety of measures, say, number of variables, model length (tree length), expressional complexity (visitation length) (Keijzer and Foster 2007), functional complexity (order of nonlinearity) (Vanneschi et al. 2010), measure accounting for model semantics (Kommenda et al. 2015), have been successfully used in the earlier studies to determine complexity. A simple structural complexity metric is used in this study that denotes the number of tanks/reservoirs present in the GP evolved model configuration.

2.3 Implementation

The basic statistical characteristics of Rainfall and Discharge series of training and validation datasets of Kentridge and Maimai catchments used in this study are presented in Table 3. In the case of Kentridge catchment, one-minute rainfall (P) data are aggregated to five-minute data. Similarly, the discharge (Q) and potential evaporation (E) are averaged to five-minute data. In this study, two types of simulations are conducted using GP based model generation framework to model rainfall-runoff processes.

- Simulations using synthetic Kentridge discharge data, two preselected Tank model configurations R2T and R4T and NS0 as optimization objective.
- Simulations using real/observed datasets of Kentridge and Maimai catchments, TANK function, Madsen and NS0 as optimization objectives.

The purpose for structuring experimentation as presented below is to firstly establish whether GP is capable of finding the relationship for which there is a proof that the optimal solution exists. Only after satisfactory results using synthetic data are achieved, a more challenging second set of simulations using real data are pursued. Table 4 shows GP settings for both synthetic and real data simulations. Fifty independent GP runs are carried out for each list of settings which are evaluated for the selection of optimal model configuration for the given dataset. The best model is the one that offers a good compromise of training, testing fitness values and complexity. Feasible non dominated model configurations (Pareto optimal) are derived based on training and testing fitness values from which one best configuration is chosen based on structural complexity.

3 Results and Discussions

3.1 Simulations using Synthetic Data

The results of simulations conducted using synthetic discharge data of Kentridge catchment and settings given in Table 4 are presented in this section. The aim of this exercise is to highlight the efficiency of GP in retrieving the model used to generate the synthetic data. Q_{syn2T} represents the synthetic discharge data generated using R2T function and observed P, E of Kentridge catchment. The symbolic representation of R2T function is given in (10).

$$RI = \text{maximum}((P - E), 0) \tag{9}$$

$$Q_{syn2T} = R2T(RI, H1, A1, H2, A2, A0, B1) \tag{10}$$

where, RI represents input to the top most reservoir unit of the structure, H1 and H2 represent heights of sideflow outlets, A1, A2 and B1 represent coefficients of sideflows and A0 represent coefficient of downflow.

Figure 2 is the pictorial representation of GP evolved two tanks model to estimate Q_{syn2T} that has the best fitness value (NSE). The symbolic expressions of the target and the best GP evolved two tanks model are given in Table 5. GP two tanks model (Fig. 2) indicates that surface discharge is the only contributor to the total runoff and the base discharge is predicted as zero. This is in good agreement with the target model which is associated with very low base flow coefficient (B1=8.6e-05).

QGP_{syn2T} represents the simulated discharge of the best GP two tanks model. The GP function set used for this exercise includes R2T, R4T, +, −, *, / (Table 4). The GP evolved and target models have similar structures (R2T) but differ in terms of model parameters. Fine tuning of model parameters evolved by GP can be optionally carried out using Differential Evolution or any suitable constant optimization algorithm. The performance evaluation of the resultant GP two tanks model is presented in Table 6. GP simulated response is in good agreement with the target indicated by high values of hydrological effi-

Table 3 Characteristics of datasets used in this study

Catchment	Data	Maximum		Minimum		Mean		Standard Deviation	
		Training	Validation	Training	Validation	Training	Validation	Training	Validation
Kentridge	Rainfall (P) in mm/5min	11.15	15.36	0	0	0.33	0.34	0.91	1.08
	Discharge (Q) in mm/5min	4.46	4.95	1e-3	0	0.18	0.1	0.38	0.46
Maimai	Rainfall (P) in mm/h	19.7	17.8	0	0	0.27	0.28	1	1
	Discharge (Q) in mm/h	6.66	8.5	1e-4	1e-4	0.16	0.17	0.39	0.42

Table 4 Settings for simulations using GP based model generation framework

Parameter	Value
Independent Runs	50
Population Size	500
Number of Generations	100
Initialization	Ramped Half and half
Tree Depth - Initial/Max	Synthetic data simulations: 1/1, Real data simulations: 1/2
Fitness Function	Synthetic data simulations: NS0
	Real data simulations: Madsen, NS0
Function Set	Synthetic data simulations: R2T, R4T, +,−,*,/
	Real data simulations: TANK, +,−,*,/
Terminals	Independent Variables:
	Synthetic data simulations: P, E of Kentridge dataset
	Real data simulations: P, E of Kentridge, Maimai datasets
	Random constants, Normalised model parameters: Range: 0 to 1
Dependent Variables	Synthetic data simulations: Q_{syn2T}
	Real data simulations: Q of Kentridge and Maimai datasets
Simplification	Yacas (Pinkus and Winitzki 2002)
Selection Method	Tournament (Size=4)
Probability of Crossover and Replication	0.7 and 0.05
Probability of Mutation (Constant/Node/Separation/Tree)	0.5/0.3/0.3/0.5
Complexity Computation	Structural Complexity (Number of Reservoirs)

ciency measures (NSE and KGE). Figure 3 shows the plot of synthetic Q_{syn2T} and GP simulated discharge time series QGP_{syn2T}.

Figure 3 shows that target and simulated hydrographs are well correlated (r=0.998). The peaks are well approximated whereas considerable deviation is observed with respect to very low flows. This is because the fitness function NS0 is sensitive to high flows and masks

Fig. 2 Best GP evolved two tanks model

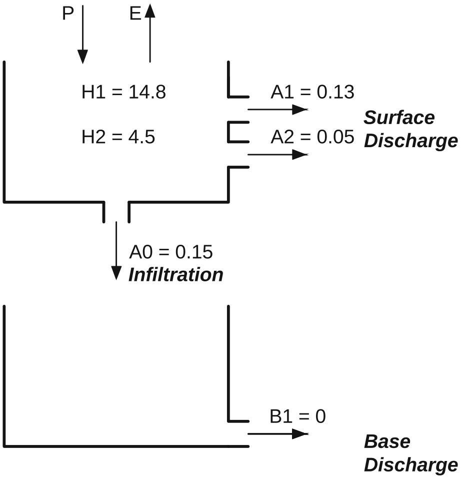


Table 5 Equation of the best GP two tanks model

Model	Equation
Target two tanks model	Q_{syn2T}= R2T(RI,1.37,0.17,11.25,0.02,0.14,8.6e-05)
GP two tanks model	$Q_{GPsyn2T}$= R2T (RI,14.8,0.13,4.5,0.05,0.15,0)

good performance for others. Hence for the simulations using real data presented in the sequel, two objective functions NS0 and Madsen that focus on different parts of hydrograph are used and the outcomes are compared in order to select the most suitable configuration.

3.2 Simulations using Real Data

This section explores the ability of the proposed framework in evolving the structure and parameters of Tank model that best suit the hydrometeorological field data collected from Kentridge and Maimai catchments. The symbolic representations of a few possible configurations that can be evolved by GP using TANK function are given below.

$$QGP = TANK(RI, H1, A1, H2, A2, A0) \tag{11}$$

$$QGP = TANK(RI, H1, A1, H2, A2, A0, H3, B1, H4, B2, B0) \tag{12}$$

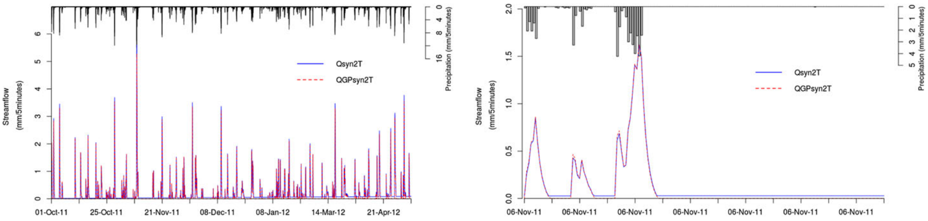
$$\begin{aligned} QGP \;=\; & TANK(TANK(RI, H1, A1, H2, A2, A0, H3, B1, H4, B2, B0), \\ & H5, C1, H6, C2, C0, H7, D1, H8, D2, D0) \end{aligned} \tag{13}$$

In Equations 11, 12 and 13, QGP represents the discharge simulated by GP evolved configurations. Equations 11 and 12 represent Tank model with one reservoir and two reservoir units in a series respectively with maximum number of outflows restricted to 3 per unit. Equation 13 represents a more complex structure that consists of a pair of two serial cascaded reservoirs each representing a zone of the catchment. The total discharge of the first set forms the input to the second set which is the closest to catchment outlet. RI represents input to the topmost reservoir unit of the first set, H1 to H8 denote heights of the sideflow outlets, A1, A2, B1, B2, C1 and C2 denote the coefficients of sideflows, A0, B0, C0 and D0 are coefficients of downflows. Fifty independent GP runs are carried out using settings in Table 4 for Kentrdige and Maimai catchments. The best out of fifty resultant models is chosen based on Fig. 4. Figure 4 shows Pareto optimal GP evolved model configurations representing the trade off between training and testing fitnesses. The labels of Pareto optimal points represent GP run indices (varying between 1 and 50) with structural complexity values (number of reservoir units) in the brackets. The best model that is less complex and offers a good compromise between performance on training and testing datasets is selected (highlighted in red). The equations of the best model configurations for Kentridge and Maimai catchments evolved using NS0 and Madsen fitness metrics are presented in Table 7.

Figure 4 shows that Tank models with two and three reservoir units have been found as optimal configurations for both Kentridge and Maimai catchments based on Madsen and

Table 6 Performance of the best GP two tanks models

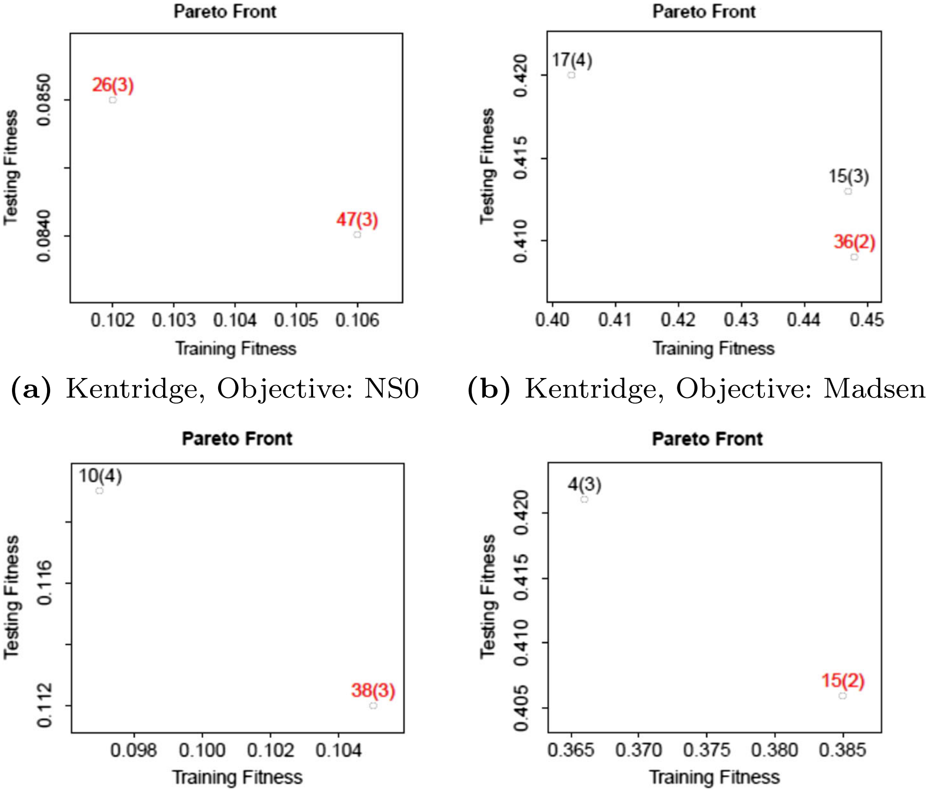
High RMSE	Low RMSE	Overall RMSE	Volume Error	NSE	KGE	r
0.055	0.022	0.05	0.041	0.986	0.772	0.998



(a) All Events (b) One selected Event

Fig. 3 Target and simulated hydrographs of synthetic Kentridge dataset

NS0 respectively. Visual selection of the best model becomes a challenging task if models on the Pareto front have similar complexity as encountered in the case of simulations using Kentridge data and NS0 as optimization objective. The performance of selected Pareto optimal models (highlighted in red) on respective testing datasets is evaluated using accuracy metrics and hydrological efficiency measures (Table 8). Table 8 shows that GP evolved configuration with model ID 47 has superior performance as compared to the other with model ID 26 for Kentridge dataset using NS0 as fitness function.



(a) Kentridge, Objective: NS0 (b) Kentridge, Objective: Madsen

(c) Maimai, Objective: NS0 (d) Maimai, Objective: Madsen

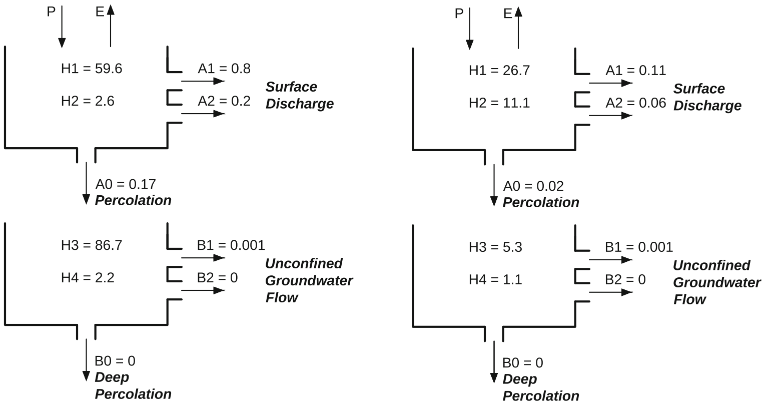
Fig. 4 Best Model selection for Kentridge and Maimai catchments

Table 7 Resultant model expressions of Real data simulations

Dataset	Objective function	Model/Run ID	Best model equation
Kentridge	NS0	26	$QGP_{KR\ NSE}$ = TANK(RI,4.7,0.03,3.9,0.05,0.57,1,0.17,0.4,0.09,0.33,100,0.11,0,0.01,0.02)
		47	$QGP_{KR\ NSE}$ = TANK(RI,0.7,0.13,3.9,0.05,0.55,0.5,0.26,0,6e-3,0.64,0,0.013,0,0,0.03
Kentridge	Madsen	36	$QGP_{KR\ Madsen}$ = TANK(RI,2.6,0.2,59.6,0.8,0.17,2.2,0,86.7,1e-3,0)
Maimai	NS0	38	$QGP_{MM\ NSE}$ = TANK(RI,10.8,0.05,22.4,0.11,0.04,0,0.07,0.1,0.001,0.224,12.6,0.001,22.4,0,0.001)
Maimai	Madsen	15	$QGP_{MM\ Madsen}$ = TANK(RI,11.1,0.06,26.7,0.11,0.02,5.3,0.001,1.1,0,0.001)

Table 8 Performance evaluation of best GP evolved models for Kentridge and Maimai catchments

Catchment	Dataset	Objective function	Model/Run ID	High RMSE	Low RMSE	Overall RMSE	Volume Error	NSE	KGE	r
Kentridge	Testing	NS0	26	0.217	0.102	0.133	0.014	0.915	0.908	0.957
Kentridge	Testing	NS0	47	0.219	0.097	0.132	0.006	0.916	0.939	0.958
Kentridge	Testing	Madsen	36	0.226	0.047	0.127	0.009	0.922	0.939	0.961
Maimai	Testing	NS0	38	0.246	0.021	0.14	0.006	0.888	0.912	0.942
Maimai	Testing	Madsen	15	0.249	0.013	0.141	0.003	0.885	0.932	0.941

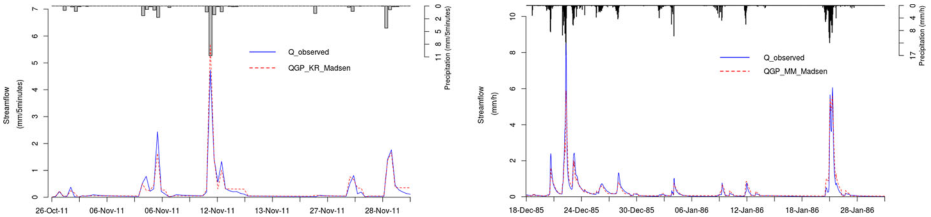


(a) Kentridge catchment **(b)** Maimai catchment

Fig. 5 Optimal Tank model configurations evolved based on Madsen metric for Kentridge and Maimai catchments

The model configurations evolved based on NS0 show slightly better performance with respect to high flows evident by lower high RMSE value in comparison to the results of simulations based on Madsen for both Kentridge and Maimai catchments. Taking different aspects of flows into account (Table 8), the configurations evolved based on Madsen can be considered optimal both in terms of accuracy and complexity for the two catchments considered in this study. The pictorial representation of thus selected configurations are presented in Fig. 5. Combined fitness functions such as Madsen that enable better performance across a range of flow characteristics can be more preferred as optimization objectives of data driven algorithms. GP results indicate that the processes contributing to the total runoff of both Kentridge and Maimai catchments are predominantly surface flows, low percolation and very little unconfined groundwater flows occurring close to ground surface. In Meshgi et al. (2015), it is established that overland flow, shallow sub-surface and baseflow contribute to total runoff of Kentridge catchment. As per (Euser et al. 2013), the best performing configuration of Maimai catchment consists of riparian, unsaturated and fast reservoirs representing surface, shallow subsurface runoff with negligible/no baseflow. Also, it is to be noted that Kentridge is a urban catchment and Maimai is a forested catchment with shallow soil column and poorly permeable subsoil with a low deep percolation rate. The result of GP is found to be in good agreement with field observations and earlier studies (Fig. 5).

Figure 6 show the observed and simulated hydrographs of Kentridge and Maimai catchments. Time steps representative of low, medium and high flows are selected from respective



(a) Kentridge catchment **(b)** Maimai catchment

Fig. 6 Observed and Simulated real data hydrographs

testing datasets for the plot. The peaks are slightly underestimated. The deviation from the observed is the greatest with respect to medium flows.

The overall trend is well captured by GP evolved models with high correlation ($r > 0.94$) and hydrologic efficiency ($NSE > 0.88$, $KGE > 0.90$) for both Kentridge and Maimai catchments.

4 Summary and Future Work

Genetic Programming (GP) framework for automatic conceptual hydrological modelling is introduced in this paper. The proposed GP framework has a customized function set which contains generic components of Sugawara Tank model in addition to the basic algebraic functions. GP evolves the most suitable Tank model configuration by deciding on model components and operational rules which include, number of series of tanks, number of reservoir units in each series, number and functions of outflows from each reservoir unit, other associated constant parameters. In this study, GP framework evolves optimal Tank model configurations to represent Rainfall-Runoff processes of two catchments with contrasting characteristics namely, Kent Ridge of Singapore and Maimai of New Zealand. Fifty independent GP runs are performed for each case and the best model that offers a good compromise between accuracy and complexity is selected. GP exhibits better performance in evolving the model structure in comparison to estimating the constant parameters. Therefore, GP can be optionally coupled with a suitable constant optimization algorithm to improve parameter optimization, which is not presented in this study. GP simulations based on NS0 and Madsen metrics suggest that combined fitness metrics (Madsen) as optimization objectives contribute to better performance of GP framework and result in configurations that perform well across multiple flow characteristics. GP models with two reservoir units evolved based on Madsen are found to be the optimal configurations for Kentridge and Maimai datasets that rightly account for dominant processes and geology of the catchments.

Overall, GP based conceptual modelling approach is promising and can be used in building hydrological models and evolving modelling strategies for catchments of varying sizes, locations and climatic conditions. The future work will focus on testing the proposed methodology on large basins, using different optimization objectives and inclusion of sophisticated model components (Fenicia et al. 2011) into GP framework, in place of simple linear reservoir elements used in the presented work.

Acknowledgements The authors would like to thank Dr. Ali Meshgi (ameshgi@gmail.com) for Kent Ridge catchment dataset and Dr. Fabrizio Fenicia (Fabrizio.Fenicia@eawag.ch) for Maimai catchment dataset. They are also grateful to reviewers for their insightful comments for improving the manuscript.

References

Arnold JG, Allen PM, Bernhardt G (1993) A comprehensive surface-groundwater flow model. J Hydrol 142(1):47–69

Babovic V (1996) Emergence, evolution intelligence: hydroinformatics. TU Delft, Delft University of Technology

Babovic V (2000) Data mining and knowledge discovery in sediment transport. Comput-Aided Civil Infrast Eng 15(5):383–389

Babovic V, Keijzer M (2000) Genetic programming as a model induction engine. J Hydroinf 2(1):35–60

Basri H (2013) Development of rainfall-runoff model using tank model: Problems and challenges in Province of Aceh, Indonesia. Aceh Int J Sci Technol 2:1

Bautu A, Bautu E (2006) Meteorological data analysis and prediction by means of genetic programming. In: Proceedings of the 5th workshop on mathematical modeling of environmental and life sciences problems constanta. Romania, pp 35–42

Charizopoulos N, Psilovikos A (2016) Hydrologic processes simulation using the conceptual model Zygos: the example of Xynias drained Lake catchment (central Greece). Environ Earth Sci 75(9):1–15

Deng Y, Cardin MA, Babovic V, Santhanakrishnan D, Schmitter P, Meshgi A (2013) Valuing flexibilities in the design of urban water management systems. Water Res 47(20):7162–7174

Dorado J, Rabuñ AL JR, Pazos A, Rivero D, Santos A, Puertas J (2003) Prediction and modeling of the rainfall-runoff transformation of a typical urban basin using ANN and GP. Appl Artif Intell 17(4):329–343

Euser T, Winsemius H, Hrachowitz M, Fenicia F, Uhlenbrook S, Savenije H (2013) A framework to assess the realism of model structures using hydrological signatures. Hydrol Earth Syst Sci 17(5):1893–1912

Fallah-Mehdipour E, Haddad OB, Marino MA (2014) Genetic programming in groundwater modeling. J Hydrol Eng 19(12):04014,031

Fenicia F, Kavetski D, Savenije HH (2011) Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development. Water Resour Res 47:11

Franchini M, Pacciani M (1991) Comparative analysis of several conceptual rainfall-runoff models. J Hydrol 122(1-4):161–219

Füssel HM (2007) Vulnerability: a generally applicable conceptual framework for climate change research. Global Environ Change 17(2):155–167

Gupta HV, Kling H, Yilmaz KK, Martinez GF (2009) Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. J Hydrol 377(1):80–91

Havlicek V, Hanel M, Máca P, Kuraz M, Pech P (2013) Incorporating basic hydrological concepts into genetic programming for rainfall-runoff forecasting. Computing 95(1):363–380

Hermanovsky M, Havlicek V, Hanel M, Pech P (2017) Regionalization of runoff models derived by genetic programming. J Hydrol 547:544–556

Keijzer M, Foster J (2007) Crossover bias in genetic programming. In: European conference on genetic programming. Springer, pp 33–44

Khu ST, Liong SY, Babovic V, Madsen H, Muttil N (2001) Genetic programming and its application in real-time runoff forecasting1. JAWRA J Amer Water Resour Assoc 37(2):439–451

Kommenda M, Beham A, Affenzeller M, Kronberger G (2015) Complexity measures for multi-objective symbolic regression. In: International conference on computer aided systems theory. Springer, pp 409–416

Koza JR (1992) Genetic programming: on the programming of computers by means of natural selection, vol 1. MIT press

Liong SY, Gautam TR, Khu ST, Babovic V, Keijzer M, Muttil N (2002) Genetic programming: a new paradigm in rainfall runoff modeling. JAWRA J Amer Water Resour Assoc 38(3):705–718

Londhe S, Charhate S (2010) Comparison of data-driven modelling techniques for river flow forecasting. Hydrol Sci J–J des Sciences Hydrologiques 55(7):1163–1174

Madsen H (2000) Automatic calibration of a conceptual rainfall–runoff model using multiple objectives. J Hydrol 235(3):276–288

McGlynn BL, McDonnel JJ, Brammer DD (2002) A review of the evolving perceptual model of hillslope flowpaths at the Maimai catchments, New Zealand. J Hydrol 257(1):1–26

Meshgi A, Schmitter P, Chui TFM, Babovic V (2015) Development of a modular streamflow model to quantify runoff contributions from different land uses in tropical urban environments using genetic programming. J Hydrol 525:711–723

Monteith J (1965) The state and movement of water in living organisms. In: Proc. evaporation and environment, XIXth Symp, pp 205–234

Muttil N, Lee JH (2005) Genetic programming for analysis and real-time prediction of coastal algal blooms. Ecol Modell 189(3):363–376

Oyebode OK, Adeyemo JA (2014) Genetic programming: principles, applications and opportunities for hydrological modelling. World Acad Sci Eng Technol Int J Environ Chem Ecol Geol Geophys Eng 8(6):348–354

Pinkus AZ, Winitzki S (2002) Yacas: a do-it-yourself symbolic algebra environment. In: Artificial intelligence, automated reasoning, and symbolic computation. Springer, pp 332–336

Refsgaard JC, Abbott M (1996) Distributed hydrological modelling. Kluwer Academic

Rowe L, Pearce A, O'Loughlin C (1994) Hydrology and related changes after harvesting native forest catchments and establishing Pinus radiata plantations. Part 1. Introduction to study. Hydrol Process 8(3):263–279

Selle B, Muttil N (2011) Testing the structure of a hydrological model using Genetic Programming. J Hydrol 397(1):1–9

Storn R, Price K (1995) Differential evolution-a simple and efficient adaptive scheme for global optimization over continuous spaces, vol 3. ICSI Berkeley

Sugawara M (1979) Automatic calibration of the tank model/L'étalonnage automatique d'un modèle à cisterne. Hydrol Sci J 24(3):375–388

Team R Core (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013

Vanneschi L, Castelli M, Silva S (2010) Measuring bloat, overfitting and functional complexity in genetic programming. In: Proceedings of the 12th annual conference on genetic and evolutionary computation. ACM, pp 877–884

Wang W, Xu D, Qiu L, Ma J (2009) Genetic programming for modelling long-term hydrological time series. In: 2009 Fifth international conference on natural computation, vol 4. IEEE, pp 265–269

Whigham P, Crapper P (2001) Modelling rainfall-runoff using genetic programming. Math Comput Model 33(6):707–721

Winkler S, Affenzeller M, Wagner S, Kronberger G, Kommenda M (2012) Using genetic programming in nonlinear model identification. In: Identification for automotive systems. Springer, pp 89–109