

# Using gene expression programming to develop a combined runoff estimate model from conventional rainfall-runoff model outputs

**Fernando, D.A. K.**<sup>1</sup>, **A.Y. Shamseldin**<sup>2</sup> and **R.J. Abrahart**<sup>3</sup>

<sup>1</sup> *Department of Civil Engineering, Unitec New Zealand, Auckland, New Zealand*  
Email: [afernando@unitec.ac.nz](mailto:afernando@unitec.ac.nz)

<sup>2</sup> *Department of Civil and Environmental Engineering, University of Auckland, New Zealand*

<sup>3</sup> *School of Geography, University of Nottingham, Nottingham, United Kingdom*

**Abstract:** In previous studies, artificial neural networks have been used to develop a model that combines simulated river flows from several individual rainfall-runoff models e.g. Shamseldin et al., 1997; Abrahart & See, 2002; Shamseldin, O'Connor, & Nasr, 2007. The combined runoff estimate model was found to perform better than the individual models in most of the cases. However, no attempts have been made to explain the inner workings of the combined models or the drivers for their success.

The research presented in this study investigates the use of gene expression programming (GEP) to develop a combination rainfall-runoff model through the process of symbolic regression. One of the additional advantages of this approach over the neural combination method is the model's ability to represent itself in the form of mathematical expressions.

The GEP model is developed using the daily simulated river flows of four other rainfall runoff models for the Chu catchment which is located in Vietnam. The four models are the linear perturbation model (LPM), the linearly varying gain factor model (LVGFM), the probability-distributed interacting storage capacity (PDISC) model, and the soil moisture accounting and routing (SMAR) models. In this paper, GeneXproTools 4.0, a powerful soft computing software package, is used to develop the combined model. The program provides transparent modeling solutions in the sense that it provides the users with the mathematical equation describing the combined model. The results reveal that combination using symbolic regression is successful and that a superior combined model can be developed using outputs from other individual models.

The structure of the combined model is also investigated in this study. The results show that the combined model is dominated by input information from the PDISC model forming the baseline estimate, to which different permutations and combinations of the remaining inputs from the other models are added.

This research, limited to one river catchment, paves the way for further investigations into GEP model development for different types of catchment. Over-fitting of the training set data during the model development observed in this study highlights the need to investigate appropriate stopping criteria.

**Keywords:** *Gene expression programming, Rainfall-runoff models, Combination models, Over-fitting*

## 1. INTRODUCTION

In the context of rainfall-runoff modelling, the combination modeling approach advocates the synchronous use of simulated discharges obtained from a number of rainfall-runoff models to produce an overall combined/integrated discharge output which can be used as an alternative to that produced by a single rainfall-runoff model. At present only a limited number of studies have dealt with the multi-model combination of hydrological models (Coulibaly et al., 2005; See & Openshaw, 2002; Shamseldin & O'Connor, 1999; Shamseldin et al. 1997). The emerging conclusion from these pioneering studies is that the combination modeling approach has tremendous potential for improving the accuracy and reliability of hydrological modelling forecasts and predictions. However, in these studies no attempts have been made to explore the nature of the combination function and their inner workings. Further, no explanation has been provided to account for the drivers behind the improvements in the modelling results essential to advance the use of combination modeling approaches in the field of hydrology.

This paper focuses on further advancing the understating about the inner workings of the multi-model combination function which can hold the key for further improvements in modelling results as well as providing guidance about the effective development of the combination models.

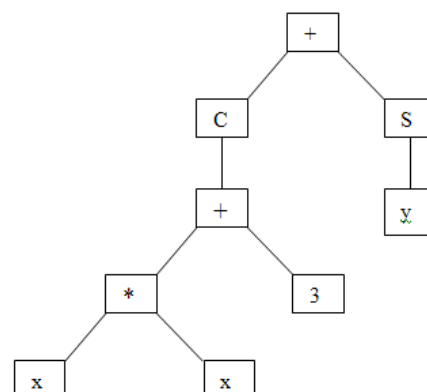
In this paper, GeneXproTools 4.0, a powerful soft computing software package is utilised to perform symbolic regression operations based on Gene Expression Programming (GEP) (Ferreira, 2001, 2006) and develop a GEP combination model using the daily estimated outputs of four rainfall-runoff models developed for the Chu catchment in Vietnam. The four rainfall runoff models, the linear perturbation model (LPM), the linearly varying gain factor model (LVGFM), the probability-distributed interacting storage capacity (PDISC) model, and the soil moisture accounting and routing (SMAR) model. Further information about these models can be found in Shamseldin et al. (1997, 2007). The GEP software package provides transparent modeling solutions in the sense that it provides its users with the mathematical equation describing the combined model.

## 2. OVERVIEW OF SYMBOLIC REGRESSION AND GEP

In broad terms, the symbolic regression is very similar to traditional parametric regression; it attempts to derive a functional relationship/model to describe the relation between dependent and independent variables. In traditional parametric regression, the form of the function relating dependent and independent variables is specified *a priori*, and thereafter, established statistical procedures are used to estimate the corresponding parameter values. In contrast, the symbolic regression is a nonparametric regression procedure, since the function relating dependent and independent variables is not specified *a priori* but is itself a product of the optimisation process. It is used to produce a constrained solution that is constructed from a number of mathematical or logical expressions - expressions selected from a large pre-selected set. The problem of symbolic regression can be viewed as an optimisation problem where the aim is to find the best combination of variables, symbols, and coefficients that will yield an optimum model, in terms of performance, according to a pre-chosen objective function. GEP, one of the optimisation techniques that can be used to solve the symbolic regression problem, is based on the principles of biological evolution similar to other evolutionary optimisation algorithms. In GEP a population of individual combined model solutions is created initially in which each individual solution is described by genes (submodels) which are linked together using a predefined mathematical operation (e.g. addition). In order to create the next generation of model solutions, individual solutions from the current generation are selected according to fitness which is based on the pre-chosen objective function. These selected individual solutions are allowed to evolve using evolutionary dynamics to create the individual solutions of the next generation. This process of creating new generations is repeated until a certain stopping criterion is met.

GEP analysis in this study uses the powerful GenXProTools<sup>®</sup> software. This tool can, for instance, given the relevant data, find the mapping function between the independent variables  $x$  and  $y$  and dependent variable  $z$  and express it in the form of a tree diagram as shown in Figure 1 where the relationship is  $z = \text{Cos}(x^2+3)+\text{Sin}(y)$ . This figure shows a single tree or a gene whereas several trees can represent more complex mapping functions.

The GenXProTools<sup>®</sup> tool was used to identify the relationship between the input variables - daily river flow



**Figure 1.** Example of a tree diagram from GEP where  $z = \text{Cos}(x^2+3)+\text{Sin}(y)$

estimates from four rainfall-runoff models, and the output variable – the actual daily river flow value.

### 3. DATA

The GEP model developed in this study used daily flow for the Chu River in Vietnam which covers a catchment area of approximately 2370 km<sup>2</sup>. This river catchment has very low flow most times of the year with significant peaky flow rates as a response to seasonal rainfall. The daily average rainfall, evaporation and discharge for this river catchment are 3.78 mm/day (average calculated for training set only), 2.54 mm/day and 1.64 mm/day respectively for the 10 years long study period. The flow rate used here is intentionally made independent of the catchment area and expressed in terms of mm/day with a view to compare different sized catchments in the future. The daily river flow records used in the study begin on the January 1<sup>st</sup> 1965; the first 8 years of data comprising a training data set of 2847 input/output vectors and the following 2 years of data comprising the testing set of 730 vectors were used in the analysis.

#### 3.1. GEP settings

In order to develop the combined model in GenXProTools<sup>®</sup>, the following parameter settings were used:

- General Settings:
  - Independent variables:
    - Input d(0) : LPM estimate (mm/day)
    - Input d(1) :LVGFM estimate (mm/day)
    - Input d(2) : PDISC estimate (mm/day)
    - Input d(3) : SMAR estimate (mm/day)
  - Dependent variable: Observed flow (mm/day)
  - Output : GEP model estimate (mm/day)
  - Number of training samples: 2847
  - Number of testing samples: 730
  - Number of chromosomes: 30
  - Head size: 8
  - Number of genes : 3. (three trees will be needed to form the final mapping function)
  - Linking function : Addition (functions on the trees will be added to form the final mapping function)
  - Constants: Two constants per gene with bounds of  $\pm 10$ .
- Fitness function: Based on Mean square error (MSE);  $Fitness = 1000.\{1/(1+MSE)\}$
- Dynamic evolution operators: Default values of mutation, inversion, transportation, recombination and transposition
- Symbolic functions: Thirteen default functions (Table 1)
- Stopping criterion: 10,000 generations

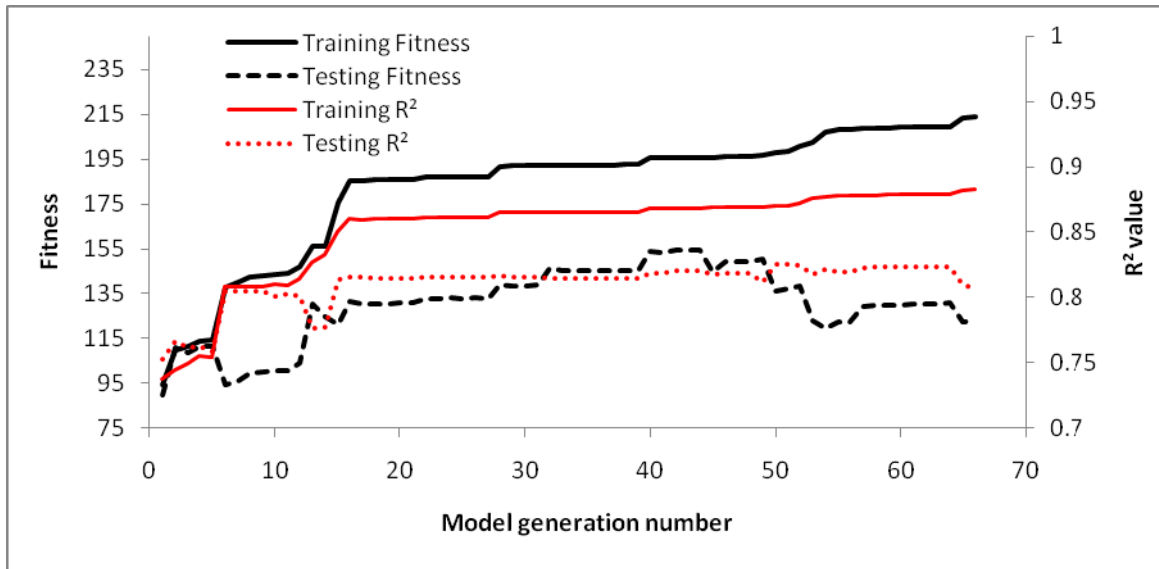
**Table 1.** Function Set

Function	Symbol
Addition	+
Subtraction	-
Multiplication	*
Division	/
Square root	Sqrt
Exponential	Exp
Natural logarithm	Ln
x to the power of 2	x2
x to the power of 3	x3
Cube root	3Rt
Sine	Sin
Cosine	Cos
Arctangent	Atan

### 4. RESULTS

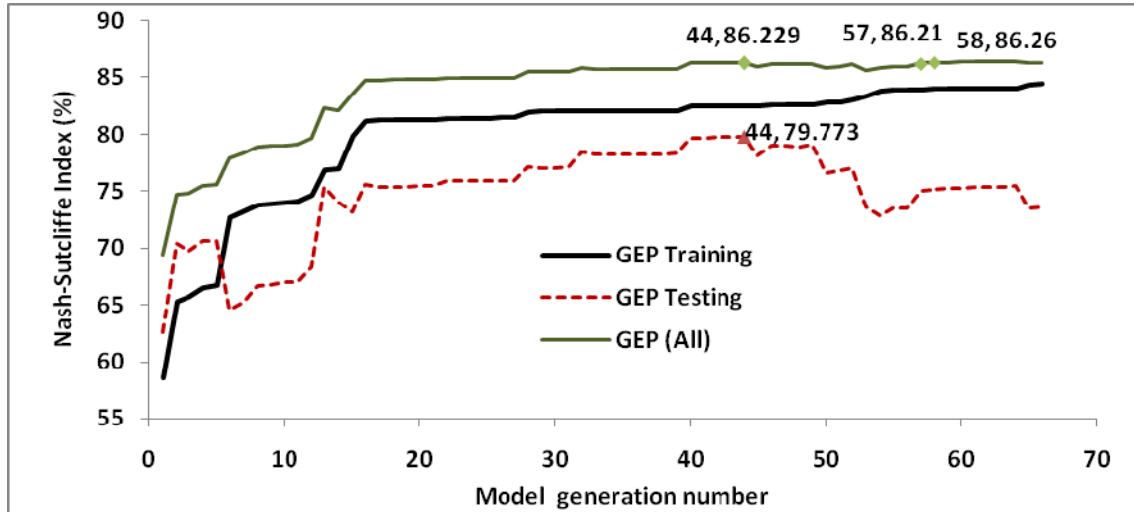
As shown in Figure 2, 66 GEP preferred models with ever-increasing fitness (for training set) were developed by the software during the evolution process that completed 10,000 generations. The figure reveals that while the fitness of the training set in the intermediate models increases monotonically, that for the testing set increases until the 44<sup>th</sup> intermediate model and thereafter begins to decay. The behavior of the standard R<sup>2</sup> values (coefficient of determination) for training and testing sets also show a similar pattern where the R<sup>2</sup> value monotonically increases for the training set, and drops somewhat for the testing set at 52<sup>nd</sup> and 64<sup>th</sup> models. A further scrutiny of the 44<sup>th</sup> model shows that although the performance of this model for the testing set is superior compared to the 66<sup>th</sup> model, it does not show superior performance for either the training set or in comparison to the original sub-models. This observation indicates that the GEP models beyond the 44<sup>th</sup> may be attempting to over-fit to the training data which is a concern. However, in this study,

the focus is on obtaining a relationship between the input and output variables and thus, the prediction accuracy of the 66<sup>th</sup> model that performed well for the overall data set was considered sufficient.



**Figure 2.** Variation of fitness of population and R<sup>2</sup> value during evolution

To capture evidence of overfitting, the Nash-Sutcliffe R<sup>2</sup> Index (NSI) for the training/testing, and the complete data sets respectively for each model generated during the GEP were compared. Figure 3 below is a plot of the NSI for the generated models which clearly shows that significant overfitting of the model to the training set occurs beyond the 44<sup>th</sup> model; all subsequent models give NSI values below 79.7736% for the testing set. The performance of the subsequent models for the overall data set too suffers temporarily (from 44<sup>th</sup> and 57<sup>th</sup> model) but as a consequence of steady improvement in the performance on the testing set, the NSI rises from the 58<sup>th</sup> model onwards.



**Figure 3.** Variation of Nash-Sutcliffe Index (NSI) for the models observed over the GEP process

The components of the final model (the 66<sup>th</sup>) and its parameters are given in Table 2. This reveals that the model is made up of three components with appropriately selected mathematical expressions. These three components are additive. All the inputs – d(0), d(1), d(2), d(3) - have been used in forming at least one component of the model. Of the functions available in GenXProTools<sup>®</sup> (Table 1), only a few, namely, addition, subtraction, multiplication, to the power of 3, and Sin have been required to fully express the GEP model. The input d(2), which is the PDISC (probability-distributed interacting storage capacity) model estimate, forms the most important part of the model solely representing one of the three components. The other two models are combinations of the inputs with interspersed constants. It is remarkable that the 44<sup>th</sup> model also contains PDISC as a single gene suggesting that the subsequent over-fitting that occurs does not affect the selection of PDISC as the single independent model.

These expression trees, when represented by mathematical expressions, can be listed by the model component equations as follows:

$$\text{varTemp} = d(2) \tag{1}$$

$$\text{varTemp} = \text{varTemp (From Eq(1))} + (G2C0 - (((G2C0^{**}3) * \sin(d(3))) * ((d(3) - d(1)) * d(0)))) \tag{2}$$


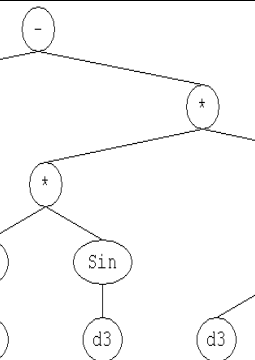
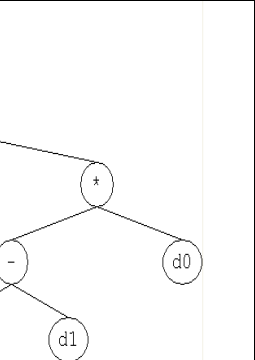
$$\text{varTemp} = \text{varTemp (From Eq2)} + (G3C0 - (((G3C0^{**}3) * (G3C1 - d(1))) * ((G3C0 - d(1)) + d(2)))) \tag{3}$$

where the constants (incidentally identical G2C0 and G3C0) are as follows:

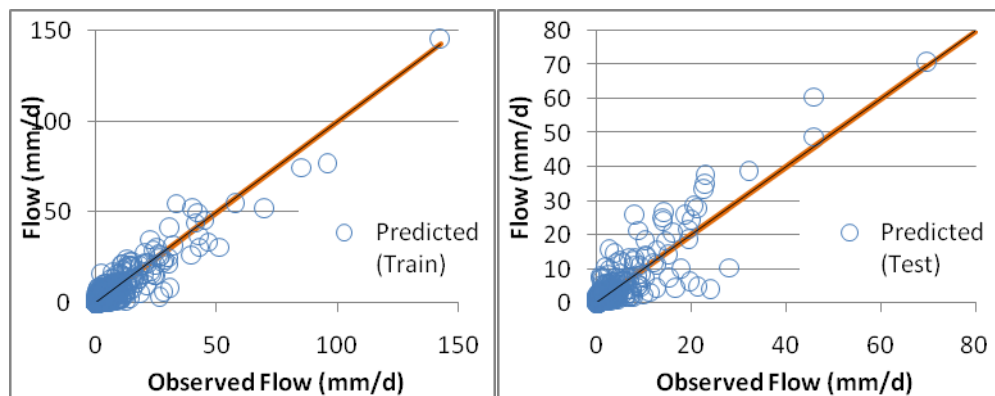
$$G2C0 = -0.230285,$$

$$G3C0 = -0.230285, G3C1 = -9.996369.$$

**Table 2.** The GEP model details

Gene No.	Genetic Expression Tree of model component	Independent variables	Constants
G1		d(2) = PDISC	
G2		d(0)=LPM d(1) = LVGFM d(3) = SMAR	C0=-0.230285
G3		d(1) = LVGFM d(2) = PDISC	C0 = -0.230285 C1 = -9.996369

The river flows estimated by the GEP model are plotted for the training and testing data sets in Figure 4. As can be seen in the figure, GEP model estimates closely correlate to the observed values. It appears that the GEP model tends to underestimate the medium-high flow values for the training set and overestimate them for the testing set.



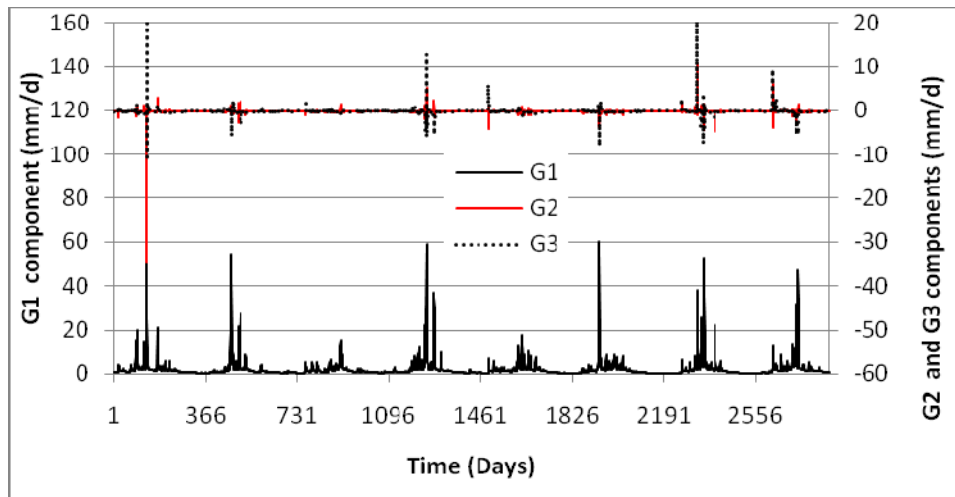
**Figure 4.** Observed vs. GEP model-predicted flow in Chu River for training and testing data sets

The GEP model performance is compared to the individual models that provided the input to the GEP model in Table 3. The NSI shows that the GEP model for the training set (84.43%) outperforms all the individual models and for the testing set the 44<sup>th</sup> GEPM model is ranked first, followed by LVGFM and then the 66<sup>th</sup> GEMP. It should be noted that, overall (for training and testing data) the 66<sup>th</sup> GEMP outperforms all others with an NSI of 86.25%.

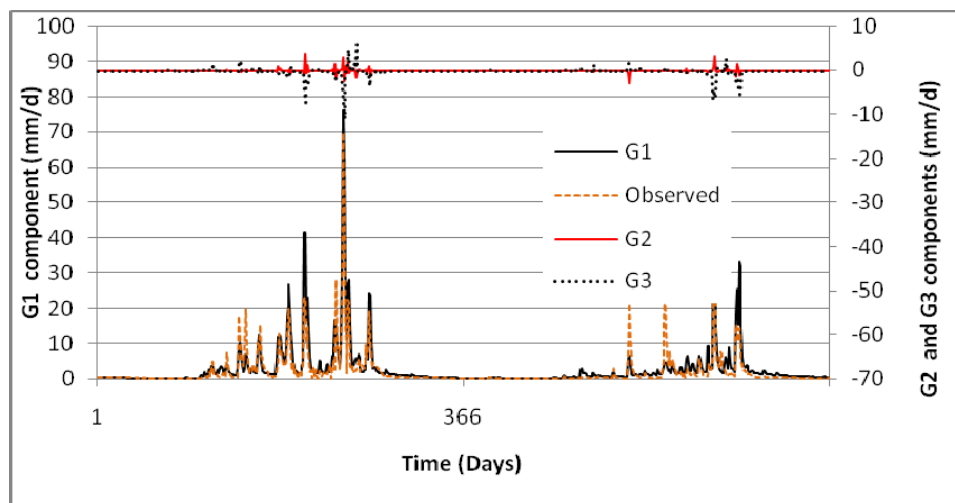
**Table 3.** Comparison of the individual models with GEP Model. (NSI: Nash Sutcliffe Index)

Model	GEPM (66 <sup>th</sup> )	GEPM (44 <sup>th</sup> )	LPM	LVGFM	PDISC	SMAR
NSI for training set (%)	84.43	82.44	63.1	83.18	79.91	75.9
Rank of Model	1	3	6	2	4	5
NSI for testing set (%)	73.60	79.77	70.28	75.34	64.44	71.82
Rank of Model	3	1	5	2	6	4

The equations identified above, when used to re-generate the numerical flow values contributed by each model component, help to identify the contribution from each sub-model and its relevance. The contributions from each gene given by Equations (1) to (3) are plotted below in Figures 5 and 6 where the contribution from Gene 1 (G1) is plotted on the primary y-axis while the contributions from Gene 2 and Gene 3 (G2 and G3 respectively) are plotted on the secondary y-axis.



**Figure 5.** Gene contributions to GEP model output for training data set



**Figure 6.** Gene contributions to GEP model output for testing data set

On Figure 6, the observed flow is also plotted on the primary axis to illustrate the significance of the G1 component. These plots reveal that while tG1 acts as the predominant signal, G2 and G3 provide “corrective” signals. It can also be observed that with the exception of a limited number of days, the corrective signals are banded within a narrow range of approximately  $\pm 20$  mm/d for the training set and  $\pm 10$  mm/d for the testing set while the bulk of the flow estimate comes from G1.

## 5. DISCUSSION AND CONCLUSIONS

The study uncovers some interesting facts from an application of the GEP technique. Some useful observations and conclusions from the study can be summarised as in the following paragraphs.

GEP can be successfully used to combine model outputs from other basic rainfall-runoff models to develop one with greater accuracy. However, as seen in the results above in Table 3, although the GEP model ranks highest for the training set (and the complete data set), its performance for the testing data set alone is not that impressive, ranking only second. It is evident that the GEP generated combination models begin to over-fit to the training data set. Future work must include a procedure to terminate the GEP when the fitness of the model for a validation set begins to deteriorate; tests must also go beyond 10,000 iterations to determine if over-fitting to training data continues or better model parsimony is eventually achieved. The presence of different potential stopping points is indicated by the fitness and  $R^2$  values that signal the need for such work.

The combination allows an insight into the components that make up the model in terms of mathematical expressions thereby making the GEP model unlike its “black-box” counterparts that have been used in the past to develop combination models. The mathematical expressions generated by the programming process can be subsequently applied to other data sets not used in the model development as well as to further investigate the contributions from each of the sub-models.

The GenXProTools<sup>®</sup> is an easy to run efficient tool based on the GEP technique which inherits the combined virtues of Genetic Programming and Genetic Algorithms. GEP, because it operates on genotype expressions as opposed to expression trees which are somewhat cumbersome to manage, does not suffer from some of the difficulties and the associated inefficiencies of GP. The time to simulate 100,000 evolutions is approximately 2.75 hours on a Desk-Top PC (Pentium (R), CPU 3.4GHz, 3.24 GB RAM).

The PDISC model emerges as a principal base-line forecast onto which permutations/combinations of other submodels are compounded. Discovering the reason for the selection of this low ranked model will be useful.

## ACKNOWLEDGMENTS

Financial support from Unitec New Zealand present this paper at MODSIM09 conference in Cairns, Australia is thankfully acknowledged.

## REFERENCES

- Abrahart, R. J., & See, L. M. (2002). Multi-model data fusion for river flow forecasting: an evaluation of six alternative methods based on two contrasting catchments. *Hydrology and Earth System Sciences*, 6(4), 655-670.
- Coulibaly, P., Hache, M., Fortin, V., & Bobee, B. (2005). Improving Daily reservoir inflow forecasts with model combination. *ASCE Journal of Hydrologic Engineering*, 10(2), 91-99.
- Ferreira, C. (2001). Gene Expression Programming: A New Adaptive Algorithm for Solving Problems. *Complex Systems*, 13(2), 87-129.
- Ferreira, C. (2006). *Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence*. Germany: 2nd Edition, Springer-Verlag.
- See, L., & Openshaw, S. (2000). A hybrid multi-model approach to river level forecasting. *Hydrological Sciences Journal* 45(4), 523-536.
- Shamseldin, A. Y., O'Connor, K. M., & Liang, G. C. (1997). Methods for combining the outputs of different rainfall-runoff models. *Journal of Hydrology*, 197(1-4), 203-229.
- Shamseldin, A. Y., & O'Connor, K. M. (1999). A Real-Time Combination Method for the Outputs of Different Rainfall-Runoff Models. *Hydrological Sciences Journal*, 44(6), 895-912.
- Shamseldin, A. Y., O'Connor, K. M., & Nasr, A. E. (2007). A comparative study of three neural network forecast combination methods for simulated river flows of different rainfall-runoff models. *Hydrological Sciences Journal*, 52(5), 896-916.
- Steeb, W.H., Hardy, Y., & Stoop, R. (2005). *The Nonlinear Workbook*. Singapore: World Scientific Publishing Company Inc.