UNIVERSITY OF GAZİANTEP
GRADUATE SCHOOL OF
NATURAL & APPLIED SCIENCES

MODELLING OF DISCHARGE-SEDIMENT YIELD RELATIONSHIP IN MIDDLE EUPHRATES BASIN USING GENETIC PROGRAMMING

M. Sc. THESIS
IN
CIVIL ENGINEERING

BY
N. ERSİN TALU
FEBRUARY 2010

Modelling of Discharge- Sediment Yield Relationship in Middle Euphrates Basin Using Genetic Programming

M.Sc.
in
Civil Engineering
University of Gaziantep

Supervisor
Assist.Prof.Dr. Aytaç GÜVEN

by

Necip Ersin TALU

February 2010

T.C.
UNIVERSITY OF GAZİANTEP
GRADUATE SCHOOL OF
NATURAL & APPLIED SCIENCES
CIVIL ENGINEERING DEPARTMENT

Name of the thesis: Modelling of Discharge-Sediment Yield Relationship in Middle Euphrates Basin Using Genetic Programming
Name of the student: N. Ersin TALU
Exam date: 26.02.2010

Approval of the Graduate School of Natural and Applied Sciences

Prof.Dr. Ramazan KOÇ
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

Assoc.Prof.Dr. Mustafa GÜNAL
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

Assist.Prof.Dr. Aytaç GÜVEN
Supervisor

Examining Committee Members	signature
Assoc.Prof.Dr. Mustafa GÜNAL	______________
Assoc. Prof.Dr. Abdulkadir ÇEVİK	______________
Asist.Prof.Dr. Aytaç GÜVEN	______________
Asist.Prof.Dr. Mustafa ŞAHMARAN	______________
Asist.Prof.Dr. Oğuzhan YILMAZ	______________

ABSTRACT

MODELING OF DISCHARGE-SEDIMENT RELATIONSHIP IN MIDDLE EUPHRATES BASIN USING

TALU, N. Ersin
M.Sc. in Civil Eng.
Supervasor: Assist. Prof. Dr. Aytaç GÜVEN
January 2010, 69 pages

This study investigates the use of Gene-expression programming (GEP) in modeling of suspended sediment yield (SSY) based on monthly averaged river discharge (Q) measurements in Middle Euphrates Basin. Conventional sediment rating curve (SRC) and multiple linear regression (MLR) techniques are often applied to determine the average relationship between SSY and Q. However, these methods are observed to generally underestimate or overestimate the amount of sediment. In the last two decades, new methods, under the name of "soft computing", have been proposed as alternative to the conventional methods. Artificial neural networks (ANNs), genetic programming (GP), support vector machines (SVMs) are some of the most widely validated branches of soft computing techniques. Except GP, the other methods generally work as "black-box" models, which are implicit that can't be simply used by other investigators. Therefore it is still necessary to develop an explicit model for the discharge–sediment relationship. The aim of this study is to develop explicit models by using GEP. GEP based explicit models, which is predicting the monthly suspended sediment yield, were compared to the rating curves and MLP techniques. The daily discharge and suspended sediment yield data from five stations on Euphrates River in Middle Euphrates Basin were used in validation of the proposed GEP, SRC and MLP models. The results indicate that the proposed GEP formulations perform superior to SRC and MLP models and these formulations are quite practical for use of water engineering society.

Keywords: Suspended sediment, discharge, gene-expression programming, rating curve, multiple linear regression

ÖZET

ORTA FIRAT HAVZASINDA DEBİ-SEDİMENT YÜKÜ İLİŞKİSİNİN GENETİK PROGRAMLAMA KULLANARAK MODELLENMESİ

TALU, N. Ersin
Yüksek Lisans Tezi, İnş Müh. Bölümü
Tez Yöneticisi: Yrd.Doç.Dr. Aytaç GÜVEN
Şubat 2010, 69 sayfa

Bu çalışma, Orta Fırat Havzası'nda aylık ortalama askıda sediment-debi ilişkisinin modellemesinde, Gen-expresyon programlama (GEP) tekniğinin kullanımını araştırmaktadır. Sediment anhtar eğrileri (SRC) ve çoklu lineer regresyon (MLR) gibi konvensiyonel metotlar, genelde debi ve askıda sediment yükü arasındaki ortalama ilişkiyi elde etmede kullanılmaktadır. Ancak, bu metotların genelde askıda sediment miktarını ya çok daha fazla ya da çok daha az tahmin ettikleri gözlemlenmiştir. Son yirmi yıl boyunca, konvensiyonel metotlara alternatif olarak, "esnek hesaplama" adı altında yeni metotlar geliştirilmiştir. Yapay Sinir Ağları (YSA), genetic programlama (GP), destek vector makineleri (DVM) en çok kullanılan esnek hesaplama tekniklerinden bazılarıdır. GP hariç diğer metotlar, içeriğinde ne tür hesaplamaların diğer araştırmacıların bilmediği "kara-kutu" mantığıyla çalışır. Bu yüzden halen, sediment-debi ilişkisi için açık formülasyonlar geliştirmek gerekmektedir. Bu çalışmada, GEP kullanarak açık formülasyonlar geliştirmek hedeflenmiştir. Aylık ortalama askıda sediment yükünü tahmin eden GEP tabanlı açık formülasyonlar, sediment anahtar eğrileri (SRC) ve MLP teknikleri ile karşılaştırılmıştır. Orta Fırat Havzasında Fırat Nehri üzerinde beş istasyondan elde edilen günlük debi (Q) ve askıda sediment yükü (SYY) veriler kullanılarak geliştirilen GEP, SRC ve MLP modelleri karşılaştırılmaktadır. Sonuçlar GEP formülasyonlarının SRC ve MLP modellerinden daha üstün performans sergilediğini ve bu formülasyonların su mühendisliği alanında kullanım için çok pratik olduğunu göstermiştir.

Anahtar kelimeler: Askıda sediment, debi, gen-expresyon programlama, anahtar eğrisi, çoklu lineeer regresyon

ACKNOWLEDGEMENTS

Firstly, I would like to express my deep gratitude to my supervisor, Assist. Prof. Dr. Aytaç Güven, for his invaluable guidance, continous advice and providing continous encouragement. He provided me stimulation and constructive criticism throughout the development of this research.

Lastly, but by no means least, I would like to thank to my parents, my wife Serpil Talu and my son Tuna Talu for their continuous encouragement and stimulation during the period of this study.

CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

F_c	the log-transformation bias correction factor
f	functional symbol
SSY	suspended sediment yield
SSY_{t-n}	time lagged SSY
$\overline{SSY}(t)$	the mean component the deterministic component of SSY
$\varepsilon(t)$	the deviation form the mean or the stochastic component of SSY
Q	water discharge
Q_{t-n}	time lagged Q
R^2	coefficient of determination
RMSE	root mean square error
$\overline{X}$	mean
S_x	skewness
C_v	standard deviation
C_{sx}	coefficient of variation
X_{max}	maximum
X_{min}	minimum

CHAPTER 1

INTRODUCTION

1.1 General

Sediment is transported based on the strength of the flow that carries it and its own size, volume, density, and shape. Stronger flows will increase the lift and drag on the particle, causing it to rise, while larger or denser particles will be more likely to fall through the flow. Rivers and streams carry sediment in their flows. This sediment can be in a variety of locations within the flow, depending on the balance between the upwards velocity on the particle (drag and lift forces), and the settling velocity of the particle (Liu, 2001).

Depending on the size and mode of transportation, total sediment load in rivers and streams are classified as bed load, suspended load (yield) and wash load (see Figure 1.1). The details on classification of these three modes are given in detail in any book related to the sediment transport such as Graf (1971), Gyr and Hoyer (2006) and Annandale (2006).

Suspended sediment yield can be defined as the fine-grained sediment that remains in the water during transportation. The suspended load is generally made up of lighter-weight, finer-grained particles such as silt and clay. Most of the sediment in a stream is carried as suspended load. It does not contribute greatly to stream erosion, since it is not in frictional contact with the stream bed.

Understanding the sediment transport phenomenon by streams is an important aspect related to issues of interactions between sediment and water quality, reservoir siltation, water pollution, channel navigability, soil erosion and soil loss, fish and invertebrate habitat, malfunctioning of hydropower plants, instream mining, river restoration and river aesthetics (Walling, 1977; Williams, 1989; Francke et al., 2008).

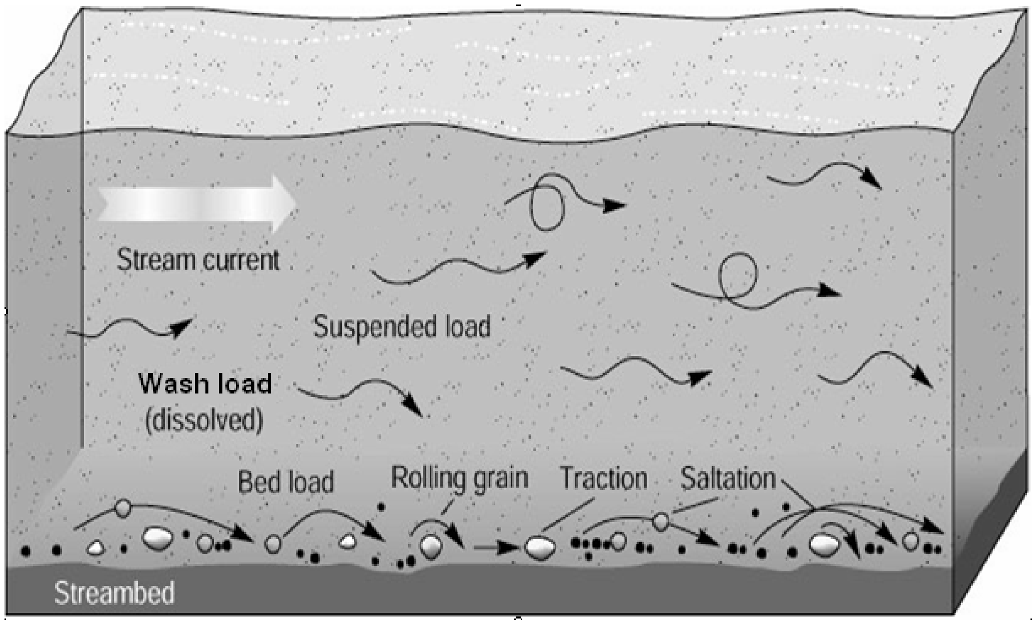


Figure 1.1 Modes of sediment transport in rivers and streams (USGS web site)

Fine sediment has long been identified as an important vector for the transport of nutrients and contaminants such as heavy metals and micro-organics. Suspended sediment is important in its own right, since its presence or absence exerts an important control on geomorphological and biological processes in rivers and estuaries (Cigizoglu, 2002a). Suspended sediment yield (SSY) at a given point for a certain time can be represented as.

$$SSY(t) = \overline{SSY}(t) + \varepsilon(t) \quad (1.1)$$

where, $\overline{SSY}(t)$ is the mean component of the deterministic component of SSY, and $\varepsilon(t)$ is the deviation form of the mean or the stochastic component of SSY. It is clear that estimation of SSY contains both deterministic and stochastic nature (Cigizoglu, 2002a). The deterministic models can be distinguished as empirical and conceptual. These models usually require long data records, so that average annual sediment yield can be determined. The conceptual models combine the mechanics of sediment transport with empirical relationships. Both the empirical and conceptual models approximate the physical processes controlling sediment yield (Singh et al., 1988; Cigizoglu, 2002a).

Direct measurement of the sediment load in a stream, which is the most reliable method, is very expensive and thus is not done for as many streams as the measurement of water discharge. On the other hand, most of the sediment transport equations require detailed information on the flow and sediment characteristics and generally do not agree with each other making it difficult to choose the best equation for a given stream. Because of these problems researchers are always looking for simpler and easy to use relationships between sediment load and water discharge or drainage area of the stream (Ozturk, 1996; Ozturk et al. 2008). In other words, the transport mechanism is complex and the available transport models are based on simplifying assumptions that often lead to large prediction errors. It is widely accepted that an accurate expression of the transport processes through a deterministic mathematical framework is difficult or even impossible (Bagnold, 1980). An alternative approach may be to use soft computing technique, which is especially attractive for modeling processes about which adequate knowledge of the physics is limited (Bhattacharya et al., 2005).

When incomplete data is existent in the periodic and stochastic components of a hydrologic time series, it becomes very complex and further characteristics are needed for its understanding and description. The studies related to the analysis of incomplete river flow data are limited compared with continuous flow series (Cigizoglu, 2002b). In the present study, there are some gaps in available discharge and suspended sedimen data. There are several statistical methods for filling these gaps. In this study, the gaps are filled by taking mean of previous and next data, and also by interpolation and extrapolation where needed.

Techniques such as artificial neural networks (ANNs) and evolutionary algorithms (genetic algorithms, genetic programming) have already entered the modeling arena and are currently understood as the principal means for merging various sources of information. This, in very few words, constitutes a new direction that can be identified as that of data mining in hydroinformatics. It can be expected that in the years to come we shall concentrate our efforts more and more on the analysis of the data we acquire from natural or artificial sources and that we shall mine for knowledge from the data so acquired (Babovic, 2000).

Evolutionary algorithms effectively provide an alternative approach to problem solving, where solutions of the problem are evolved rather than the problems being solved directly. Genetic programming (GP), which was first proposed by Koza (1992), is a branch of evolutionary algorithms, which creates computer programs that can be converted to symbolic form. Gene-expression programming (GEP) is a new variant of GP, which evolves computer programs of different sizes and shapes encoded in linear chromosomes of fixed length. The chromosomes are composed of multiple genes, each gene encoding a smaller subprogram. Further, the structural and functional organization of the linear chromosomes allows the unconstrained operation of important genetic operators such as mutation, transposition, and recombination (Ferreira, 2001a,b).

This research explores the use of GEP in modeling suspended sediment yield (SSY)-mean discharge (Q) relation in Euphrates River on Middle Euphrates Basin. First attempt is to directly estimate the monthly averaged SSY values taken from sediment gauges on Euphrates River based on the corresponding discharge values and in this way, to obtain SSY rating curves based on GEP technique. The second attempt is to investigate the optimal input set containing time-lagged SSY (SSY_{t-n}; n:1,2,…,d) and Q (Q_{t-n}) in order to estimate the SSY value for present time. In this method, the author tries to obtain time-series models for each station, which gives the monthly averaged SSY. The database for modeling studies has been taken from field measurements of Turkish Electrical Power Resources Survey and Development (EİEİ). Finally, the proposed GEP models are compared to conventional rating curve and multiple regression techniques. The results are presented in tabular form, and also in time history and scatter diagram forms.

1.2 Objective of Thesis

This thesis aims to investigate the use of gene-expression programming (GEP) as an alternative tool for the empirical modeling for monthly averaged suspended sediment yield (SSY) based on corresponding discharge (Q) values for the same period, and also based on time-lagged SSY and Q values as time-series model. The study will be based on field data taken from Turkish Electrical Power Resources Survey and Development (EİEİ), which correspond to monthly average SSY and Q values measured at five stations on Euphrates River in Middle Euphrates Basin. An extensive literature survey on both conventional and soft computing techniques was performed. Among a wide range of soft computing (SC) techniques, Neural Networks, Genetic programming and Neuro-fuzzy approaches were investigated. Two GEP models were developed based on (1) Q as single input parameter, and (2) time-lagged Q and SSY as multi-input parameters. Additionally, sediment rating curve (SRC) and multi linear regression (MLR) models were developed based on the same data for the same periods. The accuracy of GEP, SRC and MLR models were evaluated with field measurements.

1.3 Layout of Thesis

The contents of each chapter are expressed as:

- Chapter 2 contains literature review on conventional and soft computing applications on water resources engineering data and more specifically, on sediment yield estimation.
- Chapter 3 includes overview on genetic programming and main computing techniques: gene-expression programming, sediment rating curve and multi linear regression that will be used in the thesis.
- Chapter 4 presents the study area and the field data used in present study.
- Chapter 5 gives the main steps of the modeling techniques used in this study, and also the formulations of each developed model for the five stations.
- Chapter 6 presents the training and evaluation results of the proposed GEP, SRC and MLR models and discusses on the results.
- Chapter 7 emphasizes conclusions based on the findings of the thesis and suggestions for future work.

CHAPTER 2

LITERATURE REVIEW

2.1 Conventional Methods

A number of attempts have been made to relate the amount of sediment transported by a river to flow conditions such as discharge, velocity and shear stress. However, none of the equations derived have received universal acceptance. Usually, either the weight or the concentration of sediment is related to the discharge. These two forms are often used interchangeably (Aytek and Kişi, 2008). McBean and Al-Nassri (1988) examined this issue and concluded that the practice of using sediment load vs. discharge is misleading because the goodness of fit implied by this relationship is spurious. Instead they recommended that the regression link be established. Over the years, researchers have proposed several rating curves to determine the average relationship between discharge and suspended sediment load (Thomas, 1985; Asselman, 2000; Picoet et al., 2001; Overleir, 2004; Crowder et al., 2007).

Although standard and modified rating curve techniques have been recorded to give successful and satisfactory results, they contain idealization of the complex and nonlinear relationship between turbulent flow and suspended sediment transport phenomenon. These techniques try to fit a unique predefined power relation on field or experimental data, which can not cover the whole inflection points due to seasonal effects (Singh and Durgunoğlu, 1989; Guven, 2009). Therefore, since last two decades new techniques have been proposed as alternative, which are covered under the name of soft computing (Zadeh, 1991). Artificial neural networks (ANNs), fuzzy logic (FL), and genetic programming (GP) are the most pronounced branches of soft computing technique in estimation of water engineering data (Maier and Dandy, 2000; Kisi, 2004, Kisi 2005; Guven et al., 2006; Lohani et al., 2007; Singh et al., 2007; Guven and Gunal, 2008a,b; Azamathulla et al., 20008., Guven 2009).

2.2 Articial Neural Network Applications of Sediment Yield Estimation

There are several studies that ANNs have been successfully validated in estimation of suspended sediment yield. Cigizoglu (2004) investigated the accuracy of a single ANN in estimation and forecasting of daily suspended sediment data. Kisi (2004) used different ANN techniques for daily suspended sediment concentration prediction and estimation and he indicated that multi-layer perceptron could show better performance than the others. Kisi (2005) developed an ANN model for modeling suspended sediment and compared the ANN results with those of the rating curve and multilinear regression. Cigizoglu and Kisi (2006) developed some methods to improve ANN performance in suspended sediment estimation. Tayfur and Guldal (2006) predicted total suspended sediment from precipitation. Cigizoglu and Alp (2006) applied generalized neural network technique in modeling sediment yield in rivers.

Both fuzzy logic (FL) and artificial neural networks are designed to address the issues of nonlinearity and have been a focus of attention recently. Both approaches are non-parametric and produce range-conservative predictions but no error estimations. Fuzzy logic is more interpretable and transparent than artificial neural networks but requires a subjective or automated calibration procedure (Kisi et al. 2006). Kisi (2005) used multiple predictors (discharge and suspended sediment concentration of preceding time step) with fuzzy logic and artificial neural networks and obtained a slightly better performance with fuzzy logic, which was also reported by Lohani et al.(2007).

2.3. Genetic Programming Applications in Water Resources Engineering

It was observed that only a few studies existed in the literature related to the use of GP in the field of water resources engineering. Davidson et al.(1999) and Babovic and Keijzer (2000) determined empirical relationships for the friction in turbulent pipe flow and the additional resistance to flow induced by flexible vegetation, respectively. Cousin and Savic (1997), Savic et al.(1999), Drecourt (1999), Whigham and Crapper (1999, 2001), Babovic and Keijzer (2002) applied GP to rainfall-runoff modeling. Babovic et al. (2001) applied GP to sedimentary particle settling velocity equations. Harris et al. (2003) studied on velocity predictions in compound channels with vegetated floodplains using GP. Dorado et al. (2003) studied on prediction and modeling of the rainfall-runoff transformation of a typical urban basin using artificial neural networks (ANNs) and GP. Giustolisi (2004) determined Chezy resistance coefficient in corrugated channels by using GP. Rabunal et al. (2007) determined the unit hydrograph of a typical urban basin using GP.

Most recently, Aytek and Kisi (2008) applied GP for suspended sediment modeling using daily suspended sediment and daily discharge data taken from United States Geological Survey (USGS), Guven et. al. (2008) applied GP for estimation of reference evapotranspiration by using daily atmospheric variables obtained from the California Irrigation Management Information System (CIMIS) database. Guven and Gunal (2008) applied GEP in prediction of depth and location of maximum scour downstream of grade-control structures based on inflow, scour geometry and soil characteristics. Guven and Aytek (2009) proposed GEP as a new and alternative technique for estimations of stage-discharge relationship in rivers.

Guven (2009) developed linear genetic programming (LGP) based time-series models for estimation of daily flow rate of at a station on Schuylkill River at Berne, PA, USA. LGP is a new variant of GP (Oltean and Groşan, 2003). Azamathulla et al. (2009) predicted the local scour around a bridge pier using GEP and experimental scour data taken from literature, and they compared their GEP model with artificial neural network technique and conventional regression equations.

2.4 Genetic Programming Applications of Sediment Yield Estimation

Only few studies were observed for sediment modeling using GP approach: Babovic (2000) used experimental flume data utilized by Zyserman and Fredsoe (1994) and expressed a new formulation for bed concentration of suspended sediment. The proposed formulation gave successful results in predicting the measured sediment concentration values.

Kizhisseri et al. (2005) used GP methodology to explore a better correlation between the temporal pattern of fluid field and sediment transport by utilizing two datasets; one from numerical model results and other from Sandy Duck field data. They obtained remarkable correlation between the predictions of their GP models and the experimental ones.

Singh et al. (2007) suggested combination of neural networks and genetic programming in estimation of longshore sediment transport rate. They claimed that such a combination was found to produce better results than the individual use of neural networks or genetic programming. They stated that the ability of the neural network to approximate a non-linear function coupled with the efficiency of the genetic programming to make an optimum search over the solution domain seems to result in a better prediction.

Aytek and Kişi (2008) proposed GP as a new approach for the explicit formulation of daily suspended sediment–discharge relationship for two stations of U.S. Geological Survey (USGS). They used the daily streamflow and suspended sediment data from two stations on Tongue River in Montana as case studies. They obtained that the results of the proposed GP formulation performed quite well compared to sediment rating curves and multi-linear regression model.

As it is seen from the past studies, soft computing methods have been successfully applied in sediment transport issue. However, these black-box models do not present a physically interpretable solution. In this study it is aimed to use GEP technique as alternative to the other soft computing techniques, but more significantly GEP is introduced as closed from solution for sediment transport problem. In this context, this study is will be a pioneering study that presents GEP technique in modeling suspended sediment problem.

CHAPTER 3

THEORETICAL BACKGROUND

3.1 Genetic Programming

Genetic programming (GP) is a variant of genetic algorithm (GA) technique, which was first proposed by John Koza (1992). GP enables computers to solve problems without being explicitly programmed. GP solves problems based on the Darwinian principle of reproduction and survival of the fittest and analogs of naturally occurring genetic operations such as crossover (sexual recombination) and mutation. In a sexual reproduction of a DNA is a copy of the parent's DNA, possibly with some random changes, known as mutations. In the sexual reproduction, DNA from both parents is inherited by the new individual. Often about half of each parent's DNA is copied to the child where it joins with the DNA copied from the other parent. The child's DNA is usually different from that in either parent.

GP starts with an initial population of randomly generated computer programs composed of functions and terminals appropriate to the problem domain. The functions may be standard arithmetic operations, standard programming operations, mathematical functions, logical functions, or domain-specific functions. Depending on the particular problem, the computer program may be flowchart-valued, integer-valued, real-valued, complex-valued, vector-valued, symbolic-valued, or multiple-valued. Genetic programming contain a "population" of trial solutions to a problem, typically each individual in the population is modeled by a string representing its DNA. This population is "evolved" by repeatedly selecting the "fitter" solutions and producing new solution from them, the new solutions replacing existing solutions in the population (Koza, 1992).

New individuals are created either asexually (i.e. copying the string) or sexually (i.e. creating a new string from parts of two parent strings). In genetic programming the individuals in the population are computer programs. To ease the process of creating new programs from two parent programs, the programs are written as trees. New programs are produced by removing branches from one tree and inserting them into another. This simple process ensures that the new program is also a tree and so is also syntactically valid. Fig.3.1 shows the steps of GP, which reproduces computer programs to solve problems.

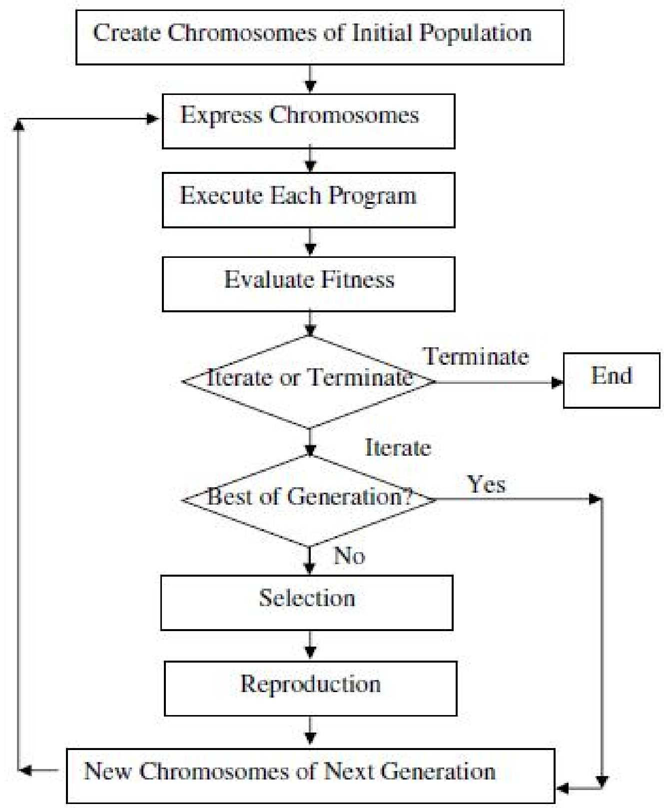


Figure 3.1. Brief flowchart of GP (Koza, 1992)

3.2 Gene-expression Programming

Gene expression programming (GEP) is an extension to GP that evolves computer programs of different sizes and shapes encoded in linear chromosomes of fixed length (Ferreira, 2001a,b). The chromosomes are composed of multiple genes, each gene encoding a smaller subprogram. Further, the structural and functional organization of the linear chromosomes allows the unconstrained operation of important genetic operators such as mutation, transposition, and recombination. One strength of the GEP approach is that the creation of genetic diversity is extremely simplified as genetic operators work at the chromosome level. Another strength of GEP consists of its unique, multigenic nature, which allows the evolution of more complex programs composed of several subprograms. As a result, GEP surpasses the old GP system in 100–10,000 times (Ferreira, 2001a,b; Ferreira, 2006).

The fundamental difference between GA, GP, and GEP is due to the nature of the individuals, namely in GAs the individuals are linear strings of fixed length (chromosomes); in GP the individuals are nonlinear entities of different sizes and shapes (parse trees); and in GEP the individuals are encoded as linear strings of fixed length (the genome or chromosomes), which are afterwards expressed as nonlinear entities of different sizes and shapes (i.e., simple diagram representations or expression trees). Thus, the two main parameters of GEP are the chromosomes and expression trees (ETs). The process of information decoding (from the chromosomes to the ETs) is called translation, which is based on a set of rules. The genetic code is very simple where there exists one-by-one relationships between the symbols of the chromosome and the functions or terminals that they represent. The rules also very simply determine the spatial organization of the functions and terminals in the Ets and the type of interaction between sub-ETs (Ferreira 2001a,b; Ferreira, 2006). That is the reason why two languages are used in GEP: the language of the genes and the language of ETs. A significant advantage of GEP is that it makes possible to infer exactly the phenotype given the sequence of a gene, and vice versa, which is termed as Karva language. For example, an algebraic expression [a+(a/b)]-[a*b] can be represented by a two-gene chromosome or an ET, as shown in Fig.3.2.

GEP technique starts the solution of a specific problem with the random generation of the chromosomes of the initial population. Then the chromosomes are expressed and the fitness of each individual is evaluated. The individuals are then selected according to fitness to reproduce with modification, leaving progeny with new traits. The individuals of this new generation are, in their turn, subjected to the same developmental process: expression of the genomes, confrontation of the selection environment, and reproduction with modification. The process is repeated for a certain number of generations or until a solution has been found (Ferreira, 2001a,b).

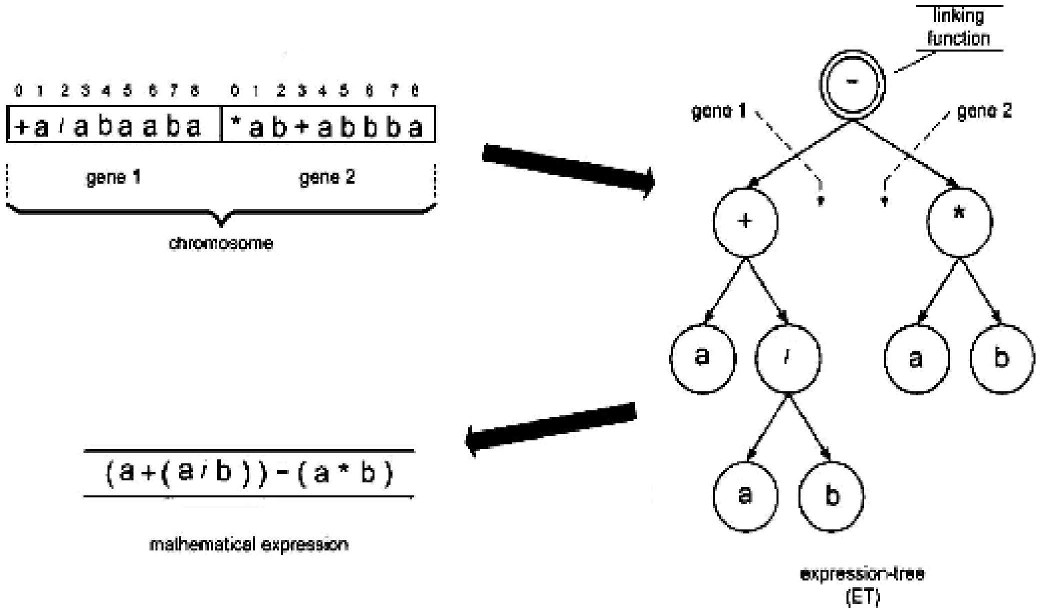


Figure 3.2. Definition sketch of a two-gene chromosome in terms of expression tree (ET), Karva language and mathematical expression (Guven and Gunal, 2008a)

3.3 Conventional Methods for Suspended Sediment Modeling

3.3.1 Sediment Rating Curve (SRC)

The functional relationship between suspended sediment yield (SSY) and discharge (Q) can be established by field measurement of SSY and Q and thereafter, can be expressed as a rating curve. Ideally, a rating curve describes a unique functional relationship between sediment yield and discharge; therefore, it is obtained as a smooth and continuous curve with reasonable degree of sensitivity.

The relation can be estimated by a sufficient number of measurements suitably distributed throughout the range in discharge, taking into account the shape of the sediment-discharge relationship. The number and spacing of the observations are made to conform to the relative frequency of sediment yield at the various discharges; that is, the number of observations at various sub-ranges is in proportion to the probable occurrence of sediment yield at these same ranges, covering the whole range of sediment yield for which the relation is plotted.

The sediment yield-discharge relation may be expressed by an equation (Herschy, 1995) of the form of

$$SSY = aQ^{n} \tag{3.1}$$

which is the equation of a parabola where SSY is the suspended sediment yield, Q is the discharge, a and n are constants. This equation may be transformed by logarithms to

$$\log SSY = \log a + n \cdot \log(Q) \tag{3.2}$$

which is in the form of the equation of a straight line

$$y = n'x + a' \tag{3.3}$$

where n' is the gradient and a' is the intersection of the line on the y axis. By plotting SSY against Q , on double logarithmic graph paper, a straight line is obtained.

After log-transformation to the arithmetic domain and application of the Ferguson (1986) correction factor, the sediment load occurring at a specific discharge can be estimated using the following expression:

$$SSY = F_{c} \cdot a \cdot Q^{n} \tag{3.4}$$

where F_c is the log-transformation bias correction factor. Specifically,

$$F_{c} = e^{MSE/2} \tag{3.5}$$

where e is the exponential function and MSE is the mean square error of the regression equation ($MSE=1/N\Sigma[\log(SSY)-\log(SRC1)]^{2}$). In the present study, first sediment rating curve (Eq. (3.1)) is denoted as SRC1 and the second one with bias correction factor (Eq. (3.4)) is denoted as SRC2.

3.3.2 Multiple Linear Regresion

Multiple linear regression generalizes the simple linear regression model by allowing for many terms in a mean function rather than just one intercept and one slope (Weisberg, 2005). The general multiple linear regression model with response Y and terms X1, . . .,Xp have the form

$$E(Y/X) = \hat{a}0 + \hat{a}1X1 + \cdots + \hat{a}pXp \tag{3.6}$$

The symbol X in E(Y|X) means that all the terms on the right side of the equation are conditioning. Similarly, when we are conditioning on specific values for the predictors x1, . . ., xp that we will collectively call **x**, we write

$$E(Y \mid X = x) = \hat{a}0 + \hat{a}1x1 + \cdots + \hat{a}pxp \tag{3.7}$$

where the âi are unknown parameters have to be estimated. When p = 1, X has only one element, and we get the simple regression problem (in present case p=1, X has only the discharge as input). When p=2, the function corresponds to a plane in three dimensions and when p>2, the fitted function is a hyperplane, the generalization of a p-dimensional plane in a (p+1) dimensional space. In this study, the unknown parameters β0, β1, β1,…,β pare calculated using least squares method.

An advantage of linear regression is that it is easy to implement. The obtained regression coefficients allow the interpretation of the predictors’ influence. Linear regression models are computationally efficient and can also predict confidence intervals for the obtained coefficients and the predicted data, if the underlying statistical assumptions are met (Francke et al, 2008).

CHAPTER 4

STUDY AREA AND DATA USED IN MODELLING STUDIES

4.1. Introduction

In this study, the Middle Euphrates Basin is under consideration. The reason for focusing on this part of Euphrates Basin is in that the sediment stations in the other two sub-basins have so many gaps, and moreover a number of them are closed and data of current periods are not available. Additionally, the majority of dams have been constructed in Upper and Lower Euphrates basins, which affect the sediment yield of the rivers they have been constructed on. However, no dam structures are observed in Middle Euphrates Basin, especially between the sediment stations considered in this study.

4.2 Euphrates Basin

The Euphrates River has its springs in the highlands of Eastern Turkey and its mouth at the Persian Gulf. It is the longest river in Southwestern Asia with 2,700 km, and its actual annual volume is 35.6 billion cubic meters. The Euphrates river is formed in Turkey by two major tributaries; the Murat and the Karasu. These two streams join together and form the Euphrates River and the Keban Dam Lake. The sediment measurement stations on Murat River are under consideration in this study.

Euphrates follows a south-eastern route to enter Syria at Karkamış point. After entering Syria, the Euphrates continues its southeastern course and is joined by two more tributaries, the Khabur and the Balikh. Both of these tributaries have their sources in Turkey and they are the last bodies of water that contribute to the river. After entering Iraq, the river reaches the city of Hit, where it is only 53 m above sea level. From Hit to the delta in the Persian Gulf, for 735 km, the river loses a major portion of its waters to irrigation canals and to Lake Hammar.

The Euphrates Basin is divided into three sub-basins as: Upper Euphrates, Lower Euphrates and Middle Euphrates. Turkish Electrical Power Resources Survey and Development (EİEİ) has totally 39 sediment measurement stations in Euphrates Basin, 16 of them being recently closed. The suspended sediment data of these stations are commonly intermittent. Generally, both discharge and sediment data of the same period are not available, thus the intermittent data can not be filled.

4.3 Data Collection

The data set used in this study was obtained from Turkish Electrical Power Resorces Survey and Development (EİEİ). The time series of monthly averaged daily streamflow and suspended sediment data from five stations on Murat River in Middle Euphrates Region (see Figure 4.1) are used: the upstream station (station no: 2122) below Tutak Town, Hınıs Creek station (station no:2177) at Adıvar Village, Akkonak station (station no: 2174) on Göçmenler Field at Akkonak Village, Göynük Creek station (station no: 2164) at Bingöl City, and the downstream station Palu (station no: 2102) at Elazığ City. The drainage areas at these sites are 5.882 km^2 for the upstream station, 2.995 km^2 for Hınıs Creek, 17.435 km^2 for Adıvar, 2232 km^2 for Göynük Creek, and 25.516 km^2 for the downstream station Palu. Information on the monthly time series for these stations can be acquired with cost from the EİEİ web server (http://www.eie.gov.tr/turkce/YEK/HES/hidroloji/sedim.html).

The data of January 01, 2005–December 30, 2007 (66% of the whole data) were chosen for training, and the data of January 01, 2007–December 30, 2008 (33% of the whole data) were chosen for testing the proposed GEP, SRC and MLR models. Unfortunately, there are some missing data in each station records which cannot be used in training and testing periods. In order to rely on the original data, no filling was done for modeling discharge-suspended sediment yield. However, during the time series modeling of suspended sediment yield, the unavailable data was filled with mean of previous and/or next data points. The estimated results of the filled data points are not shown in figures (Figures.6.1-6.20) that show the model performance in training and testing periods.

The scatter plots of the stations data are given in Figs.4.2-4.6. It can be seen that there is a nonlinear and scattered relationship between discharge and sediment data for each station. Figs. 4.2-4.6 seem to indicate the presence of outliers. For instance, in the station 2102, the suspended sediment yield values of 226,521 ton/day and 306,239 ton day/1 are observed while the other sediment values are below 50,000 ton/day. In station 2122 data set, a suspended sediment load value of 289,631 ton/day is observed while the other values are below 40,000 ton/day. Similar outliers are also observed in the other stations, which are generally used in training period. These values give an additional difficulty to the models in prediction and estimation periods. The models trained using such outliers generally give overestimations of the low sediment values (Aytek and Kişi, 2008).

The monthly averaged statistical parameters of the discharge and suspended sediment data for the stations are given in Table 4.1. In this table, $\overline{X}$, S_x, C_v, C_{sx}, X_{max} and X_{min} denote the mean, standard deviation, coefficient of variation, skewness, maximum and minimum, respectively. The skewness and coefficient of variation of discharge and suspended sediment data each station are high, for both the training and testing data. In the training sediment data, X_{min} and X_{max} values fall in the ranges 5.75–649,380 ton/day. However, the testing sediment data set extremes are x_{min} = 3.77 ton/day, X_{max} =181,430 ton/day. The value of X_{min} for the training sediment data is higher than that for the corresponding testing set for the downstream station. This may cause extrapolation difficulties in estimation of low sediment values.

Table 4.1. The statistical parameters of data set for the stations

Data Set	St.	Data Type	$\overline{X}$	S_x	C_{sx}	C_v	X_{max}	X_{min}
Training	2102	Q	271.28	348.82	1.87	1.29	1439.01	33.06
		SSY	56682.9	134427	3.19	2.37	649379.7	212.4
	2164	Q	27.77	40.68	2.19	1.46	162.87	2.46
		SSY	1345.27	3321.31	3.04	2.47	13596.41	5.75
	2174	Q	153.96	185.33	1.74	1.20	718.61	20.99
		SSY	17497.3	55793.9	5.11	3.19	306238.6	67.83
	2177	Q	29.65	37.12	1.99	1.25	136.13	6.22
		SSY	2605.53	8378.34	4.68	3.22	43074.83	45.07
	2122	Q	56.14	84.28	2.86	1.50	399.85	5.02
		SSY	13947.8	54447.3	5.17	3.90	289630.6	36.35
Testing	2102	Q	178.25	216.92	1.62	1.22	652.65	19.72
		SSY	26012.7	54712.5	2.75	2.10	181429.8	218.6
	2164	Q	19.15	29.72	2.47	1.55	102.46	2.14
		SSY	1499.8	3474.82	2.17	2.32	10153.2	3.77
	2174	Q	88.15	109.22	1.59	1.24	317.82	13.63
		SSY	6786.6	15047.5	2.66	2.22	45483.98	112.0
	2177	Q	19.47	28.90	2.49	1.48	92.51	3.94
		SSY	4056.3	11651.1	3.00	2.87	35118.6	19.76
	2122	Q	28.73	32.85	1.26	1.14	91.29	4.62
		SSY	3536.6	6904.57	1.91	1.95	18964.7	25.39

Note: Q in m^3/s and SSY in ton/day; $C_v=(S_x/\overline{X})$

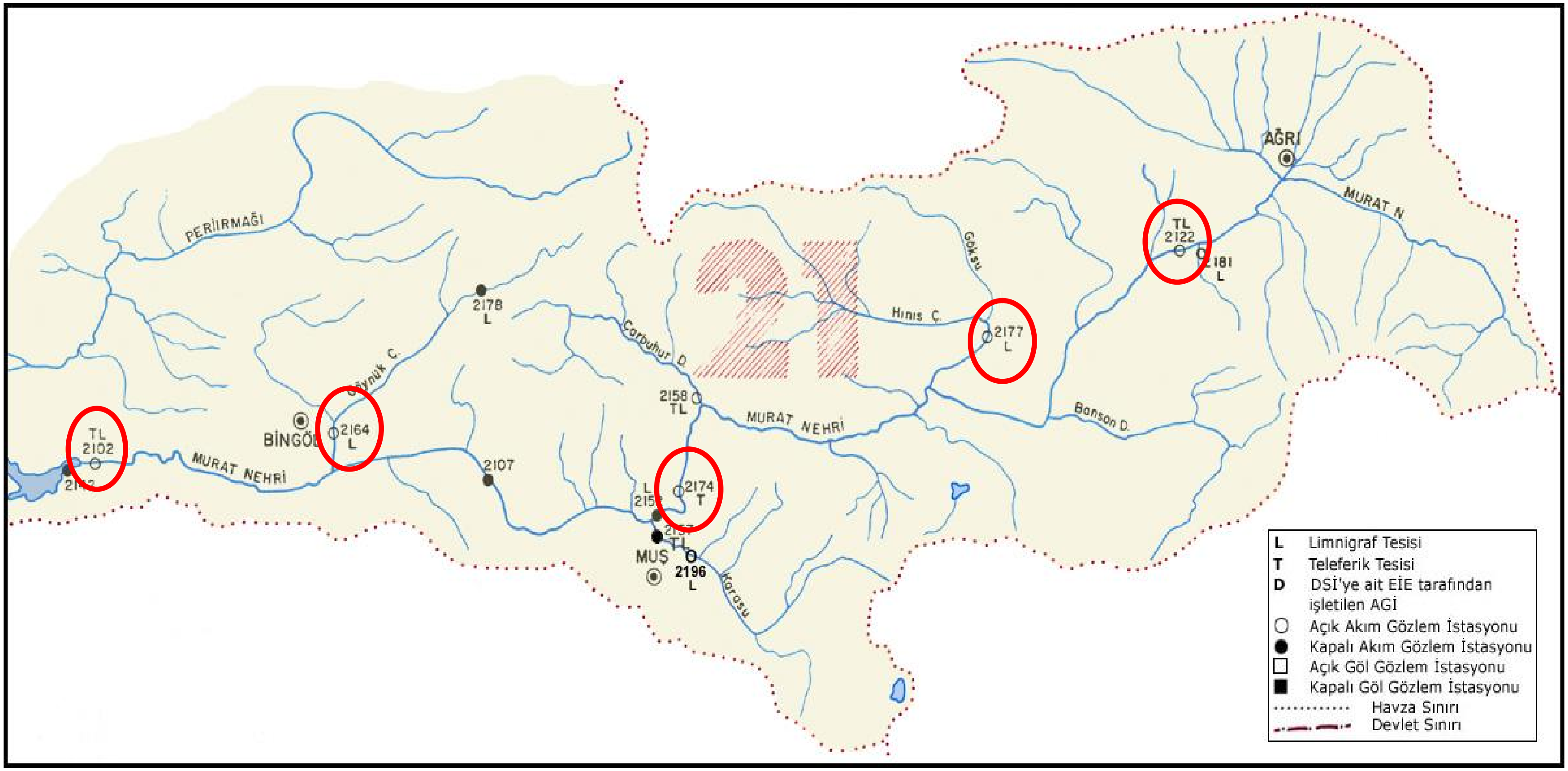


Figure 4.1 Suspended sediment and discharge stations in Middle Euphrates Basin

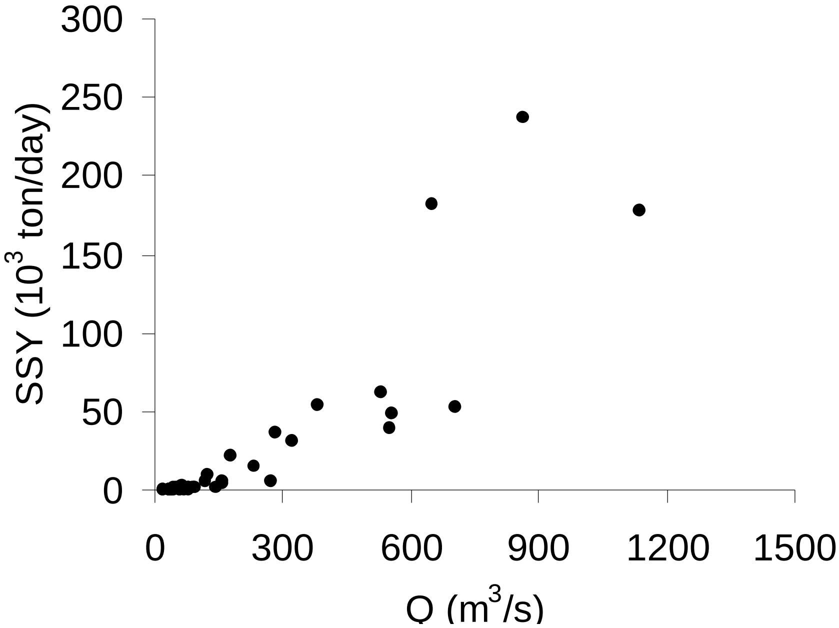


Figure 4.2 Scatter plots of station 2102 data

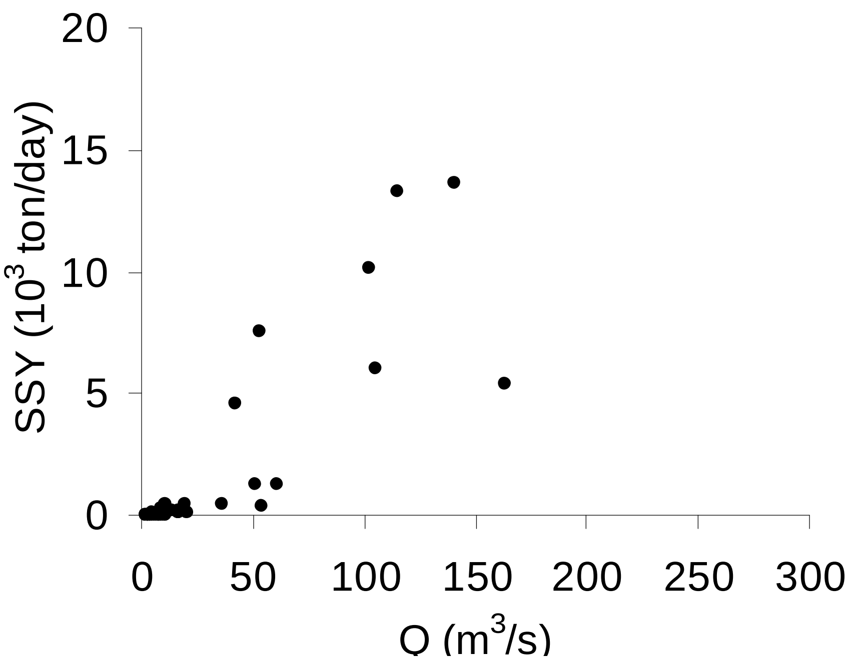


Figure 4.3 Scatter plots of station 2164 data

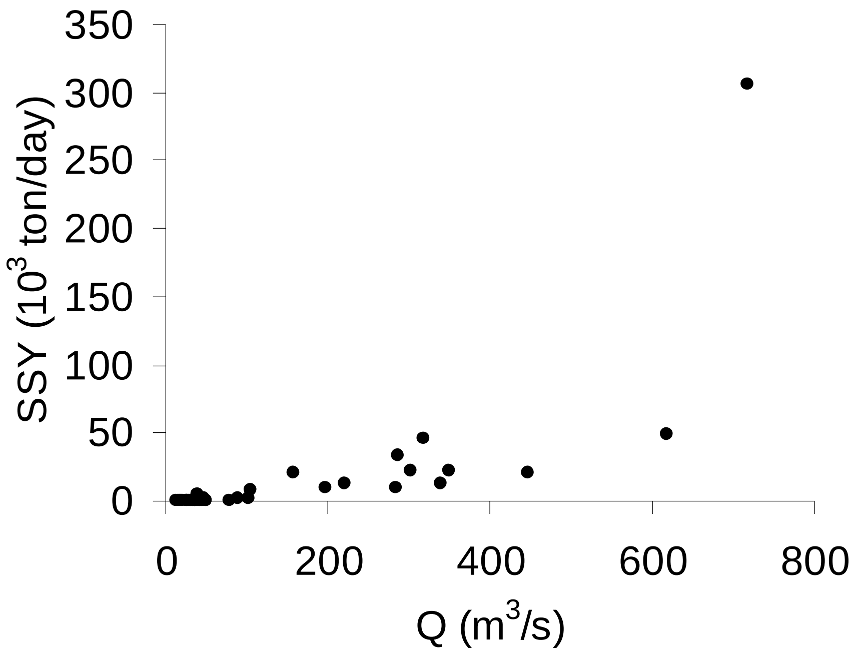


Figure 4.4 Scatter plots of station 2174 data

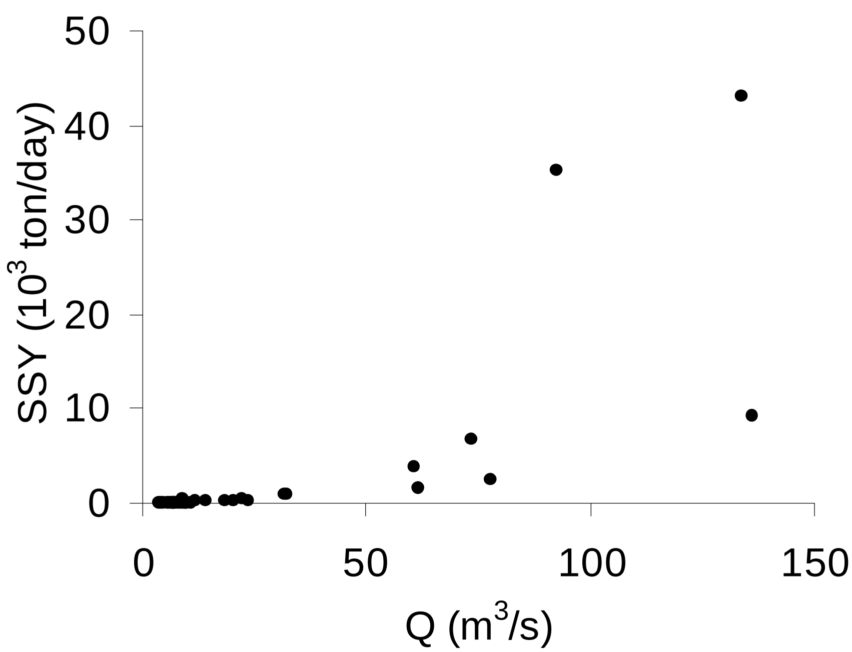


Figure 4.5 Scatter plots of station 2177 data

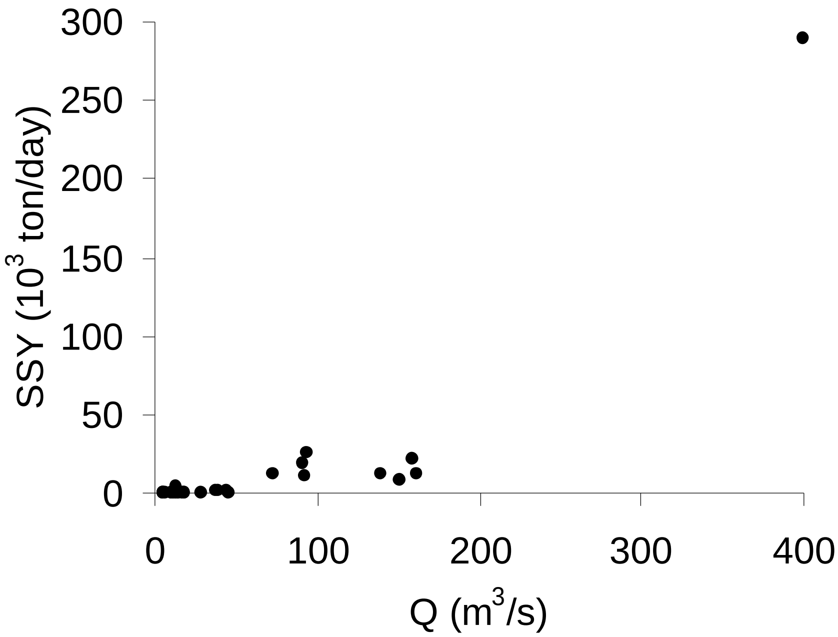


Figure 4.6 Scatter plots of station 2122 data

CHAPTER 5

MODEL DEVELOPMENT

5.1 Introduction

An important issue in model development studies is to find the optimal input set which best represents the considered problem. In the present study, the same issue is to find what degree of time-lagged Q (Q, Q_{t-1}, Q_{t-2},…., Q_{t-n}) and SSY(SSY_{t-1}, SSY_{t-1}, ….,SSY_{t-n}) will be used in prediction of SSY. The most common way is to investigate the correlation between each time-lagged input parameter and the output (SSY). Referring to Tayfur and Guldal (2006), the cross-correlation between SSY and time-lagged discharges (Q, Q_{t-1}, Q_{t-2},…., Q_{t-n}) and time-lagged SSY (SSY_{t-1}, SSY_{t-1}, ….,SSY_{t-n}) was evaluated. Table 5.1 shows the cross-correlation (R) values between SSY and and each time-lagged input parameters. As it can be seen in the table, R value is higher than 0.1 only for inputs Q, Q_{t-1} and SSY_{t-1}. Therefore it was decided to select Q, Q_{t-1} and SSY_{t-1} as input parameters in prediction of SSY.

Two functional relations are used in this study, the first one is the basic discharge-sediment yield relation (Eq. 5.1), and the second one is the relation between SSY and time-lagged Q and SSY (Eq.5.2). The same relationship (Equations 5.1 and 5.2) were used for all model developments: GEP, SRC and MLR. The models developed by using this relation in Equation 5.1 were coded as GEP1, SRC1 and MLR1, and those developed with the relationship in Equation 5.2 were coded as GEP2 and MLR2. Because SRC can be developed only for single Q as an input, it can not be used for modeling the relation in Equation 5.2. Modified SRC, proposed by Ferguson (1986), models were coded as SRC2 in this study.

$$SSY = f(Q) \tag{5.1}$$

$$SSY = f(Q, Q_{t-1}, SSY_{t-1}) \tag{5.2}$$

5.2 GEP Model Development

There are five major steps in preparing to use gene expression programming, and the first is to choose the fitness function. For this problem, the fitness f_i of an individual program i is defined by the following expression:

$$f_i = \sum_{j=1}^{C_t} \left(M - \left| C_{(i,j)} - T_j \right| \right) \tag{5.3}$$

where M is the range of selection, $C_{(i,j)}$ the value returned by the individual chromosome i for fitness case j (out of C_t fitness cases) and T_j is the target value (SSY) for fitness case j. If $|C(i,j) - Tj|$ (the precision) less or equal to 0.01, then the precision is equal to zero, and $f_i = f_{max} = C_t M$. For the present study, M = 100 and, therefore, f_{max} = 1000. The advantage of this kind of fitness function is that the system can find the optimal solution for itself (Ferreira, 2001a).

The second major step consists in choosing the set of terminals T and the set of functions F to create the chromosomes. In this problem, two terminal sets are used based on two input sets, namely T_1 that consists of single independent variable, i.e., T = {Q}, and T_2 that consists of three independent variables, i.e., T = {Q, Q_{t-1}, SSY_{t-1}}. The choice of the appropriate function set is not so obvious, but a good guess can always be done in order to include all the necessary functions (Guven and Gunal, 2008a). In this case, four basic arithmetic operators (+, -, *, /), and some basic functions (√, ln(x), log(x), e^x, 10^x, power) are used.

The third major step is to choose the chromosomal architecture, i.e., the length of the head and the number of genes. Different length of the head, h values were tried for each model and it was observed that h=8, and three genes per chromosome for three-input models (GEP2) and h=6 and two genes per chromosome for single input models (GEP1) gave the best results. The fourth major step is to choose the linking function. In this case, addition and multiplication were tried for each GEP model and generally, sub-ETs linked by multiplication gave better results compared to those linked by addition. And finally, the fifth major step is to choose the set of genetic operators that cause variation and their rates. Table 5.2 lists the combination of all genetic operators (mutation, the three kinds of transposition, and the three kinds of recombination) used in this study. Parameter for the training of the GEP model is given in Table 5.2.

The expression tree (ET) presentations of GEP models for each station (Figs.5.1-5.9) are given in the Appendix. The corresponding simplified explicit formulations of GEP1 model for each station are determined as follows:

GEP1 for station 2102:

$$SSY = -2.81Q^2 + 14.371Q^{1.5} - 14.362Q - 267.03Q^{0.5} - 2183.715 \tag{5.4}$$

GEP1 for station 2164:

$$SSY = 2Q - e^{0.061Q} + \frac{Q^3}{10+Q} + 33.166 \tag{5.5}$$

GEP1 for station 2174:

$$SSY = e^{0.017Q - 0.14} - 0.142Q^2 + 107.52Q - 1478.067 \tag{5.6}$$

GEP1 for station 2177:

$$SSY = 4.33Q^2 - 5 \times 1.065^Q - 5.903Q - 29.70 \tag{5.7}$$

GEP1 for station 2122:

$$SSY = Q^2\left(0.089(Q+1.56)^{0.5} - 2\right) + 99.811Q \tag{5.8}$$

The simplified explicit formulations of the corresponding GEP2 models for the five stations are determined as follows.

GEP2 for station 2102:

$$SSY = \frac{Q}{S_{t-1}}(Q_{t-1} - Q)\frac{(Q+9.015)S_{t-1}}{(-6.987Q_{t-1})} - 3.04\left(Q^{1.5} + 2Q_{t-1}\right) + 28.314Q - 9.221 \tag{5.9}$$

GEP2 for station 2164:

$$SSY=\frac{S_{t-1}^{2}-9.9S_{t-1}}{Q-S_{t-1}-7.687}+S_{t-1}+57.23Q-313.519-2.476Q_{t-1}+\frac{Q(Q_{t-1}-S_{t-1})-S_{t-1}}{S_{t-1}-9.274Q+0.543} \quad (5.10)$$

GEP2 for station 2174:

$$SSY=\frac{Q}{Q_{t-1}}(16.61Q-4.506Q_{t-1})+\frac{S_{t-1}-(Q\times Q_{t-1}-S_{t-1})}{717.837-Q}-(Q_{t-1}(S_{t-1}-9.982Q))^{0.5} \quad (5.11)$$

GEP2 for station 2177:

$$SSY=\frac{Q^{3.5}}{7.31(S_{t-1}\times Q_{t-1})^{0.5}}+\frac{Q(1.835-Q_{t-1})}{Q_{t-1}-8.284-Q_{t-1}^{0.5}}+1.233Q^{0.5}(S_{t-1}+Q)^{0.5}+4.966Q \quad (5.12)$$

GEP2 for station 2122:

$$SSY=1.313Q^{2}+-393.413Q+9626.95+\frac{13313.51Q}{S_{t-1}+9.118}+56.51Q_{t-1} \quad (5.13)$$

It should be noted that the proposed GEP formulation is valid for the ranges of Train set used in this study, thus it covers almost all possible cases that can be faced in practice.

5.2 Development of SRC and MLR Models

The model parameters of SRC1, SRC2, MLR1 and MLR2 models for the five stations are given in Table 5.3 As it is seen from the Table 5.3, only “a” parameter is modified in SRC2 with respect to corresponding SRC1 model, whereas “n” values remain the same. MLR models were developed based using regression analysis toolbox of Microsoft Office 2003 Excel software ®. SRC models were developed by using curve fit toolbox of Microsoft Office 2003 Excel ®.

Table 5.1 Correlation matrix

	Q	SSY	Q_{t-1}	Q_{t-2}	Q_{t-3}	SSY_{t-1}	SSY_{t-2}	SSY_{t-3}
Q	1							
SSY	**0.738**	1						
Q_{t-1}	0.616	**0.331**	1					
Q_{t-2}	0.221	0.091	0.616	1				
Q_{t-3}	-0.0004	-0.038	0.221	0.616	1			
SSY_{t-1}	0.335	**0.196**	0.738	0.331	0.091	1		
SSY_{t-2}	0.051	-0.001	0.335	0.738	0.331	0.196	1	
SSY_{t-3}	-0.0451	-0.051	0.051	0.335	0.738	-0.001	0.196	1

Table 5.2 Parameters of GEP models used in this study

p_1	Function set	+, -, *, /, √, e^x, x^2,x^3, power
p_2	Chromosomes	20-30
p_3	Head size	6-8
p_4	Number of genes	2-3
p_5	Linking function	Addition, Multiplication
p_6	Fitness function error type	MAE(Mean absolute error)
p_7	Mutation rate	0.044
p_8	Inversion rate	0.1
p_9	One-point recombination rate	0.3
p_{10}	Two-point recombination rate	0.3
p_{11}	Gene recombination rate	0.1
p_{12}	Gene transposition rate	0.1

Table 5.3 Comparison of results obtained by standard rating curve (SRC1), Ferguson (1986)'s method (SRC2)

Station	Method	Model	Eqn. No.
2102	SRC1	$SSY=0.211Q^{2.006}$	(5.14)
	SRC2	$SSY=1.329*0.211Q^{2.006}$	(5.15)
	MLR1	$SSY=274.816*Q-17870.1$	(5.16)
	MLR2	$SSY=318.23*Q-108.395*Q_{t-1}+0.12*SSY_{t-1}-7154.2$	(5.17)
2164	SRC1	$SSY=1.269Q^{1.736}$	(5.18)
	SRC2	$SSY=1.480*1.269Q^{2.006}$	(5.19)
	MLR1	$SSY=69.173*Q-575.69$	(5.20)
	MLR2	$SSY=71.089*Q-4.609*Q_{t-1}-0.002*SSY_{t-1}-471.713$	(5.21)
2174	SRC1	$SSY=0.591Q^{1.816}$	(5.22)
	SRC2	$SSY=1.547*0.591Q^{1.816}$	(5.23)
	MLR1	$SSY=216.685*Q-15863.2$	(5.24)
	MLR2	$SSY=239.94*Q-54.28*Q_{t-1}+0.096*SSY_{t-1}-11951.4$	(5.25)
2177	SRC1	$SSY=1.508Q^{1.836}$	(5.26)
	SRC2	$SSY=1.163*1.508Q^{1.836}$	(5.27)
	MLR1	$SSY=167.766*Q-2369.18$	(5.28)
	MLR2	$SSY=204.304*Q-42.91*Q_{t-1}-0.166*SSY_{t-1}-1404.62$	(5.29)
2122	SRC1	$SSY=1.832Q^{1.817}$	(5.30)
	SRC2	$SSY=1.624*1.832Q^{1.817}$	(5.31)
	MLR1	$SSY=551.435*Q-17011.5$	(5.32)
	MLR2	$SSY=632.2*Q-140.141*Q_{t-1}+0.13*SSY_{t-1}-12857.6$	(5.33)

CHAPTER 6

RESULTS AND DISCUSSION

6.1 Training Results of Estimation of SSY based on Q

Table 6.1 lists the statistical performance of each model in prediction of SSY in training period, in terms of coefficient of determination (R^2) and the root mean square error (RMSE). Figs. 6.1-6.5 show the observed and estimated sediment yield values of the five stations for training period, respectively, in the form of hydrograph and scatter plot. As seen from Fig.6.1, the GEP1 model approximates the corresponding observed suspended sediment values better (R^2=0.517, RMSE=92250.03 ton/day) than the rating curve SRC1 (R^2=0.422, RMSE=102048.95 ton/day), SRC2 (R^2=0.422, RMSE=107275.78 ton/day) and multiple regression MLR1 (R^2=0.509, RMSE=92923.04 ton/day) techniques.

$$R^2 = \left(\frac{\sum_{i=1}^{N} (m_i - m')(p_i - p')}{\sqrt{\sum (m_i - m')^2 \sum (p_i - p')^2}} \right)^2 \tag{6.1}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (m_i - p_i)^2}{N}} \tag{6.2}$$

where m_i is the measured value and p_i is the predicted output, m' and p' are the mean values of m_i and p_i and N is number of total number of data.

MLR1 performs better than SRC1 and SRC2 models, however, MLR1 method provides negative approximations for some of the low sediment values, whereas the other model estimates are all positive. SRC2 seems to have a better accuracy than the SRC1 model.

Parallel findings are observed in Figs. 6.2-6.5 and Table 6.1, which show the predicted and SSY values for stations 2164, 2174, 2177 and 2122, respectively. The overall performance of GEP1 outperforms the other techniques, which indicate that GEP1 model learned much better than the non-linear relationship given in Eqn.5.1 compared to the other models.

Table 6.1 Statistical performance of the models in training period

Station	RMSE (ton/day)				R^2			
	GEP1	SRC1	SRC2	MLR1	GEP1	SRC1	SRC2	MLR1
2102	**92250.03**	102048.95	107275.78	92923.04	**0.517**	0.422	0.422	0.509
2164	**1020.54**	2027.21	1851.53	1739.72	**0.907**	0.685	0.685	0.718
2174	**4946.32**	33526.49	38081.90	38081.90	**0.993**	0.682	0.682	0.518
2177	**1159.64**	7718.94	6812.44	6812.44	**0.980**	0.636	0.636	0.552
2122	**3931.14**	36518.35	25541.82	27852.25	**0.995**	0.943	0.943	0.729

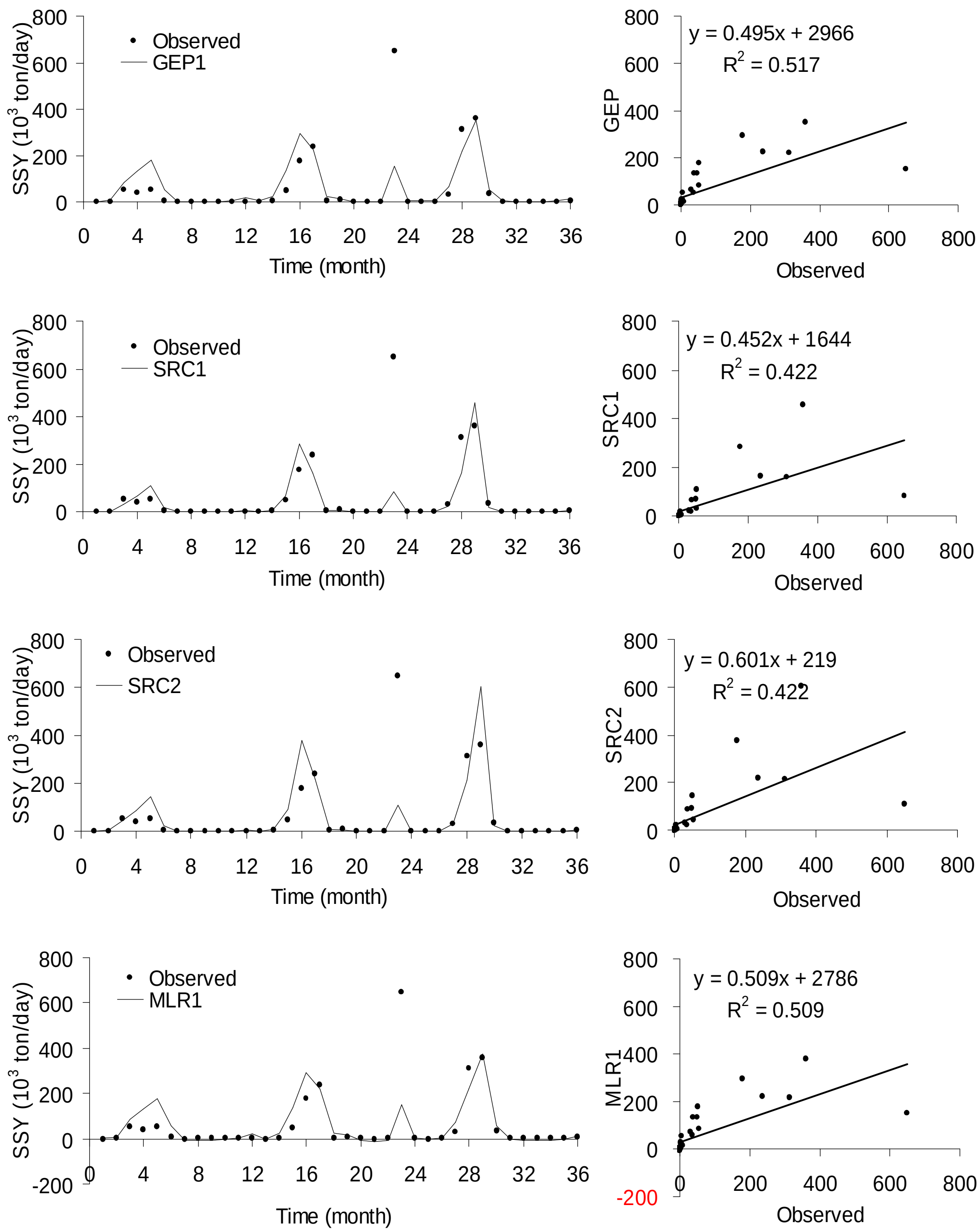


Figure 6.1 The observed and estimated suspended sediments of station 2102 in training period.

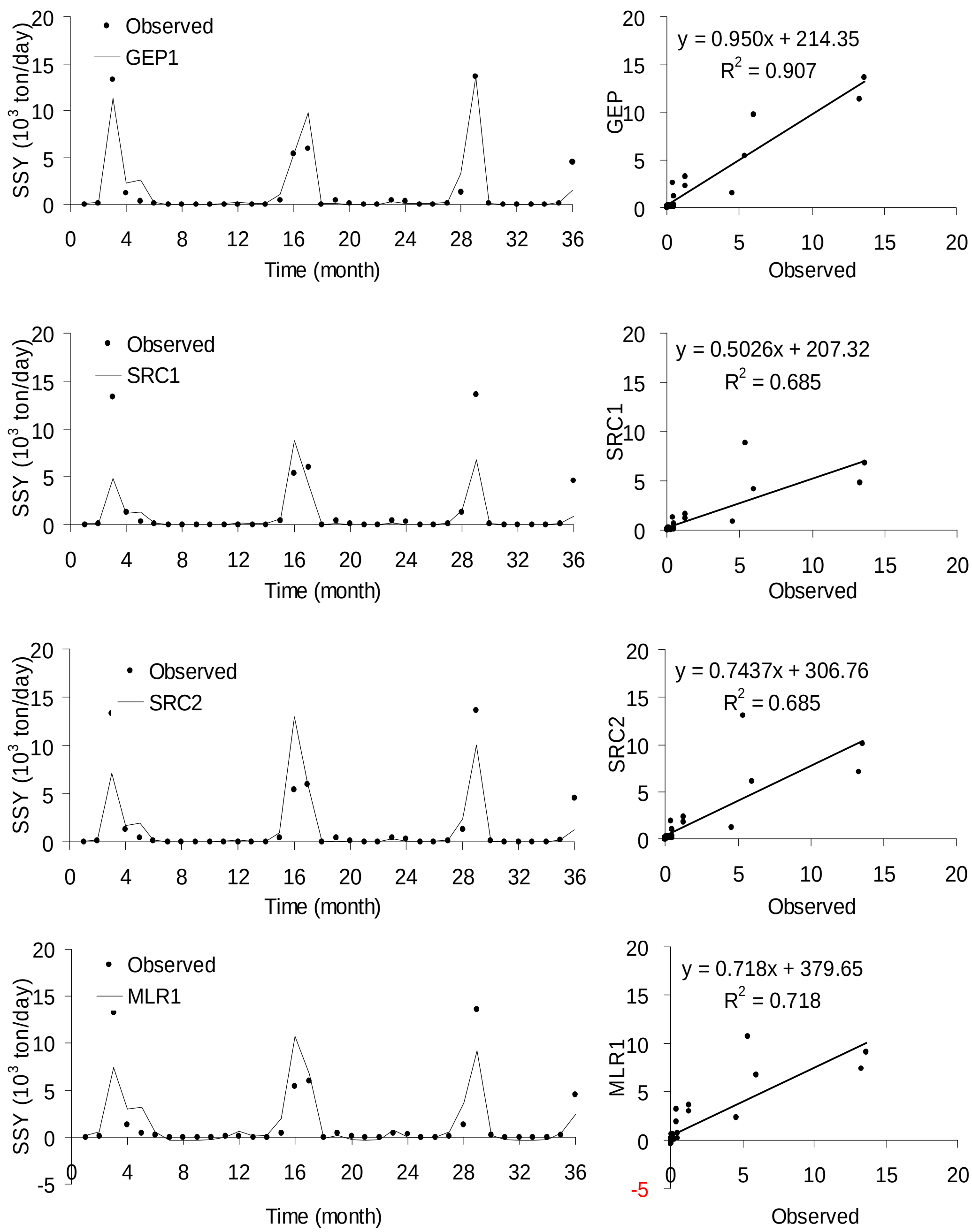


Figure 6.2 The observed and estimated suspended sediments of station 2164 in training period.

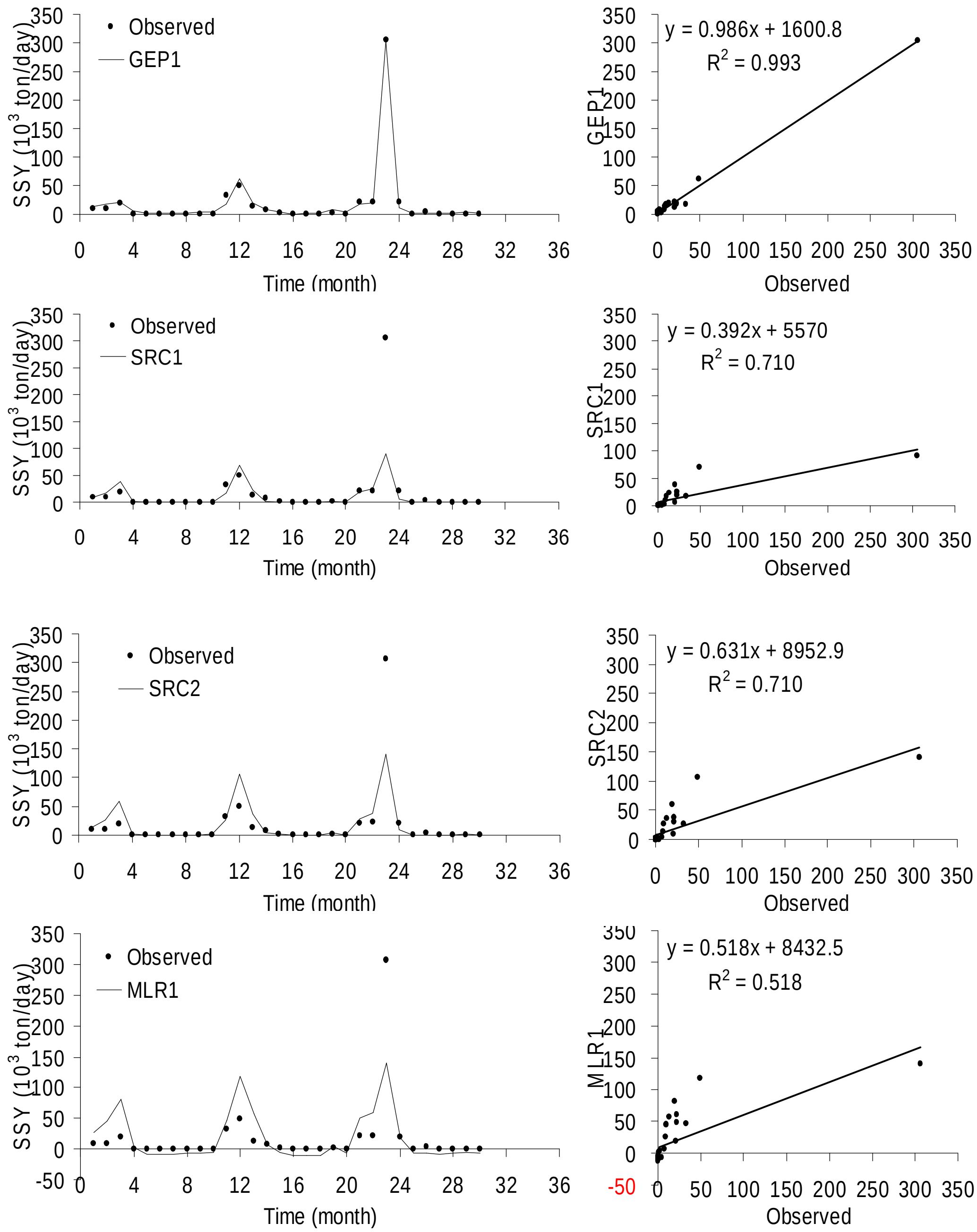


Figure 6.3 The observed and estimated suspended sediments of station 2174 in training period.

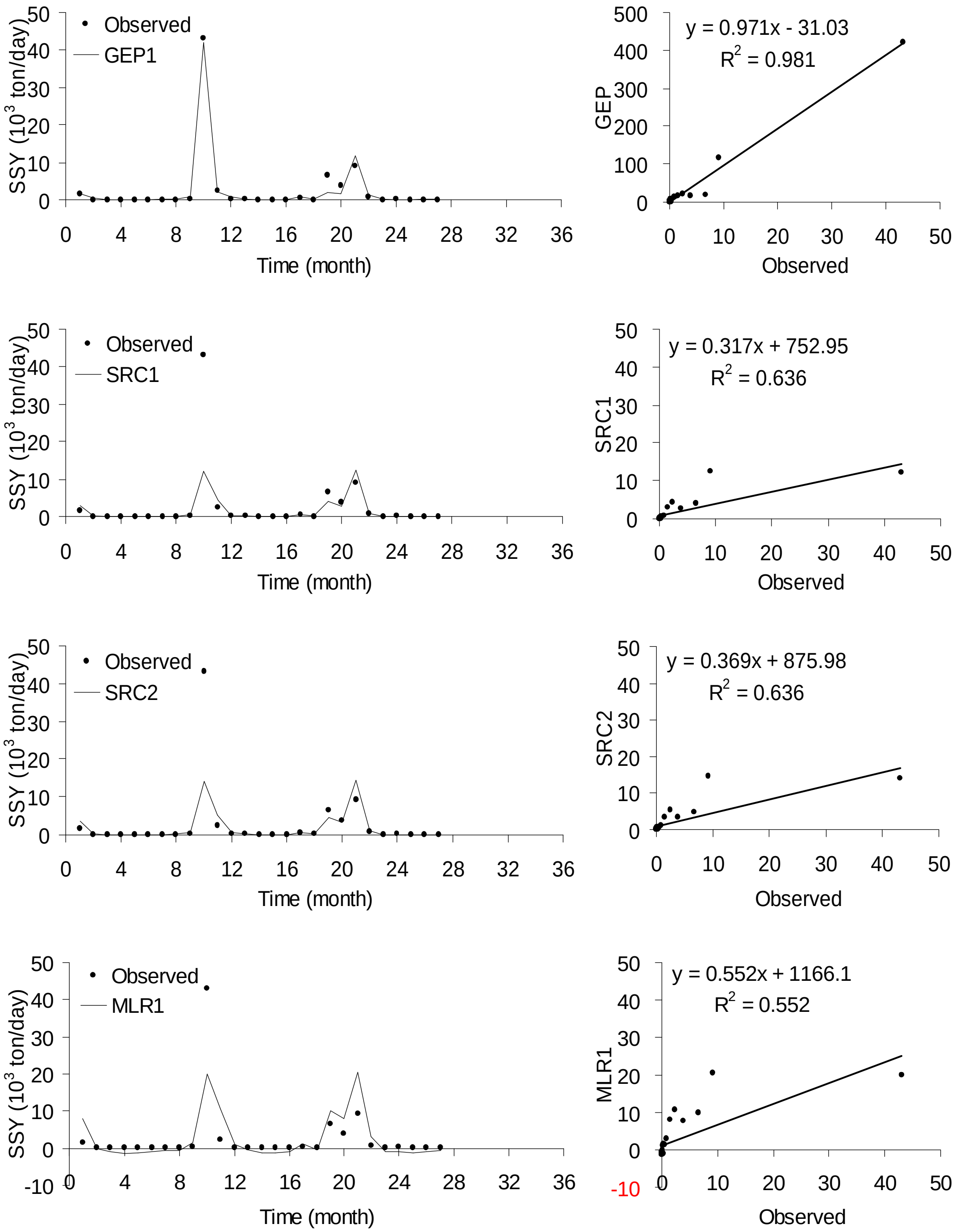


Figure 6.4 The observed and estimated suspended sediments of station 2177 in training period.

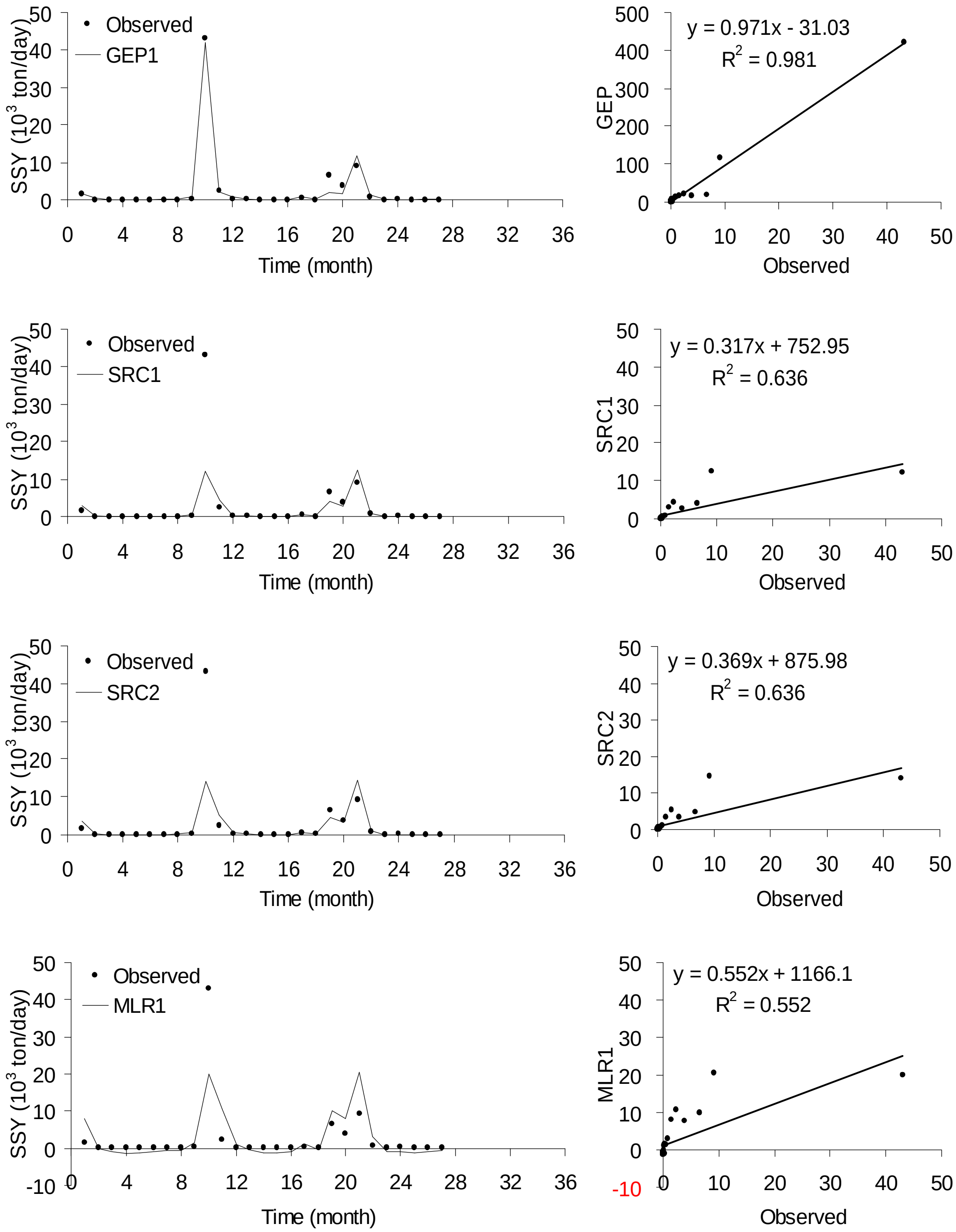


Figure 6.5 The observed and estimated suspended sediments of station 2122 in training period.

6.2 Testing Results of Estimation of SSY based on Q

Table 6.2 lists the statistical performance of each model in prediction of SSY in testing period for the five stations, in terms of R^2 and RMSE. Figs.6.6-6.10 show the observed and estimated sediment yield values of the five stations for testing period, respectively, in the form of hydograph and scatter plot. As seen from Fig. 6.6, the rating curve SRC1 (R^2=0.915), SRC2 (R^2=0.915) models predict the corresponding observed suspended sediment values better than GEP1 model (R^2=0.855) and multiple regression MLR1 (R^2=0.825) technique. However, Table 6.2 indicates that the predictions of GEP1 model are much closer to observed ones due to much lower error (RMSE=22307.9 ton/day) compared to SRC1 (RMSE=26966.8 ton/day), SRC2 (RMSE=22882.43 ton/day) and MLR1 (RMSE=24311.49 ton/day) models. Although R^2 is an important statistical measure showing the linear correlation between the observed and predicted values, sometimes it contradicts with error measures such as RMSE, which directly show how accurate the predictions of a predictive technique by measuring the error between the predicted and observed values.

As in training period, MLR1 method is observed to provide negative approximations for some of the low sediment values in testing period, whereas the other model estimates are all positive. SRC2 (RMSE=22882.43 ton/day) seems to have a better accuracy than the SRC1 (RMSE=26966.8 ton/day) model.

Parallel findings are observed in Figs. 6.7-6.10 and Table 6.2, which show the predicted and SSY values for stations 2164, 2174, 2177 and 2122, respectively. The overall performance of GEP1 outperforms the other techniques with respect to RMSE criteria, but SRC1 and SRC2 models gave better correlation with the observed SSY values for the five stations, compared to GEP1 and MLR1 models (see Table 6.2). It can be deduced from the overall results that GEP1 model is more generalized compared to other techniques.

Table 6.2 Statistical performance of the models in testing period

Station	RMSE				R^2			
	GEP1	SRC1	SRC2	MLR1	GEP1	SRC1	SRC2	MLR1
2102	**22307.9**	26966.8	22882.43	24311.49	**0.855**	0.915	0.915	0.825
2164	**1485.92**	2554.55	2067.79	1688.70	**0.839**	0.871	0.871	0.934
2174	**9356.41**	4812.63	11490.66	11490.66	**0.720**	0.942	0.942	0.838
2177	**253.81**	9329.52	7463.42	7463.42	**1.000**	0.987	0.987	0.911
2122	**3694.12**	4749.37	3155.93	12177.70	**0.875**	0.961	0.961	0.875

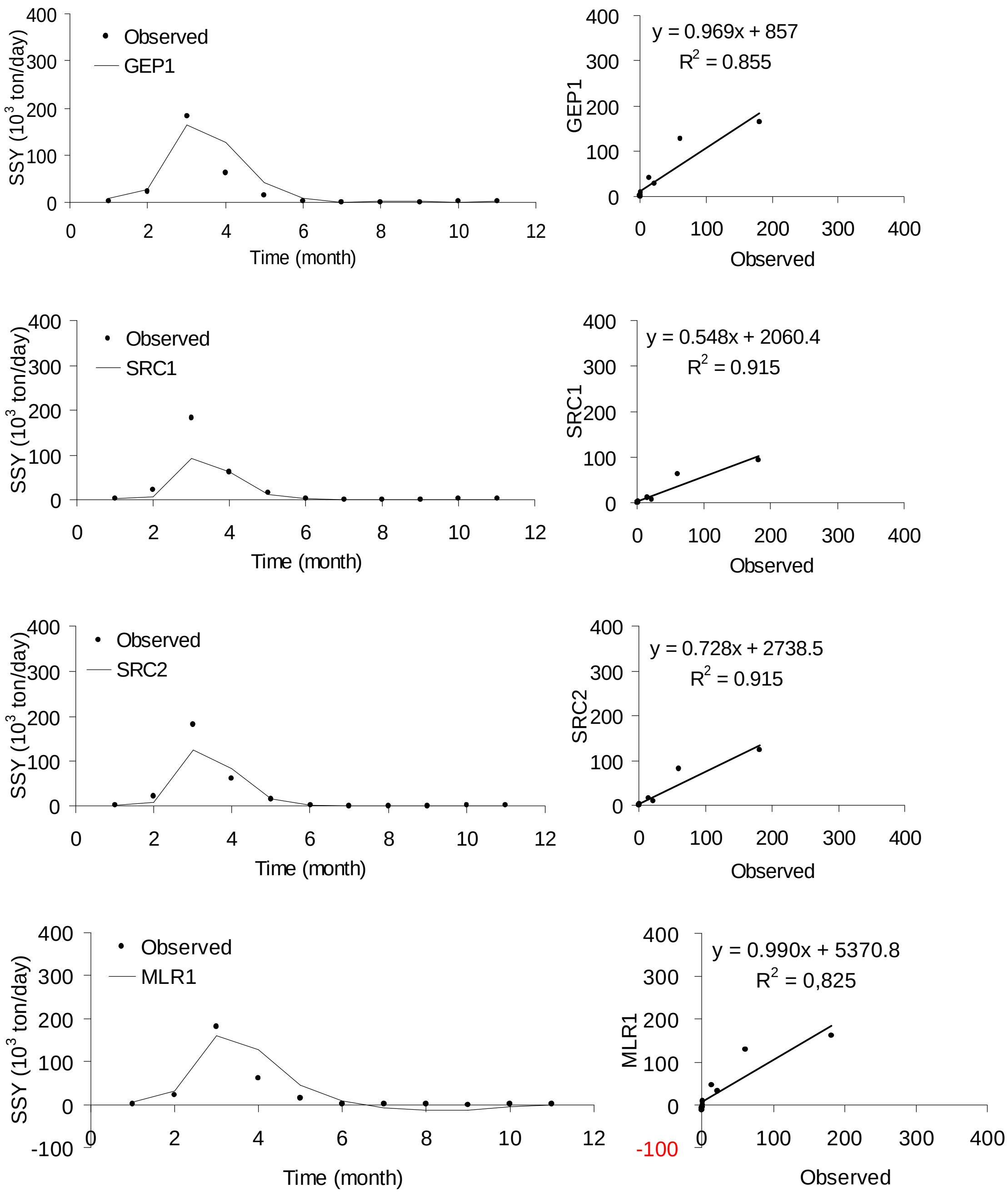


Figure 6.6 The observed and estimated suspended sediments of station 2102 in testing period.

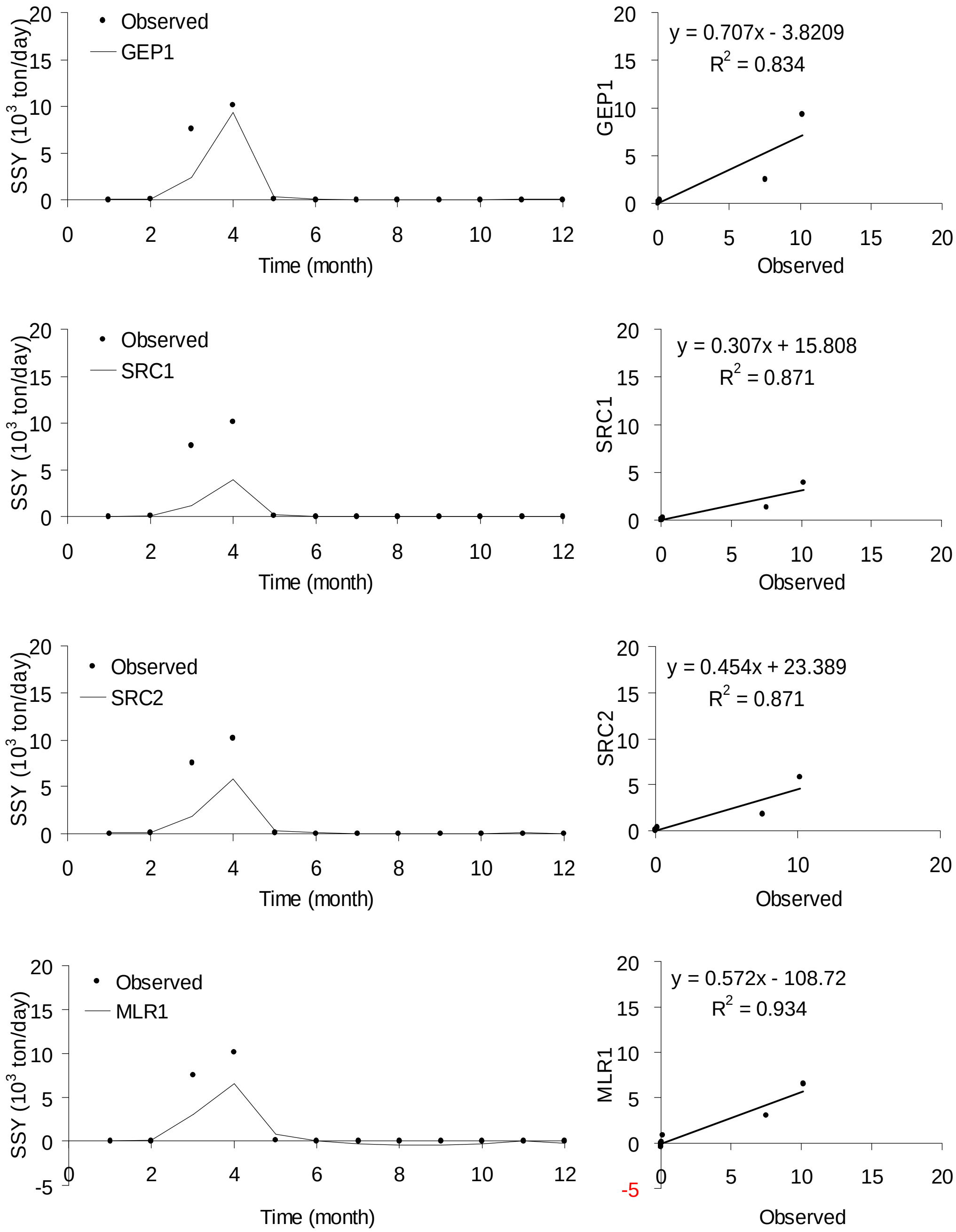


Figure 6.7 The observed and estimated suspended sediments of station 2164 in testing period.

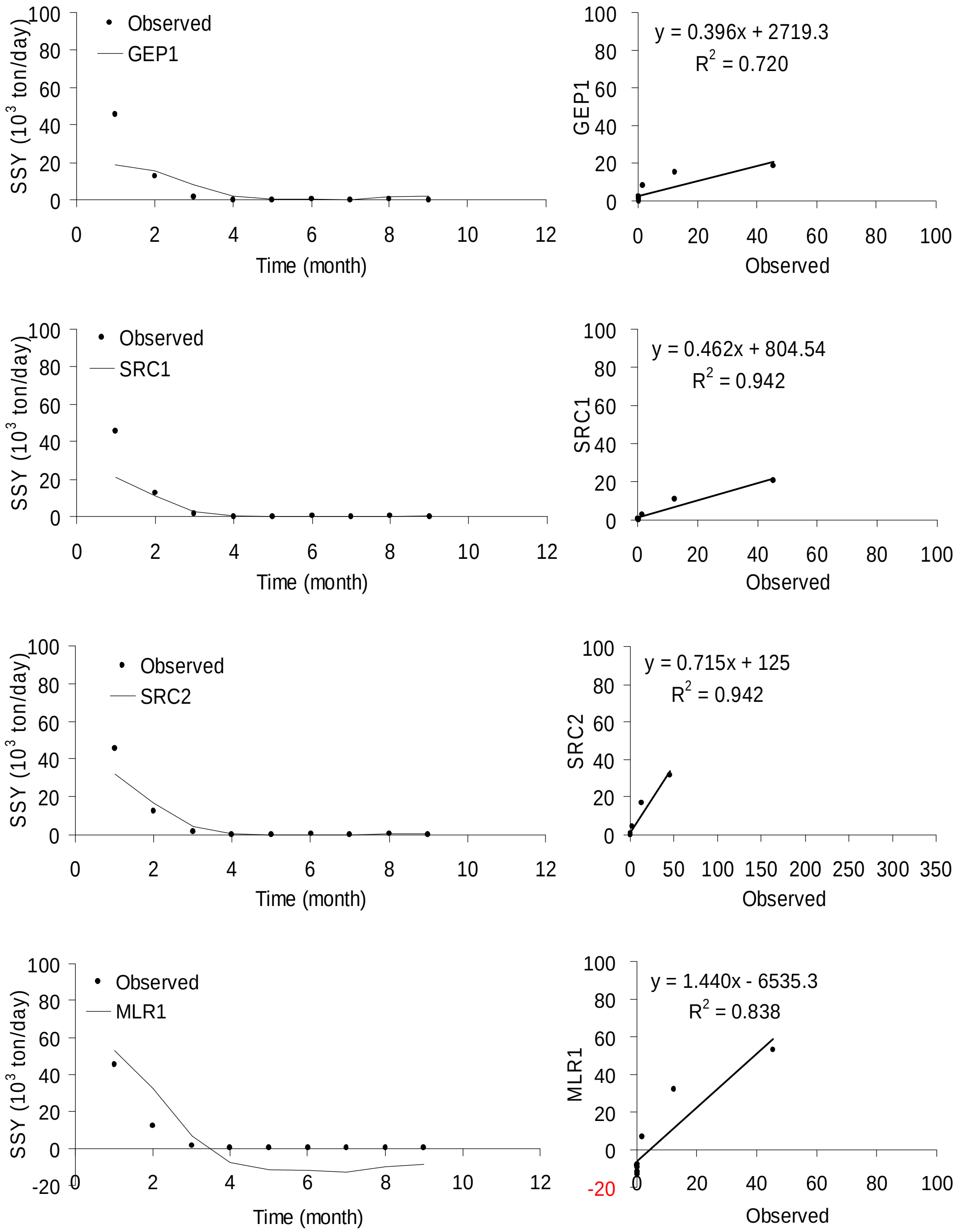


Figure 6.8 The observed and estimated suspended sediments of station 2174 in testing period.

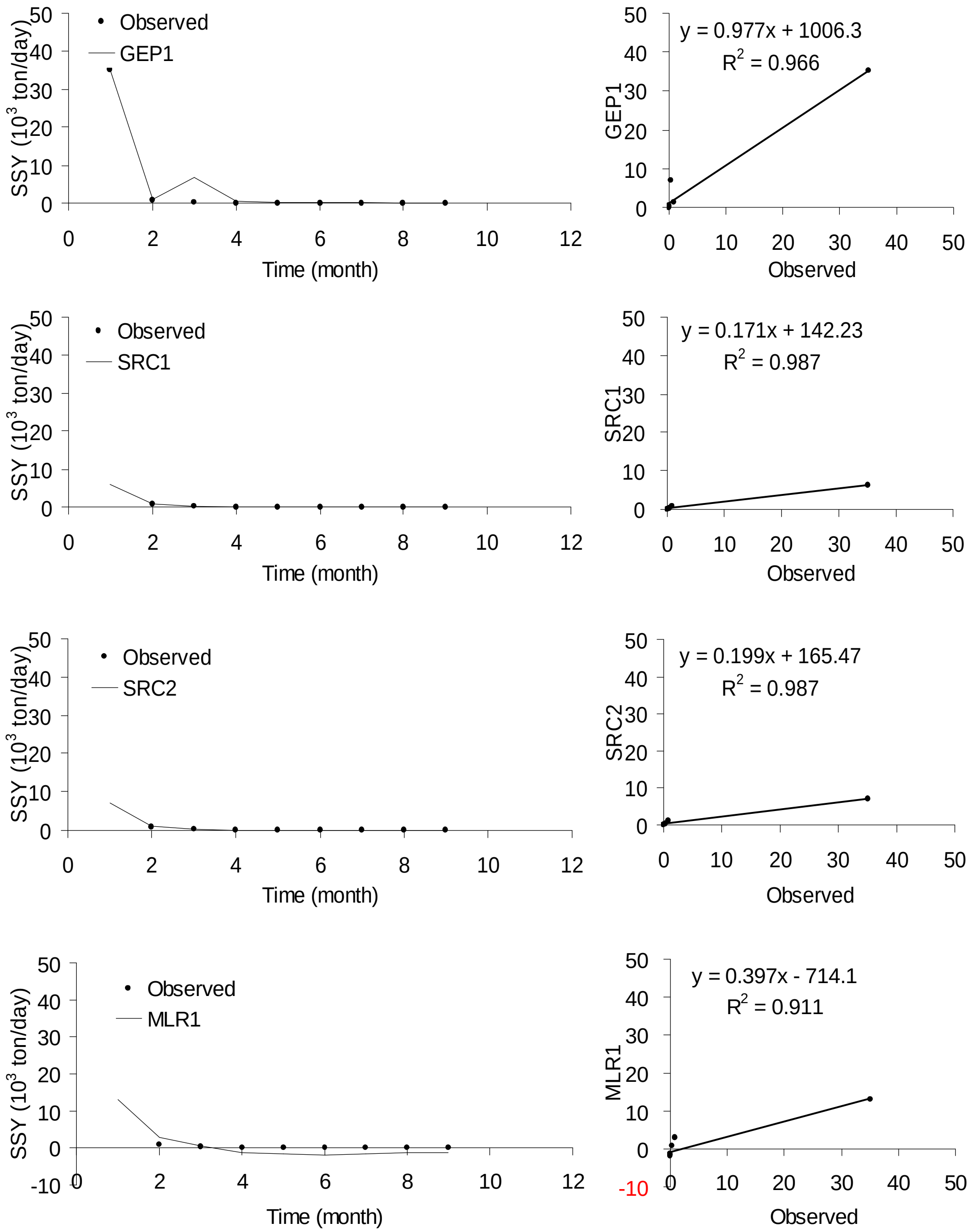


Figure 6.9 The observed and estimated suspended sediments of station 2177 in testing period.

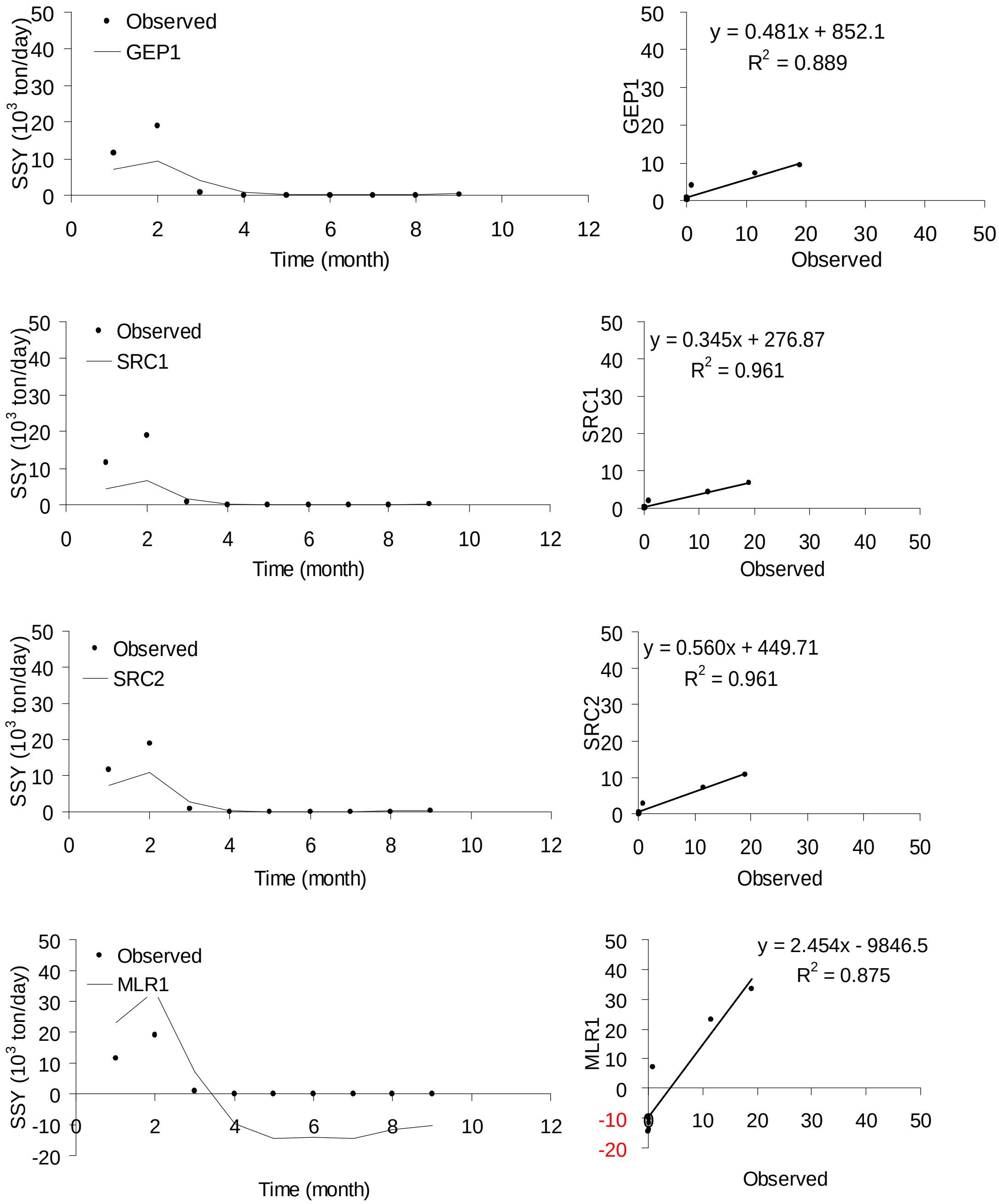


Figure 6.10 The observed and estimated suspended sediments of station 2122 in testing period.

6.3 Training Results of Estimation of SSY based on Time Series of Q and SSY

Table 6.3 lists the statistical performance of each model in prediction of SSY in training period, in terms of R^2 and RMSE. Figs.6.11-6.15 show the observed and estimated sediment yield values of the five stations for training period, respectively, in the form of hydograph and scatter plot. From Fig.6.11 and Table 6.3, it is obvious that GEP2 model approximates the corresponding observed suspended sediment values much better (R^2=0.944, RMSE=31500.3 ton/day) than multiple regression MLR2 (R^2=0.540, RMSE=89939 ton/day) technique. MLR2 method provides negative approximations for some of the low sediment values, which may be explained by very high standard error between MLR2 predictions and the observed SSY ones, and thus resulting negative values especially for low SSY values.

Parallel findings are observed in Figs.6.12-6.15 and Table 6.3, which show the predicted and SSY values for stations 2164, 2174, 2177 and 2122, respectively. GEP2 performs dominantly compared to MLR2, which indicate that GEP2 model learned much better the non-linear relationship given in Eqn.5.2, compared to MLR2.

Table 6.3 Statistical performance of the models in training period

Station	RMSE		R^2	
	GEP2	MLR2	GEP2	MLR2
2102	**31500.30**	89939.39	**0.944**	0.540
2164	**914.79**	1725.42	**0.932**	0.723
2174	**39926.49**	54062266.62	**0.995**	0.525
2177	**6038.79**	4555814.33	**0.996**	0.622
2122	**3944.68**	28535.30	**0.995**	0.722

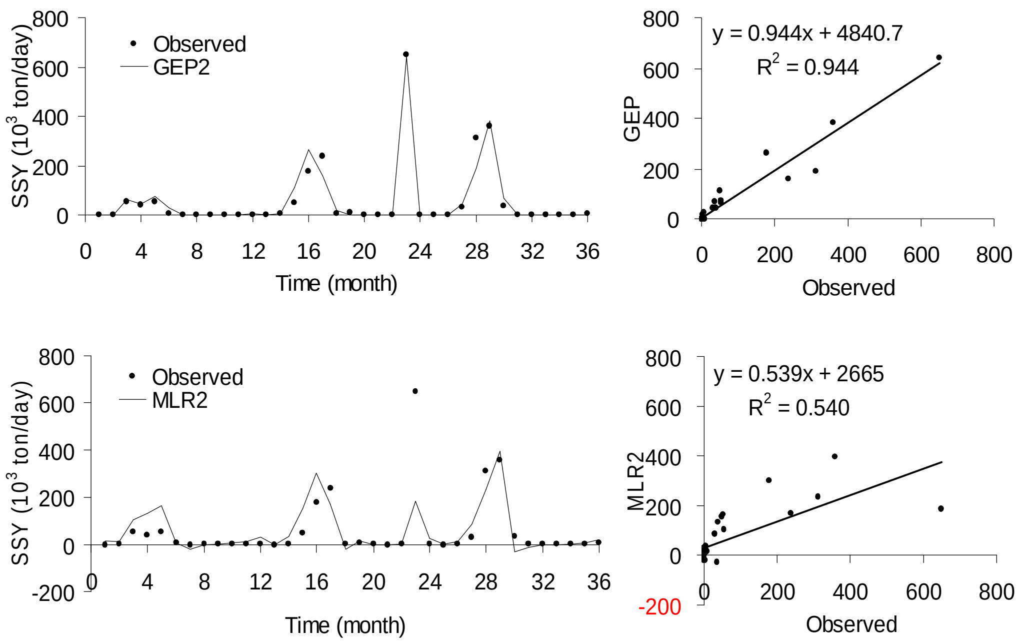


Figure 6.11 The observed and estimated suspended sediments of station 2102 in training period.

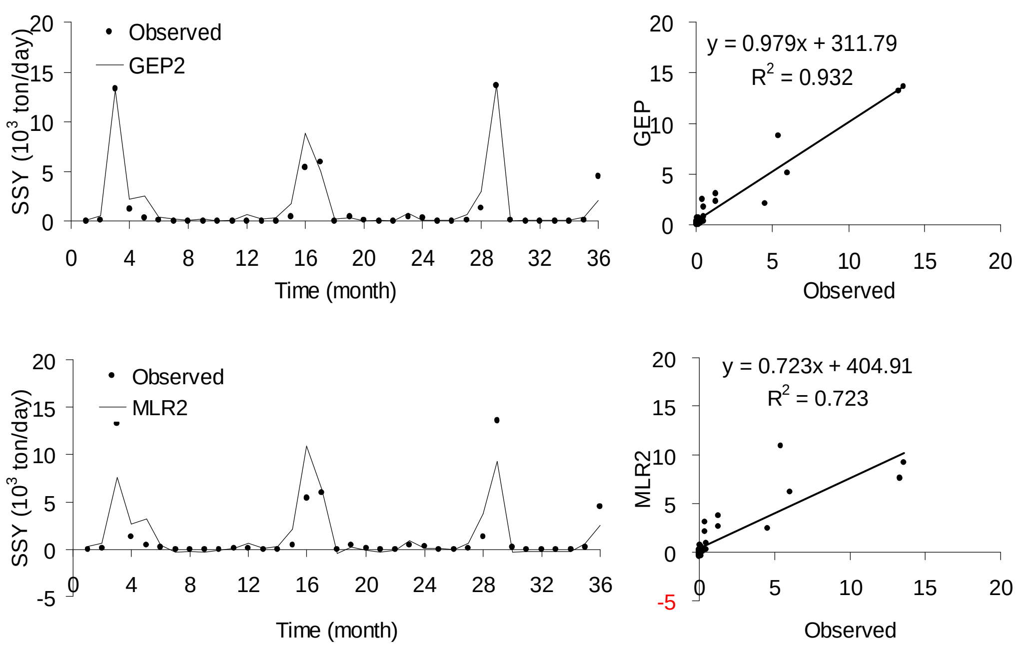


Figure 6.12 The observed and estimated suspended sediments of station 2164 in training period.

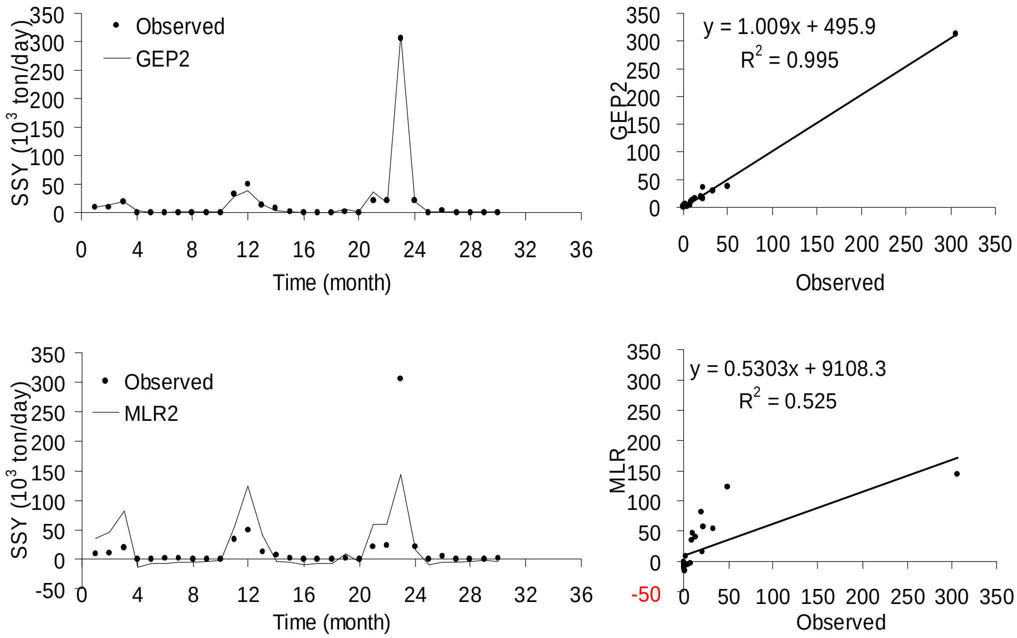


Figure 6.13 The observed and estimated suspended sediments of station 2174 in training period.

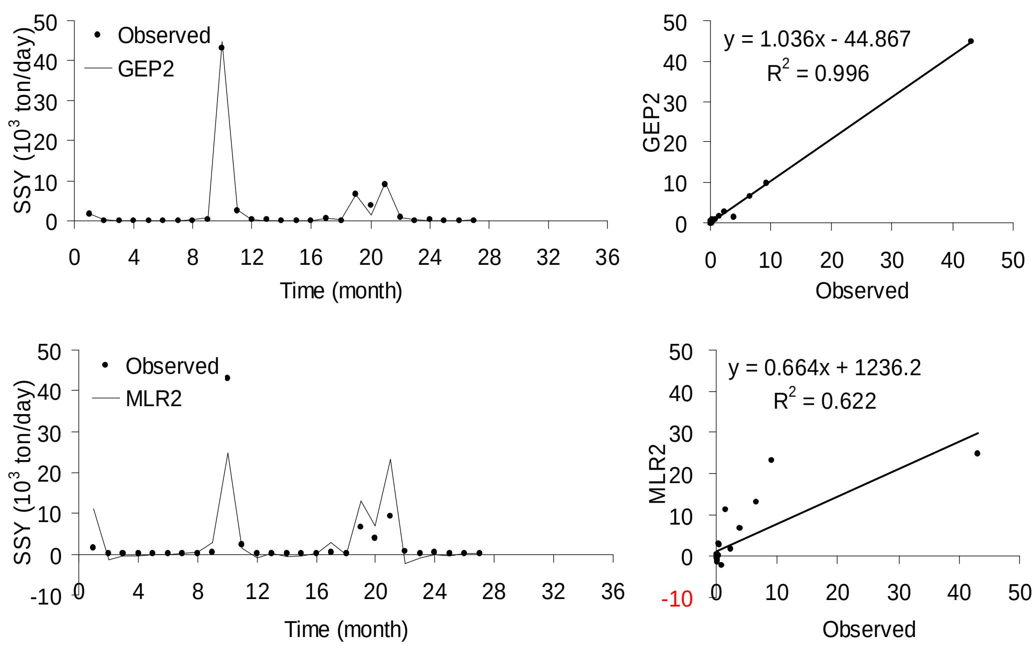


Figure 6.14 The observed and estimated suspended sediments of station 2177 in training period.

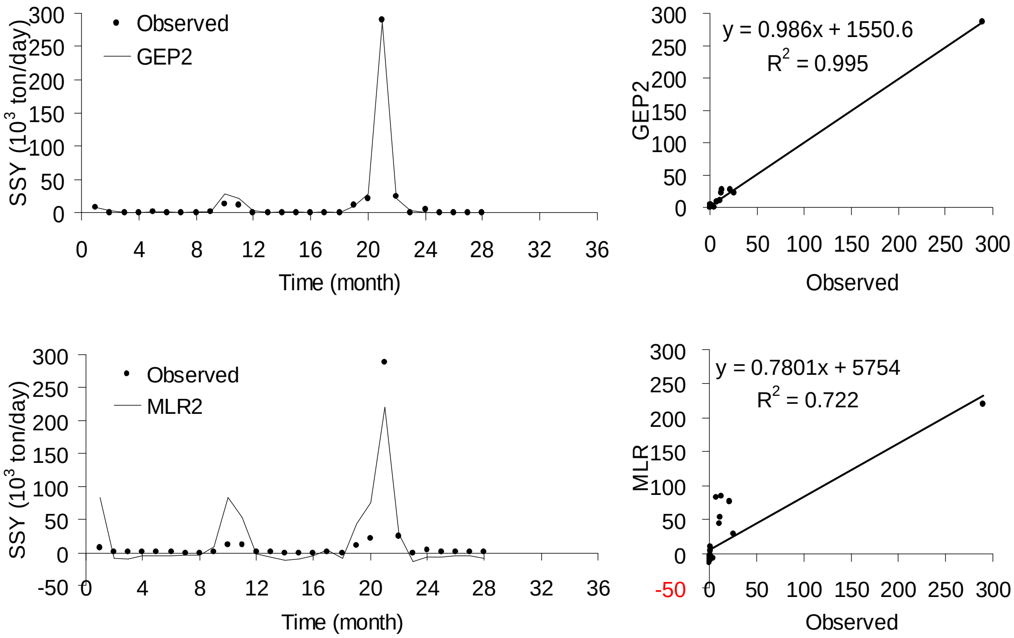


Figure 6.15 The observed and estimated suspended sediments of station 2122 in training period.

6.4 Testing Results of Estimation of SSY based on Time Series of Q and SSY

Table 6.4 lists the statistical performance of each model in prediction of SSY in testing period for the five stations, in terms of R^2 and RMSE. Figs .6.16-6.20 show the observed and estimated sediment yield values of the five stations for testing period, respectively, in the form of hydrographs and scatter plot. As seen from Fig.6.16, GEP2 model (R^2=0.992, RMSE=9537.21 ton/day) predict the corresponding observed suspended sediment values better than multiple regression MLR2 (R^2=929, ton/day, RMSE=17346.64 ton/day) technique.

As in training period, MLR2 method is observed to provide negative approximations for some of the low sediment values in testing period (see Figs.6.16-6.20.) for all stations, whereas the GEP2 estimates are all positive.

Parallel findings are observed in Figs.6.17-6.20 and Table 6.4, which show the predicted and SSY values for stations 2164, 2174, 2177 and 2122, respectively. The overall performance of GEP2 MLR1 (see Table 6.4), except for the sattion 2164, in which MLR2 (R^2=0.953, RMSE= 1630.73 ton/day) predictions to the observed ones are closer than those of GEP2 (R^2=952, RMSE=2022.64 ton/day). It can be deduced from the ovearll results that GEP2 model is more generalized compared to MLR2.

Table 6.4 statistical performance of the models in testing period

Station	RMSE		R^2	
	GEP2	MLR2	GEP2	MLR2
2102	**9537.21**	17346.64	**0.992**	0.929
2164	**2022.64**	1630.73	**0.952**	0.953
2174	**8306.64**	242981296.19	**0.996**	0.935
2177	**9663.13**	87202.52	**1.000**	0.931
2122	**2525.14**	11911.19	**0.981**	0.908

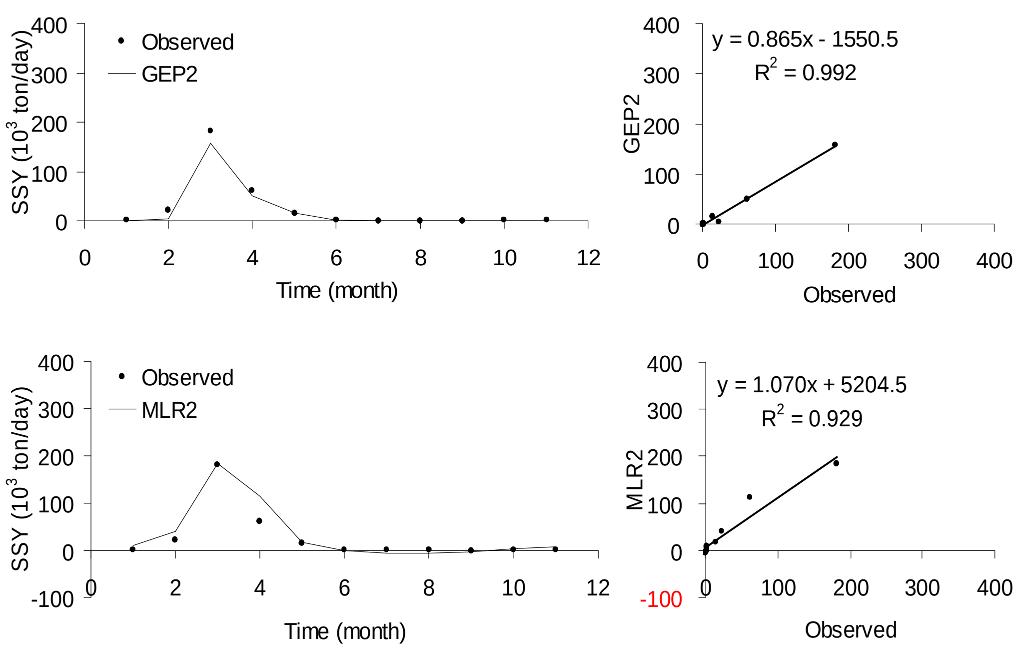


Figure 6.16 The observed and estimated suspended sediments of station 2102 in testing period.

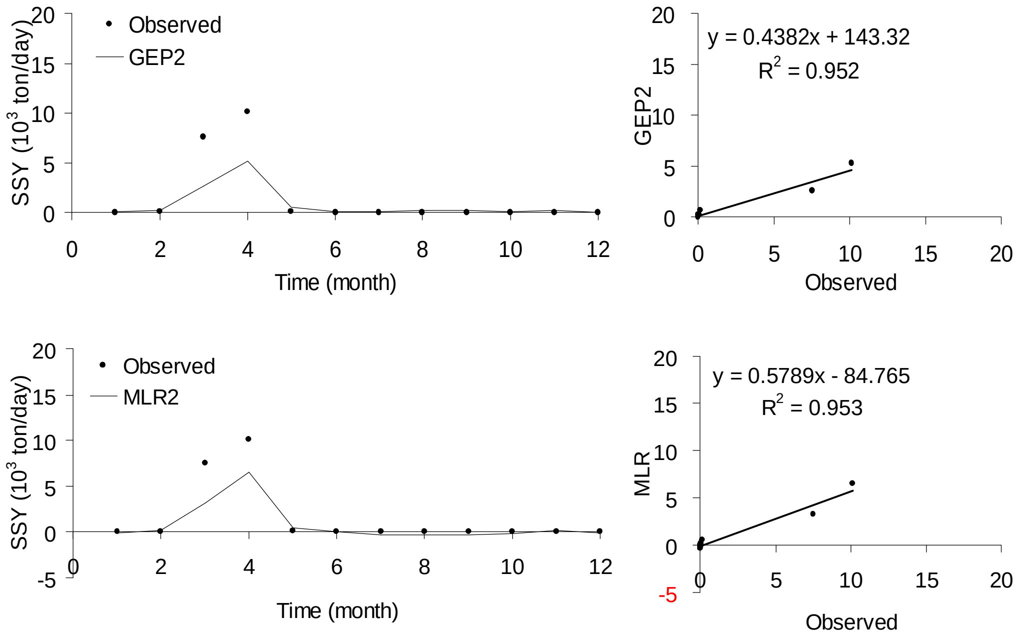


Figure 6.17 The observed and estimated suspended sediments of station 2164 in testing period.

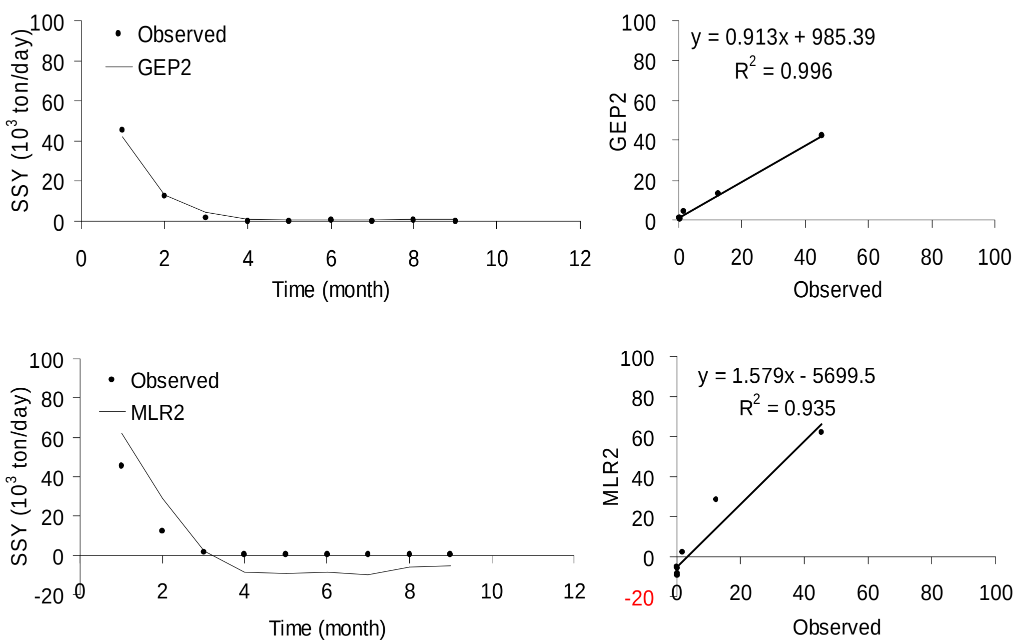


Figure 6.18 The observed and estimated suspended sediments of station 2174 in testing period.

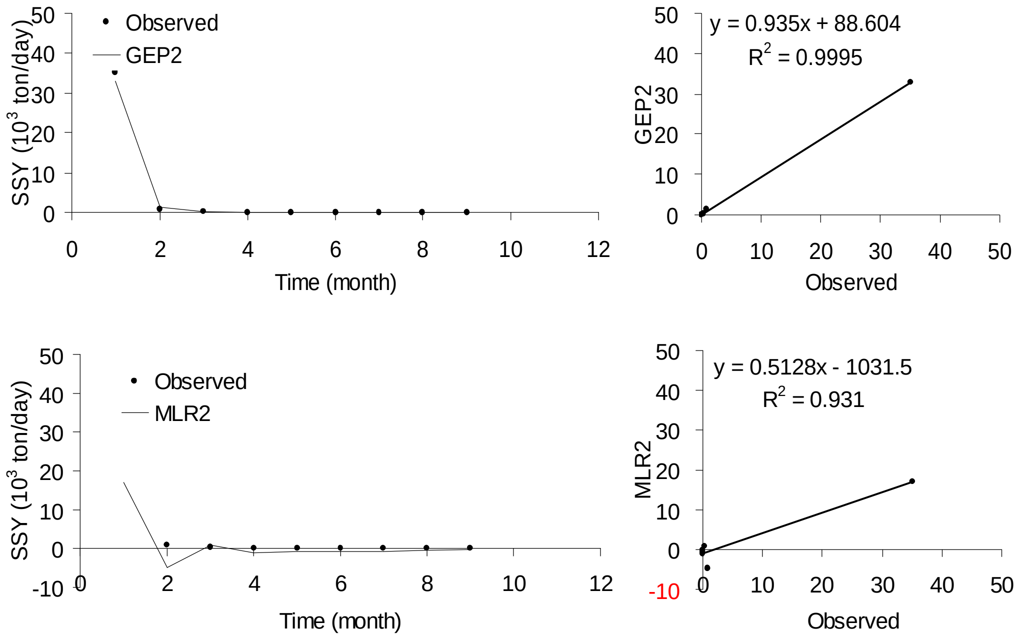


Figure 6.19 The observed and estimated suspended sediments of station 2177 in testing period.

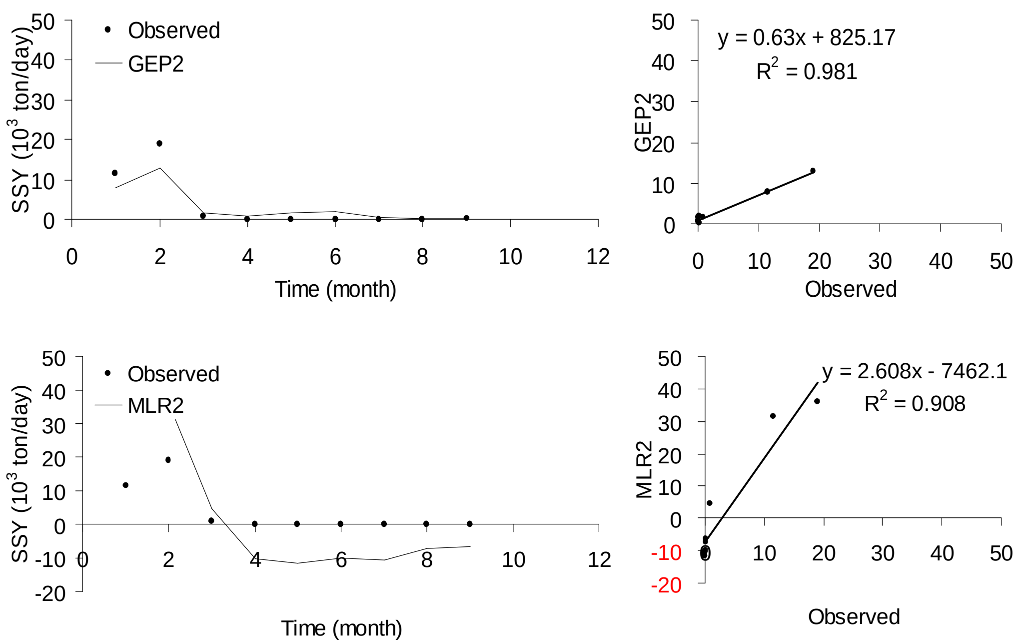


Figure 6.20 The observed and estimated suspended sediments of station 2122 in testing period.

CHAPTER 7

CONCLUSIONS AND FURTHER SUGGESTIONS

7.1 Conclusions

This study indicates the use of gene-expression programming (GEP), which is an extension to genetic programming (GP) technique, to model the discharge (Q)-suspended sediment yield (SSY) relationship. The proposed GEP model is explicit and simple that can be used by anyone not specialized with GP. The proposed models give a practical way for sediment estimation to obtain accurate results and encourage the use of GEP in other aspects of water engineering studies.

The suspended sediment estimates based on GEP models are compared with two different sediment rating curves and multi-linear regression. The results obtained with GEP models are better than those obtained using the conventional rating curve and multi-linear regression techniques and confirm the ability of this technique to provide a useful tool in solving specific problems in hydrology, such as suspended sediment estimation. The results suggest that the GEP approach may provide a superior alternative to the sediment rating curve and multi-linear regression techniques. The MLR method was found to provide negative approximations for some of the low sediment values, whereas the other model estimates are all positive.

In this study, first the relationship between suspended sediment yield (SSY) and current discharge (Q) was modeled for the five stations (given in Eqn.5.1). And as an alternative attempt, the correlation between SSY and time-lagged Q and SSY variants were investigated. The cross-correlation matrix (Table 5.1) showed that SSY can be well estimated using the current and 1 month-lagged discharge values (Q and Q_{t-1}) and the values of 1 month lagged suspended sediment yield (SSY_{t-1}). The results showed that using time-lagged Q and SSY values enhanced the generalization capacity of GEP modeling (Tables 6.1-6.4). The same findings were also recorded while MLR modeling attempts. MLR2 that employs Q, Qt-1 and SSYt-1 as input set performed better than MLR that employs only Q as input. The results are given in Tables 6.1-6.4.

The SRC2 model with a bias term is found to be much better than the unbiased SRC1 model. The study only used data from five stations on Middle Euphrates Basin and further work using more data from various areas (new stations) may be required to strengthen these conclusions.

Although the models successfully simulated measured suspended sediment yield under predominant flow conditions, errors in SSY estimates were larger during high- and low-flow conditions. This shortcoming might be improved if additional concentration data for a wider range of flow conditions were available for calibrating the model.

Filling the intermittent sediment-discharge data available for Middle Euphrates Basin is out of this study, and the gaps were filled with mean of neighbour data points. However, the proposed GEP2 model for five stations can also be used for filling the gaps in available data, providing that discharge values are existed. It is clear that for the GEP models developed based on time-lagged discharge and sediment yield data (Eqns. 5.9-5.13) the required input data other than discharge are the time-lagged suspended sediment values, which are readily calculated from available data. These models can be practically used for gap-filling in the corresponding five stations.

In the present study, gene-expression programming that is a relatively new and powerful variant of genetic programming was used for the explicit formulation of monthly averaged suspended sediment yield–discharge relationship (Eqs. 5.4-5.13). Other soft computing techniques such as Artificial Neural Networks, Fuzzy Logic, Neuro-fuzzy and Support Vector Machines have already been applied in this subject area. It is believed that the results of this study will take an important place and will fill a gap in scientific area of suspended sediment estimation using new robust techniques.

7.2 Further suggestions

Present research investigated the use of GEP in estimation of monthly averaged suspended sediment yield values five sediment measurement stations on Euphrates River in Middle Euphrates Basin. Only five stations may not be representative for the whole basin area, therefore new sediment stations should be established especially near to potential river sections for future hydraulic structures.

During the data collection period, gaps have been observed especially in suspended sediment yield data in Euphrates Basin, which is still the most potential area in Turkey for future water projects. These gaps affect the reliability of estimation methods used by both EİEİ and DSİ. Hence, a comprehensive project should be established in co-operation with EİEİ and DSİ, which would contain robust conceptual and stochastic methods to fill the gaps in past sediment and discharge data in Euphrates Basin. More importantly, the results of the project should be shared with water engineering community via internet media.

Monthly averaged sediment and discharge data may not capture the seasonal variation and also abrupt climatic changes that would cause extreme sediment yield and discharge records. During the literature review on past studies, it was observed that especially U.S. and European countries study also on daily and, more specifically, hourly discharge and sediment yield measurement of their country rivers. EİEİ and DSİ should be advised and motivated for fulfilling new projects to establish more stations and increase the period of getting sample data from flow and sediment measurement stations.

The results of this research, which has been derived based on Middle Euphrates Basin, can be used for the segments of Euphrates River passing from the other parts of Euphrates Basin, which show similar conditions.

To say that the rapid emergence of GIS has revolutionized water engineering is no exaggeration. GIS has proven to be especially useful when looking at *what if* scenarios and for developing complex simulation models. The use of GIS in extracting hydrologic characteristics and also sedimentation history of Euphrates Basin is chiefly recommended.

This research proposed one soft computing model as alternative to conventional regression based methods. A more detailed comparative study may be performed to investigate the capacity and robustness of other branches soft computing technique and also stochastic models.

REFERENCES

1. Annandale, G.W. (2006). *Scour technology*. Mc-Graw Hill, New York.
2. Asselman, N.E.M. (2000). Fitting and interpretation of sediment rating curves. J*ourna of. Hydrology*, **234**, 228–248.
3. Aytek A., and Kisi O. (2008). A genetic programming approach to suspended sediment modeling. *Journal of Hydrology,* **351**, 288-298.
4. Azamathulla, M.D., Ghani A. Ab., Guven, A. and Zakaria, N.A. (2009). Genetic programming to predict bridge pier scour. *Journal of Hydraulic Engineering* (In press).
5. Babovic, V. (2000). Data mining and knowledge discovery in sediment transport. *Comput.-Aided Civ. Infrastruct. Eng*. **15**(5), 383–389.
6. Babovic, V. and Keijzer, M. (2000). Genetic programming as a model induction engine. *Journal of Hydroinformatics*, **1**(2),35-60.
7. Babovic,V., and Keijzer M. (2002). Rainfall-runoff modeling based on genetic programming. **Nordic Hydrology**, **33**(5), 331-346.
8. Babovic, V., Keijzer, M., David Rodrigez Aguilera, and Harrington, J. (2001). Automatic discovery of settling velocity equations. D2K Technical Report 0201-1, Danish Technical Research Council (STVF), http://www.d2k.dk.
9. Bagnold, R.A. (1980). An empirical correlation of bedload transport rates in flumes and natural rivers. In: The Physics of Sediment Transport by Wind and Water. ASCE, pp. 323–345.
10. Bhattacharya, B., Price, R.K., and Solomatine, D.P. (2005). Data-driven modelling in the context of sediment transport. *Physics and Chemistry of the Earth*, **30** (2005) 297–302.
11. Cigizoglu, H.K. (2002a). Suspended sediment estimation and forecasting using articial neural networks. *Turkish J. Eng. Env. Sci*., **26** , 15 -25.
12. Cigizoglu, H. K. (2002b). “Intermitting river flow forecasting by artificial neural networks.” *14th Int. Conf. on Computational Methods in Water Resources*, S. M. Hassanizadeh, R. J. Schotting, W. G. Gray, and G. F. Pinder, eds., Elsevier, Amsterdam, The Netherlands, 1653–1660.

13. Cigizoglu, H.K. (2004). Estimation and forecasting of daily suspended sediment data by multi layer perceptrons. *Advances in Water Resources*, **27**, 185–195.
14. Cigizoglu H K and Alp, M. (2006) Generalized regression neural network in modelling river sediment yield. *Advanced Engineering Software*, **37**, 63–68.
15. Cigizoglu H.K. and Kisi O. (2006). Methods to improve the neural network performance in suspended sediment estimation. *Journal of Hydrology,* **317**(3-4), 221-238.
16. Cousin, N., Savic, D.A. (1997). A rainfall-runoff model using genetic programming. Centre for Systems and Control Engineering, Report No. 97/03, School of Engineering, University of Exeter, Exeter, United Kingdom, p. 70.
17. Crowder, D.W., Demissie, M., Markus, M. (2007). The accuracy of sediment loads when log-transformation produces nonlinear sediment load–discharge relationships. *Journal of Hydrology*, **336**, 250–268.
18. Davidson, J.W., Savic, D.A., and Walters, G.A. (1999). Method for identification of explicit polynomial formulae for the friction in turbulent pipe flow. *Journal of Hydroinformatics*, **1**(2), 115-126.
19. Dorado, J., Rabunal, J.R., Pazos, A., Rivero, D., Santos, A., Puertas, J. (2003). Prediction and modelling of the rainfall-runoff transformation of a typical urban basin using ANN and GP. *Appl. Artif. Intell*. **17**, 329–343.
20. Drecourt J.P. (1999). Application of Neural Networks and Genetic Programming to rainfall-runoff rodeling. Danish Hydraulic Institute (Hydro-Informatics Techonologies HIT), D2K-0699-1.
21. Ferguson, R.I. (1986). River loads underestimated by rating curves. *Water Resour. Res.*, **22**(1), 74–76.
22. Ferreira, C. (2001a). Gene expression programming in problem solving. *Proc., 6th Online World conference on Soft Computing in Industrial Applications* (invited tutorial).
23. Ferreira, C. (2001b). Gene expression programming: A new adaptive algorithm for solving problems. *Complex Systems*, **13**(2), 87–129.

24. Ferreira, C. (2006). *Gene-expression programming; mathematical modeling by an artificial intelligence*. Springer, Berling, Heidelberg, New York.
25. Francke, T., Lopez-Tarazon, J.A. and Schröder, B. (2008). Estimation of suspended sediment concentration and yield using linear models random forests and quantile regression forests. *Hydrological Processes*, **22**, 4892-4904.
26. Giustolisi, O. (2004). Using genetic programming to determine Chezy resistance coefficient in corrugated channels. *Journal of Hydroinform*atics, 157–173.
27. Graf, W.H. (1971). *Hydraulics of Sediment Transport*. McGraw-Hill Book Company.
28. Guven A and Gunal M (2008a). A genetic programming approach for prediction of local scour downstream hydraulic structures. *J. Irrig. and Drain. Eng.* **132**(4), 241-249.
29. Guven, A. and Gunal, M. (2008b). Prediction of local scour downstram of grade-control structıres using neural networks. *J. Hydraulic Engineering*, **134**(11), 1656-1660.
30. Guven A, Aytek A Yuce Mİ and Aksoy H. (2008). Genetic programming-based empirical model for daily reference evapotranspiration estimation; *CLEAN-Soil, Air, Water*, **36**(10-11), 905-912.
31. Guven, A. (2009). Linear genetic programming for time-series modeling of daily flow rate. *Journal of Earth System Science*, **118**(2), 137-146.
32. Guven, A. and Aytek, A. (2009). New approach for stage-discharge relationship: Gene-Expression Programming. *Journal of Hydrologic Engineering*, **14**(8), 812-820
33. Gyr, A. and Hoyer, K. (2006). *Sediment transport: A geophysical phenomenon*. Springer, AA Dordrecht, The Netherlands.
34. Harris, E.L., Babovic, V. and Falconer R.A. (2003). Velocity predictions in compound channels with vegetated floodplains using genetic programming. *Int. Journal of River Basin Management*, **1**(2), 117-123.
35. Herschy, R.W. (1995). *Streamflow measurement*, 2nd Ed., E & FN Spon, London.

36. Kizhisseri, A.S., Simmonds, D., Rafiq, Y., Borthwick, M. (2005). An Evolutionary computation approach to sediment transport modeling. In: Fifth International Conference on Coastal Dynamics, April 4–8, 2005, Barcelona, Spain.
37. Kisi, O. (2004). River flow modelling using artificial neural networks. *Journal of Hydrologic Eng*. **9**(1) 60**–**63.
38. Kisi, O. (2005). Suspended sediment estimation using neuro-fuzzy and neural network approaches. *Hydrological Sciences Journal*, **50**(4), 683–696.
39. Kisi O, Emin Karahan M, Sen Z. (2006). River suspended sediment modelling using a fuzzy logic approach. *Hydrological Processes*, **20**(20), 4351–4362.
40. Koza J.R. (1992). *Genetic programming: On the programming of computers by means of natural selection*. MIT Press, Cambridge, MA.
41. Liu, Z. (2001). *Sediment transport*. Hydraulic Laboratory of Institute of Vand, Alborg University.
42. Lohani AK, Goel K, Bhatia KKS. (2007). Deriving stage–discharge–sediment concentration relationships using fuzzy logic. *Hydrological Sciences Journal*, **52**(4), 793–807.
43. Maier H.R. and Dandy G.C. (2000). Neural networks for the prediction and forecasting of water resources variables: a review of modeling issues and applications. *Environmental Modeling and Software*, **15**(1), 101-123.
44. McBean, E.A., Al-Nassri, S. (1988). Uncertainty in suspended sediment transport curves. *J. Hydr. Eng*. ASCE, **114**(1), 63–74.
45. Oltean, M. and Groşan, C. (2003). A comparison of several linear genetic programming techniques. *Complex Systems*, **14**(1) 1**–**29.
46. Overleir, A.P. (2004). Accounting for heteroscedasticity in rating curve estimates. *J. Hydrol*., **292**, 173–181.
47. Ozturk, F. (1996). Suspended sediment yield of rivers in Turkey. Erosion and Sediment Yield: Global and Regional Perspectives. In: International Association Hydrological Sciences (Eds. D.E. Walling and B.W. Webb). Great Yarmouth, UK.

48. Ozturk, F., Yurekli, K. and Apaydın, H. (2008). Stochastic Modeling of Suspended Sediment from Yesilirmak Basin, Turkey. *International Journal of Natural and Engineering Sciences*, **2**(1), 21-27.
49. Picoet, C., Hingray, B., Olivry, J.C. (2001). Empirical and conceptual modeling of the suspended sediment dynamics in large tropical African River: The Upper Niger River Basin. *J. Hydrol.*, **250**, 19–39.
50. Rabunal, J.R., Puertas, J., Suarez, J., Rivero, D. (2007). Determination of the unit hydrograph of a typical urban basin using genetic programming and artificial neural networks. *Hydrol. Process.*, **21**, 476–485.
51. Savic A.D., Walters A.G., and Davidson J.W. (1999). A genetic programming approach to rainfall-runoff modeling. *Water Resources Management*, **13**, 219-231.
52. Singh, K.P. and Dursunoglu, A. (1989). Developing accurate and reliable stream sediment yields. Sediment and the Environment (Proceedings of the Baltimore Symposium, May 1), IAHS Publ. no. 184, pp 193-199.
53. Singh A. K, Deo, M. C. and Sanil Kumar, V. (2007) Neural network – genetic programming for sediment transport; *ICE Journal of Maritime Engineering* **160**(MA3) 113–119.
54. Tayfur, G. and Guldal, V. (2006) Artificial neural networks for estimating daily total suspended sediment in natural streams. *Nordic Hydrology*, **37**(2006), 69-79.
55. Thomas, R.B. (1985). Estimating total suspended sediment yield with probability sampling. *Water Resour. Res.*, **21**, 1381–1388.
56. Walling, D. E. (1977). Limitation of the rating curve technique for estimating suspended sediment loads, with particular reference to British rivers. In: Erosion and Solid Matter Transport in Inland Waters (Proc. Paris Symp., pp. 34-48. IAHS Publ. no. 122.
57. Williams, G.P. (1989). Sediment concentration versus water discharge during single hydrologic events in rivers. *Journal of Hydrology*, **111**(1989), 89-106.
58. Weisberg, S. (2005). *Applied Linear Regression*. 3rd Ed., John Wiley & Sons, Inc., Hoboken, NJ.
59. Whigham, P. A., and Crapper, P. F. (1999). Time series modeling using genetic programming: An application to rainfall-runoff models. *Advances*

in genetic programming, L. Spector, W. B. Langdon, Una-May O'Reilly and P.J. Angeline, eds., The MIT Pres, Cambridge, Mass., pp.89–104.

60. Whigham, P. A., and Crapper, P. F. (2001). Modeling rainfall-runoff using genetic programming. *Math. Comput. Model.*, **33**(6-7), 707–721.
61. Zadeh, L.A. (1991). Fuzzy Logic and the Calculus of Fuzzy If-Then Rules, *OMRON Technics*, **31**(4), 316- 320.
62. Zyserman, J.A., Fredsøe, J. (1994). Data analysis of bed concentration of suspended sediment. *J. Hydrol. Eng.* ASCE, **120**(9), 1021–1042.

APPENDIX

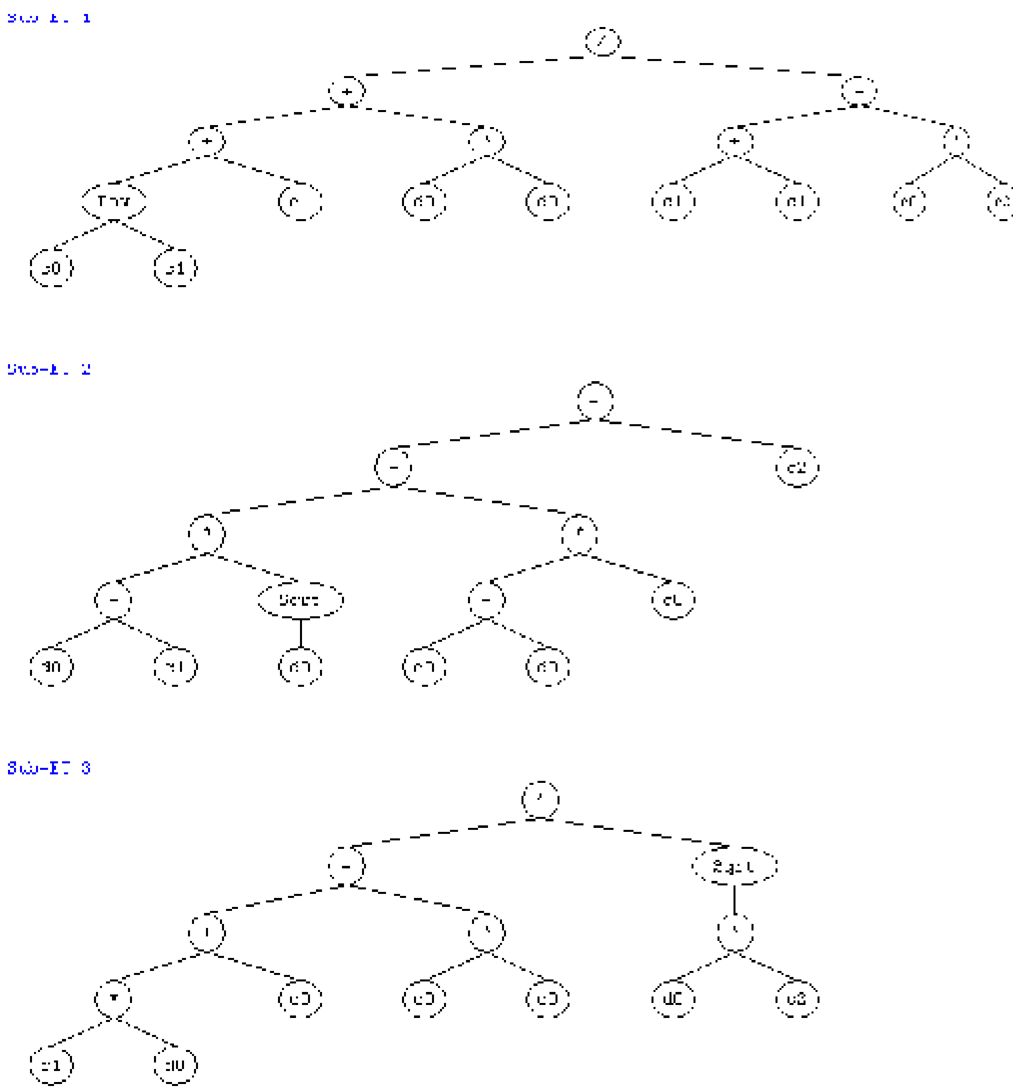


Figure 5.1 Expression tree (ET) representation of GEP1 model for station 2102

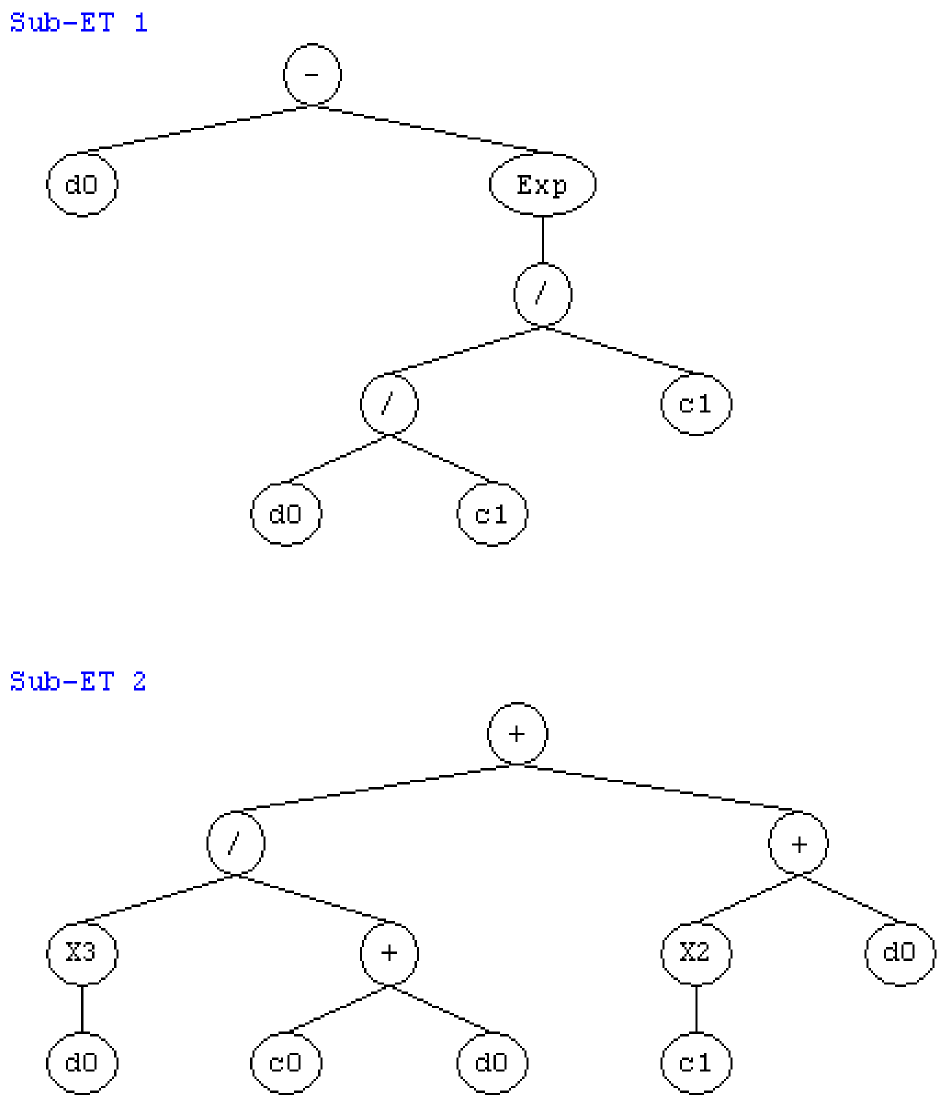


Figure 5.2 Expression tree (ET) representation of GEP1 model for station 2164

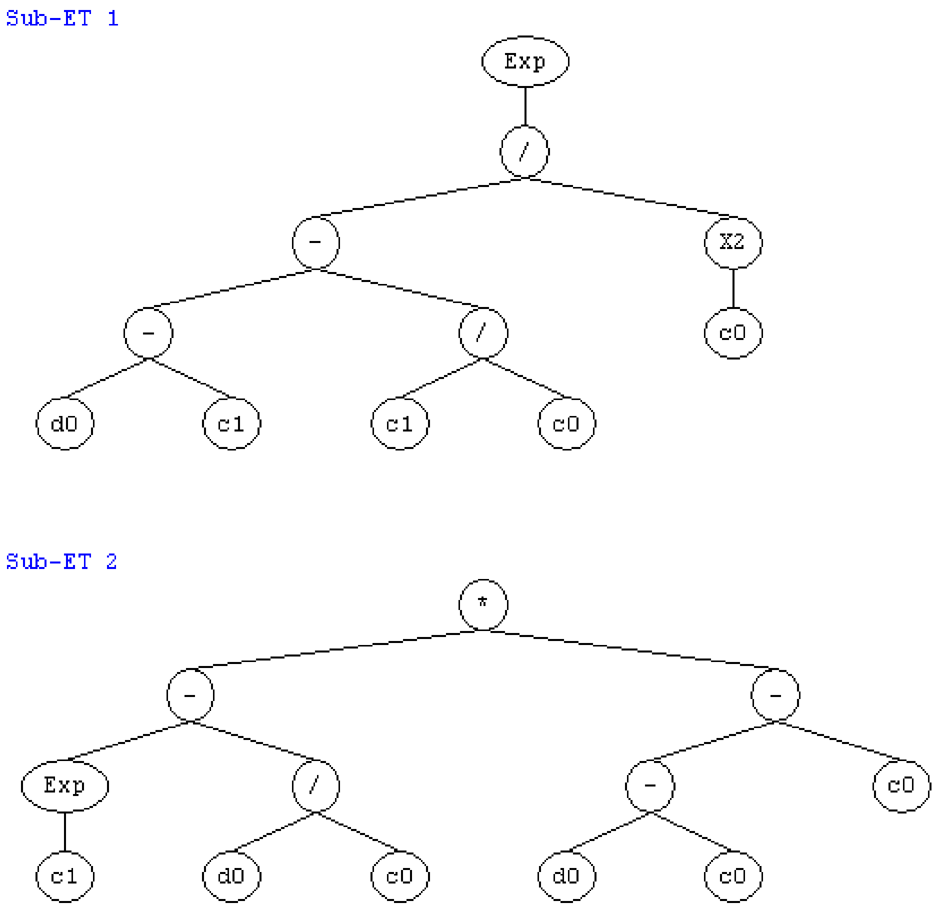


Figure 5.3 Expression tree (ET) representation of GEP1 model for station 2174

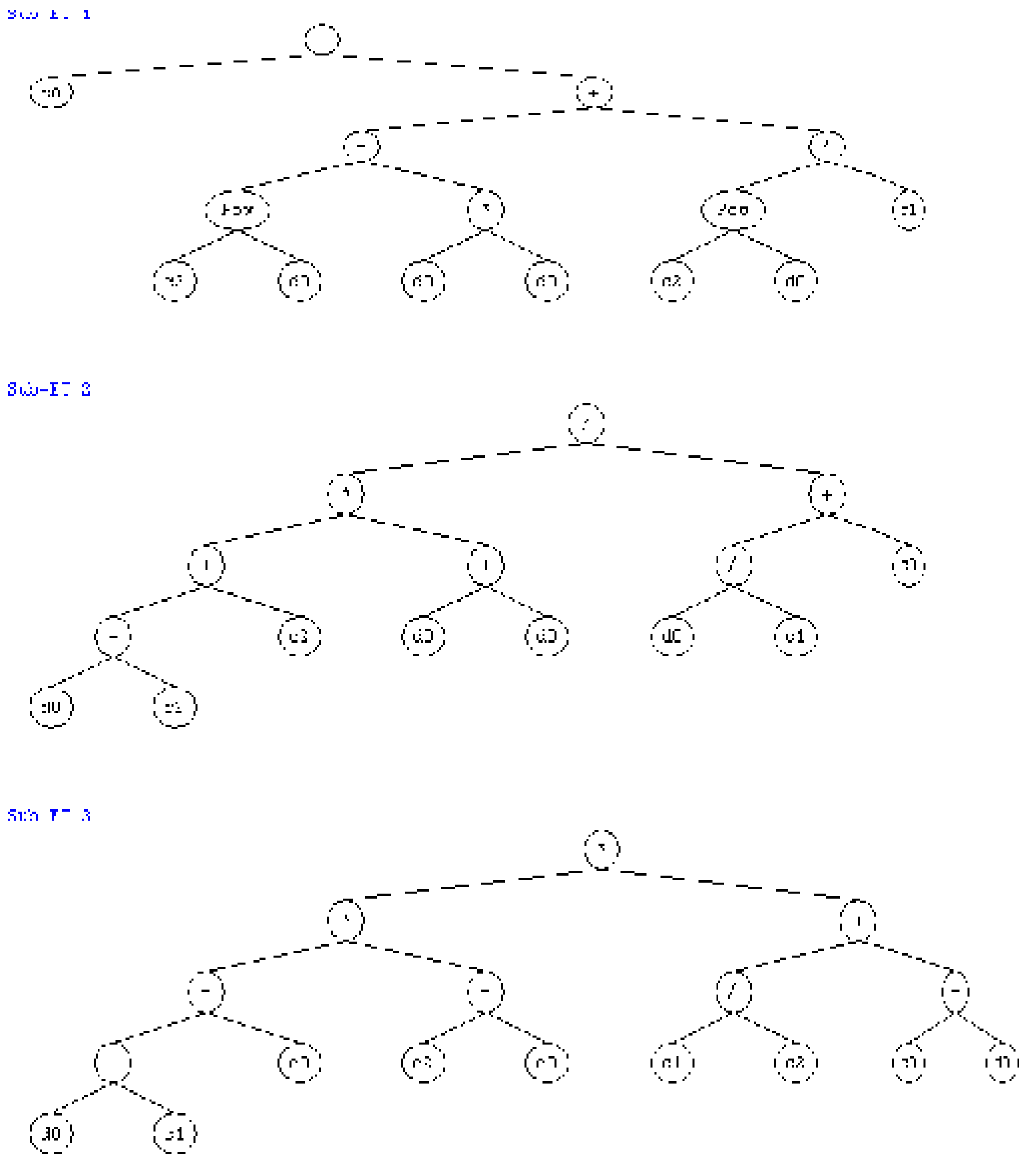


Figure 5.4 Expression tree (ET) representation of GEP1 model for station 2177

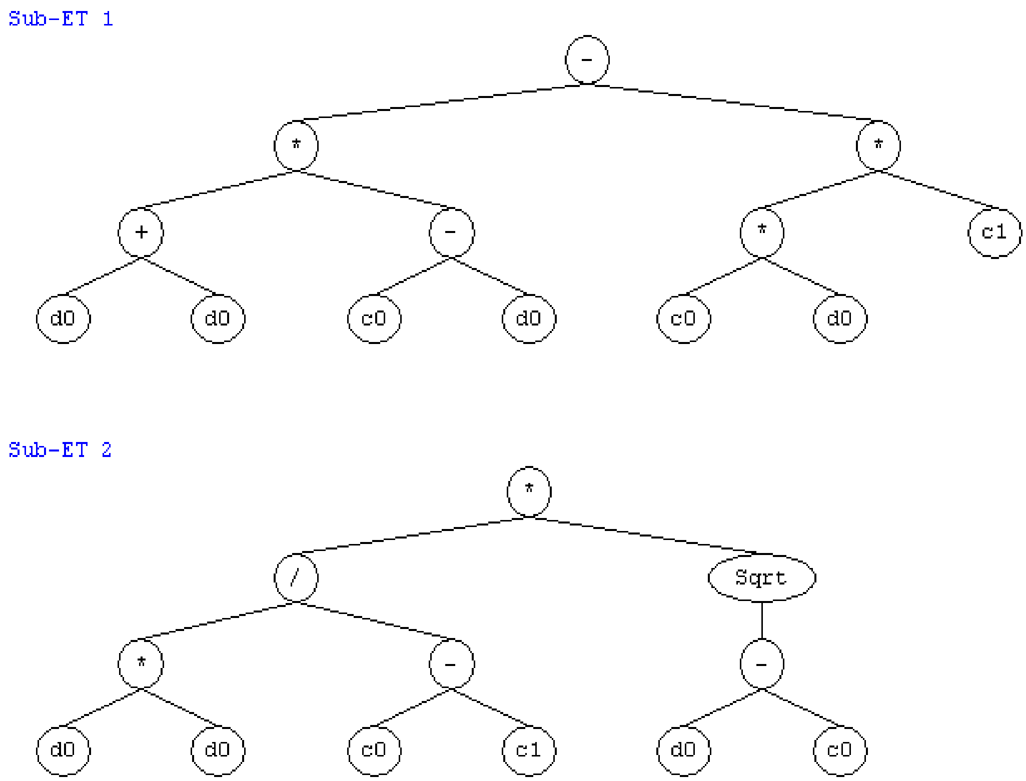


Figure 5.5 Expression tree (ET) representation of GEP1 model for station 2122

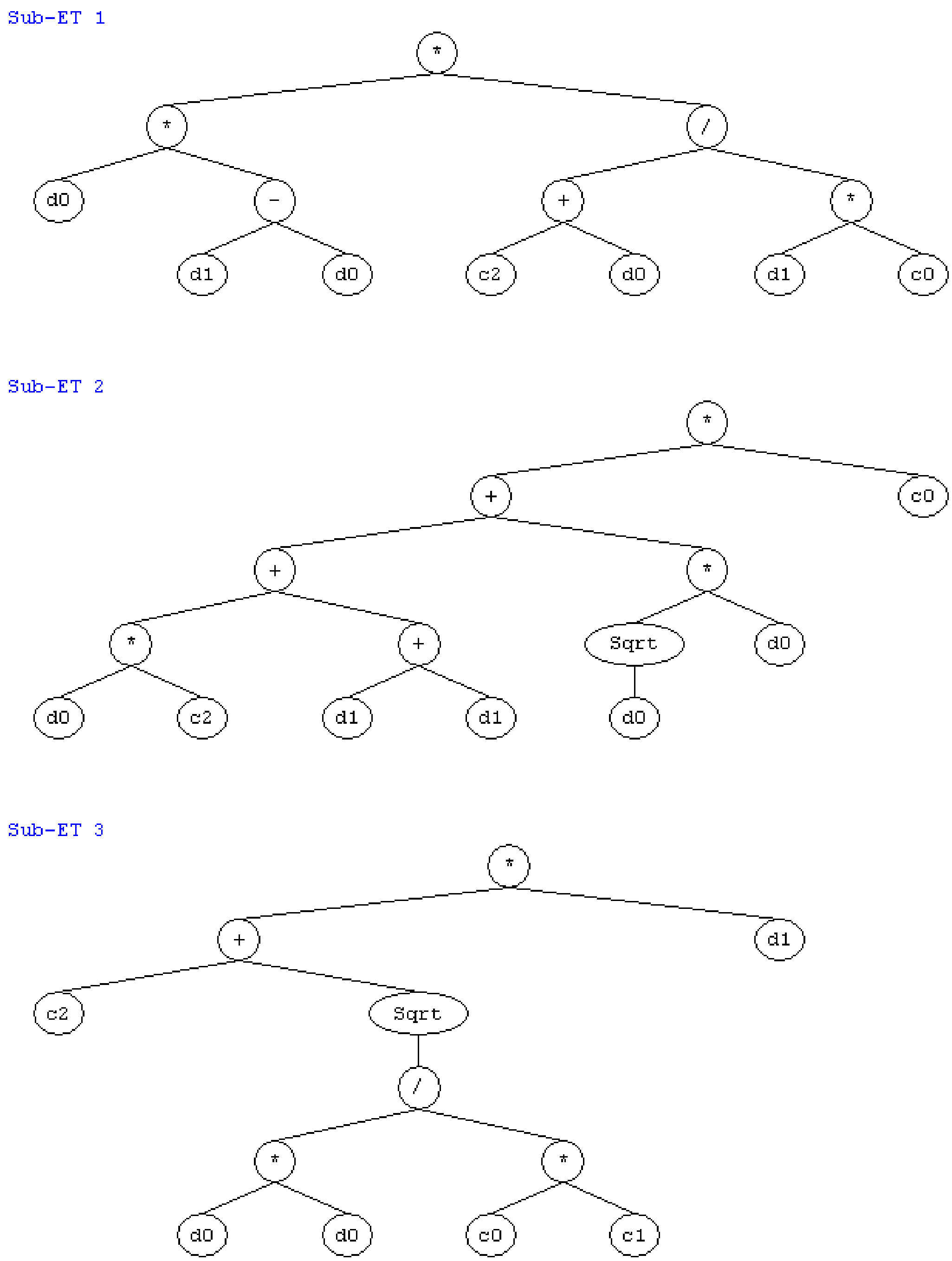


Figure 5.6 Expression tree (ET) representation of GEP2 model for station 2102

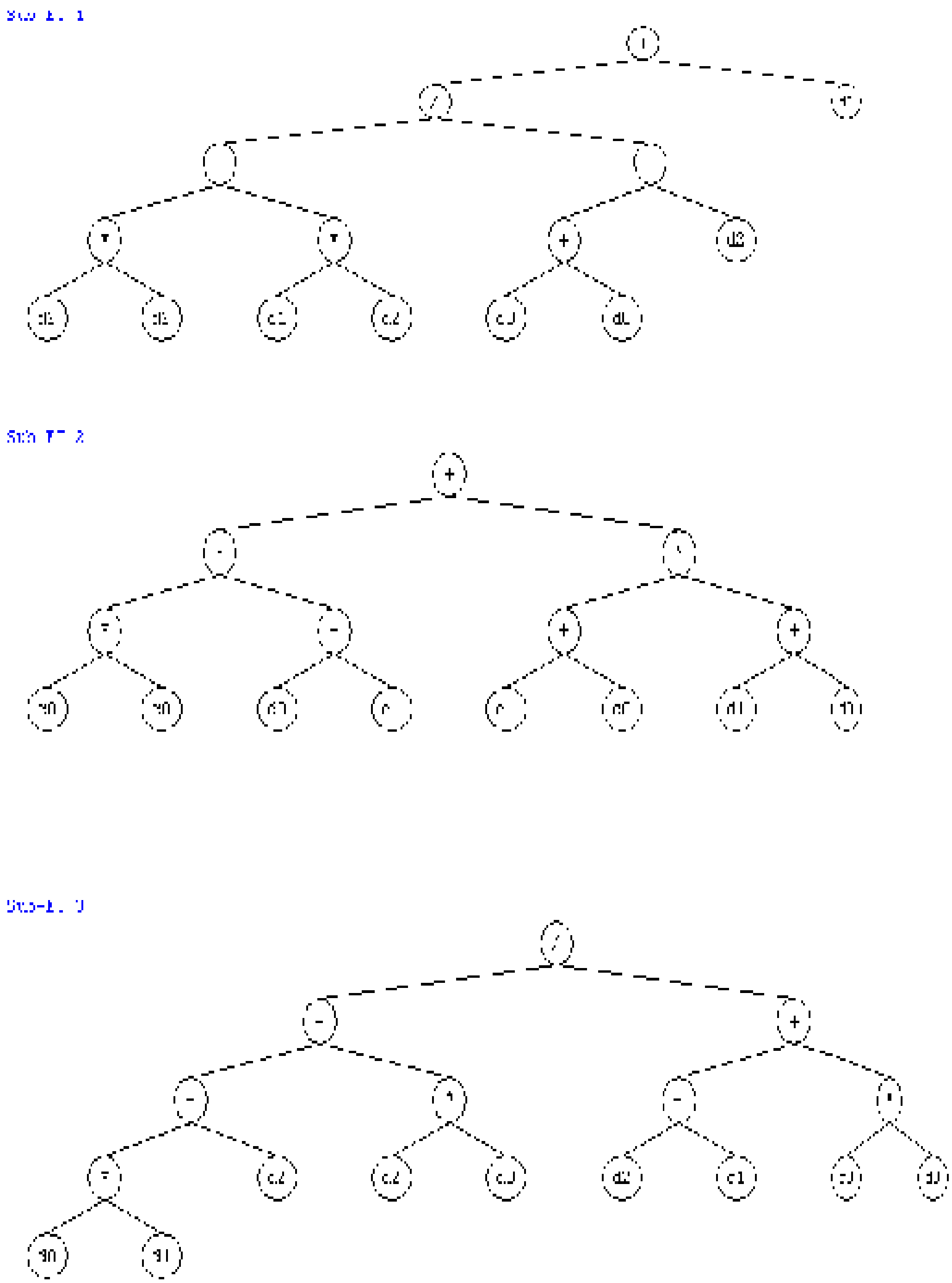


Figure 5.7 Expression tree (ET) representation of GEP2 model for station 2164

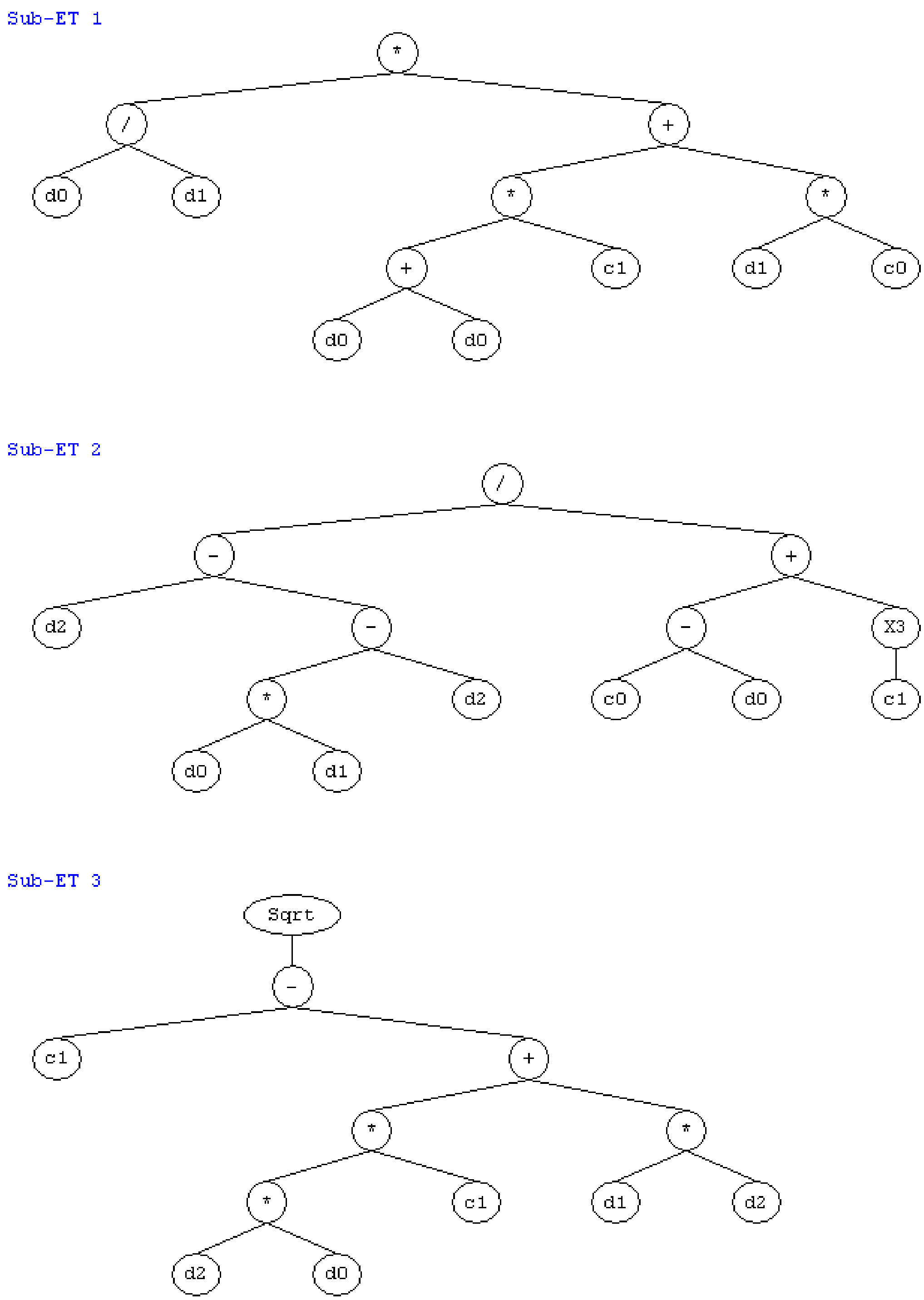


Figure 5.8 Expression tree (ET) representation of GEP2 model for station 2174

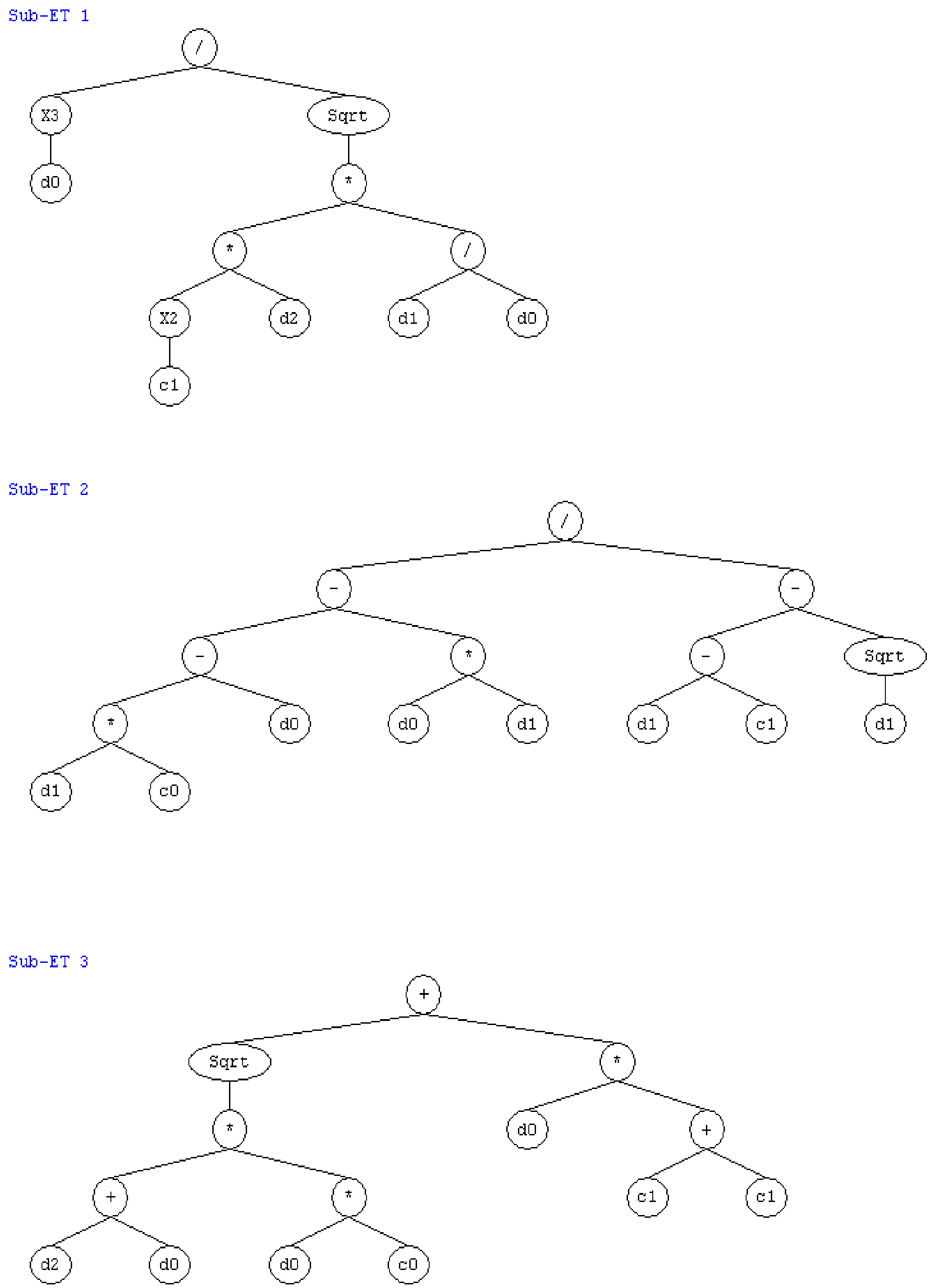
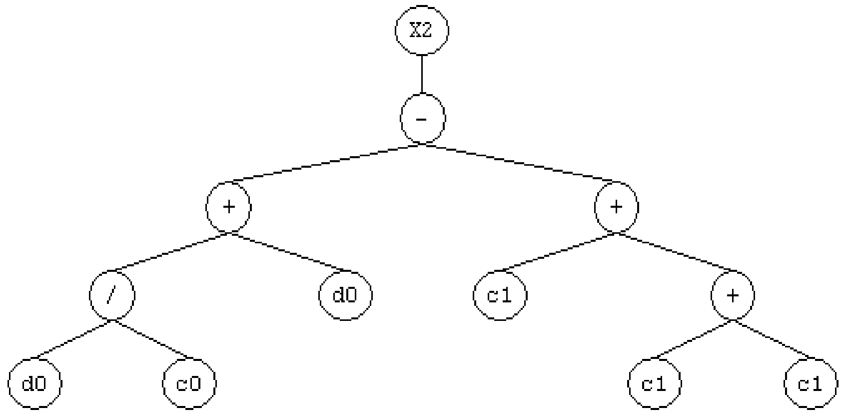
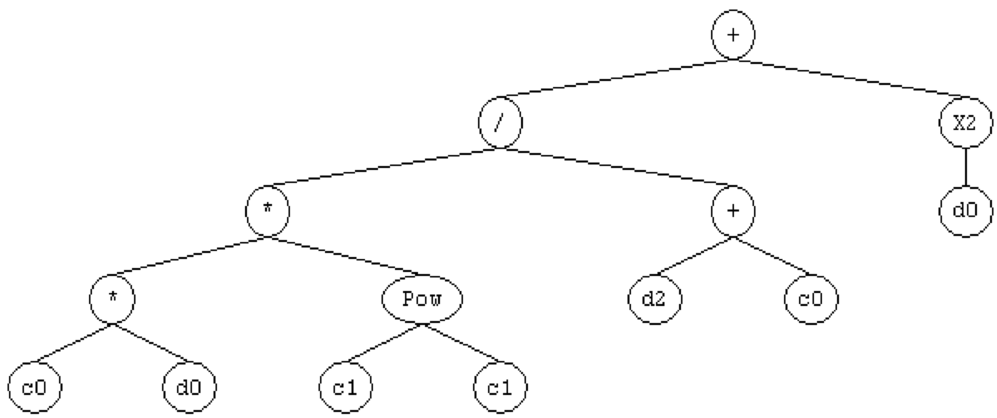


Figure 5.9 Expression tree (ET) representation of GEP2 model for station 2177

Sub-ET 1



Sub-ET 2



Sub-ET 3

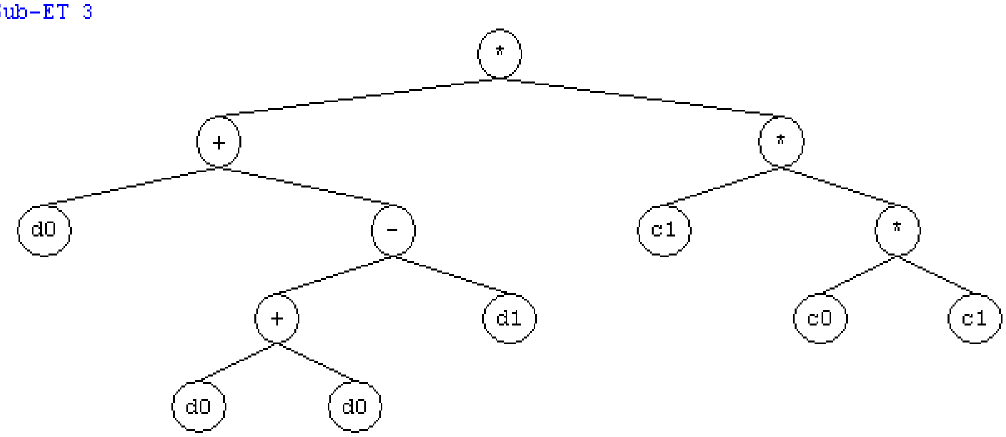


Figure 5.10 Expression tree (ET) representation of GEP2 model for station 2122