ESTIMATION OF SUSPENDED SEDIMENT YIELD FLOWING INTO INANDA DAM USING GENETIC PROGRAMMING

Submitted in fulfilment of the requirements

of the degree of

Master of Engineering

in the

Faculty of Engineering and the Built Environment at

Durban University of Technology

Adesoji Tunbosun JAIYEOLA

Approved for final submission:

Supervisor……..... Date……

Professor Josiah Adeyemo

Co-supervisor…… Date……

Professor Fred Otieno

December, 2015

ABSTRACT

Reservoirs are designed to specific volume called the dead storage to be able to withstand the quantity of particles in the rivers flowing into it during its design period called its economic life. Therefore, accurate calculation of the quantities of sediment being transported is of great significance in environment engineering, hydroelectric equipment longevity, river aesthetics, pollution and channel navigability. In this study different input combination of monthly upstream suspended sediment concentration and upstream flow dataset for Inanda Dam for 15 years was used to develop a model for each month of the year. The predictive abilities of each of the developed model to predict the quantity of suspended sediment flowing into Inanda Dam were also compared with those of the corresponding developed Sediment Rating Curves using two evaluation criteria - Determination of Coefficient (R^2) and Root-Mean-Square Error (RMSE). The results from this study show that a genetic programming approach can be used to accurately predict the relationship between the streamflow and the suspended sediment load flowing into Inanda Dam. The twelve developed monthly genetic programming (GP) models produced a significantly low difference when the observed suspended sediment load was compared with the predicted suspended sediment load. The average R^2 values and RMS error for the twelve developed models were 0.9996 and 0.3566 respectively during the validation phase. The Genetic Programming models were also able to replicate extreme hydrological events like predicting low and high suspended sediment load flowing into the dam. Moreover, the study also produced accurate sediment rating curve models with low RMSE values of between 0.3971 and 11.8852 and high R^2 values of between 0.9833 and 0.9962. This shows that sediment rating curves can be used to predict historical missing data of the quantity of suspended sediment flowing into Inanda Dam using existing streamflow datasets. The results from this study further show that the predictions from the Genetic Programming models are better than the predictions from the Sediment Raring Curve models, especially in predicting large quantities of suspended sediment load during high streamflow such as during flood events. This proves that Genetic Programming

technique is a better predictive tool than Sediment Raring Curve technique. In conclusion, the results from this study are very promising and support the use of Genetic Programming in predicting the nonlinear and complex relationship between suspended sediment load and streamflow at the inlet of Inanda Dam in KwaZulu-Natal. This will help planners and managers of the dam to understand the system better in terms of its problems and to find alternative ways to address them.

DECLARATION

I hereby declare that the work described in this thesis is my original work and has not previously been submitted in its entirety or in part for a degree in any other university. I further declare that this work does not infringe or violate the right of others, as all the sources cited or quoted are indicated and acknowledged by means of a comprehensive list of references.

Adesoji Tunbosun Jaiyeola.

DEDICATION

To The Almighty God

The King of Kings

The Ancient of Days

The Greatest of all Teachers

ACKNOWLEDGEMENTS

I give GOD all the glory for the successful completion of this study. He remains faithful even when I am not faithful. I appreciate HIM for his love and mercy throughout the duration of this study. May his name be praised for ever and ever. I thank him for every new day He brings my way.

I would like to say a big THANK YOU to my supervisor, Professor Josiah Adeyemo, for his professional excellence which contributed a great deal to the success of this study. His pieces of advice, contributions and attention to details played a major role in propelling me to overcome all the challenges I encountered in this study. I would also like to appreciate my co-supervisor Professor Fred Otieno for his fatherly role, guardiance and positive contributions during difficult times in this study.

I would like to express my heart felt gratitude to all those who contributed one way or the other, in the planning, development, production and presentation of this study. I pray that God will reward you for all your effort in making the study a success.

A big THANK YOU also goes to all friends, brothers and colleagues like Mr Oluwaseun Kunle Oyebode, Dr Oluwatosin Olofintoye, Mr Onyeka Nkwonta, Mr Joseph Kapuku Bwapwa and those I cannot mention their names because of space. They really helped me in transiting from construction work to research work.

I also acknowledge the support of the management and staff of the South African Weather Services, Umgeni waters and the Department of Water Affairs, for providing the necessary data for carrying out this study.

I want to thank Durban University of Technology for the opportunity given to me to pursue my dream and also thank Mangosuthu University of Technology for their support during this study.

Finally, I would like to appreciate the support from my family, especially from my lovely and wonderful wife Mrs Abosede Jaiyeola, and my children, Oladapo, Temiloluwa, Ebunoluwa and Oluwafemi, for their tolerance and moral support

during the course of this study. You guys are my sunshine! To my beloved mother Mrs Victoria Jaiyeola and late father Master Warrant Officer Joseph Ayoola Jaiyeola, I say a big THANK YOU for counting it worthy to send me to school.

TABLE OF CONTENTS

ABSTRACT i
DECLARATION iii
DEDICATION iv
ACKNOWLEDGEMENTS v
TABLE OF CONTENTS vii
LIST OF FIGURES xi
LIST OF TABLES xv
SYMBOLS AND ABBREVIATIONS xvi
CHAPTER 1 1
GENERAL INTRODUCTION 1
1.1 INTRODUCTION 1
1.2 PROBLEM STATEMENT 2
1.3 RESEARCH OBJECTIVE 3
1.4 RESEARCH METHODOLOGY 3
1.4.1 Genetic programming 4
1.4.2 The sediment rating curve 5
1.5 DESCRIPTION OF THE STUDY AREA - INANDA DAM 6
1.6 RATIONALE OF THE STUDY 9
1.7 SIGNIFICANCE OF THE STUDY 10
1.8 OUTLINE OF THE THESIS 11
1.9 PUBLICATIONS 12
CHAPTER 2 14
LITERATURE REVIEW 14
2.1 PREDICTIVE MODELLING TECHNIQUES IN SEDIMENT YIELD AND HYDROLOGICAL MODELLING 14
2.2 TYPES OF HYDROLOGICAL MODELS 16
2.2.1 Empirical models 16
2.2.2 Conceptual models 17
2.2.3 Physical based models 17

2.2.4 Selecting an appropriate model 18
2.3 ARTIFICIAL INTELLIGENCE IN HYDROLOGICAL STUDIES 18
2.3.1 Artificial intelligence data driven models 20
2.4 GENERAL APPROACH FOR APPLICATION OF ARTIFICIALLY INTELLIGENT DATA DRIVEN MODELS (AIDDM) 21
2.5 K-NEAREST NEIGHBOUR (KNN) METHOD BASED MODELS 22
2.6 CHAOS THEORY (CT) BASED MODELS 23
2.7 ARTIFICIAL NEURAL NETWORKS (ANN) BASED MODELS 26
2.7.1 ANN model development 26
2.8 FUZZY RULE BASED SYSTEMS (FRBS) 29
2.8.1 FRBS model evaluation and applications 30
2.9 SUPPORT VECTOR MACHINES (SVM) BASED MODELS 31
2.10 GENETIC PROGRAMMING (GP) MODEL DEVELOPMENT 32
2.10.1 GP model evaluation and applications 33
2.11 COMPARISON OF ARTIFICIAL INTELLIGENCE (AI) MODELS 34
2.12 RECENT APPLICATIONS OF ARTIFICIAL INTELLIGENCE IN WATER RESOURCES MODELING 35
2.13 RECENT APPROACHES AND FUTURE DIRECTIONS 36
2.14 SOFT COMPUTING AND ENGINEERING MODELLING 37
2.14.1 History of evolutionary algorithms 38
2.15 OVERVIEW OF GENETIC PROGRAMMING (GP) 39
2.15.1 Introduction 39
2.15.2 Initialization of genetic programming 40
2.15.3 Fitness evaluation and selection 41
2.15.4 Genetic operators 42
2.15.5 Variants of genetic programming 44
2.16 GENERAL APPLICATIONS OF GENETIC PROGRAMMING 45
2.16.1 Sediment modelling using GP approach. 46
2.17 APPLICATIONS OF GENETIC PROGRAMMING IN OTHER ENGINEERING APPLICATIONS 48
2.17.1 Genetic programming in photogrammetry 49
2.17.2 Genetic programming in medicine, biology and bioinformatics 49

2.18 EVALUATIONS OF GP THEORIES AND PRINCIPLES 50
2.18.1 Benchmarking of GP 50
2.18.2 Open issues in genetic programming 51
2.18.3 Conclusions – Advancement of GP 52
2.19 SEDIMENT RATING CURVE 52
2.19.1 Representation of the sediment rating curve 53
2.19.2 Application of sediment rating curve in sediment yield modeling 53
2.19.3 Areas of concern 54
2.19.4 Performance improvement 54
2.19.5 Advantages and disadvantages 55
2.20 CONCLUSION 55
CHAPTER 3 56
GENETIC PROGRAMMING FOR PREDICTING SUSPENDED SEDIMENT 56
3.1 INTRODUCTION 56
3.2 METHODOLOGY 57
3.2.1 GPdotNET version 4.0 software 57
3.3 STUDY AREA AND DATASETS 58
3.4 MODEL DEVELOPMENT 59
3.4.1 Input variable section 59
3.4.2 Data splitting 62
3.4.3 Fitness Function and Model Fitness 62
3.4.4 GP algorithm setup 63
3.4.5 Evaluation of the performance of the models 64
3.5 RESULTS AND DISCUSSION 65
3.6 CONCLUSION 92
CHAPTER 4 94
SEDIMENT RATING CURVE FOR SUSPENDED SEDIMENT PREDICTION AND PERFORMANCE COMPARISON BETWEEN GP AND SRC 94
4.1 INTRODUCTION 94
4.2 METHODOLOGY 95
4.2.1 Sediment rating curve (SRC) 95
4.3 IMPLICATIONS AND APPLICABILITY 96

4.3.1 Interpretation of sediment rating 103
4.4 RESULTS AND DISCUSSION 104
4.5 CONCLUSION 125
CHAPTER 5 126
CONCLUSIONS AND RECOMMENDATIONS 126
5.1 GENERAL CONCLUSIONS 126
5.2 RECOMMENDATIONS FOR FUTURE RESEARCH 128
REFERENCES 130

LIST OF FIGURES

Figure 1: The Umgeni river system 6
Figure 2: Inanda Dam catchment and its soil depth 8
Figure 3: Inlet of Inanda Dam 8
Figure 4: Purpose and significance of Inanda Dam 9
Figure 5: Layers and movement of particles in a river 14
Figure 6: GP syntax tree 40
Figure 7: Illustration of the crossover process in GP 43
Figure 8: Illustration of the mutation process in GP 44
Figure 9: Gauging stations in Inanda catchment area 58
Figure 10: Plot of average monthly streamflow dataset 60
Figure 11: Plot of average monthly suspended sediment dataset 60
Figure 12: Plots of observed and GP-predicted suspended sediment (mg/l) for January during training phase 68
Figure 13: Plots of observed and GP-predicted suspended sediment (mg/l) for January during validation phase 69
Figure 14: Plots of observed and GP-predicted suspended sediment (mg/l) for February during training phase 70
Figure 15: Plots of observed and GP-predicted suspended sediment (mg/l) for February during validation phase 71
Figure 16: Plots of observed and GP-predicted suspended sediment (mg/l) for March during training phase 72
Figure 17: Plots of observed and GP-predicted suspended sediment (mg/l) for March during validation phase 73
Figure 18: Plots of observed and GP-predicted suspended sediment (mg/l) for April during training phase 74
Figure 19: Plots of observed and GP-predicted suspended sediment (mg/l) for April during validation phase 75
Figure 20: Plots of observed and GP-predicted suspended sediment (mg/l) for May during training phase 76

Figure 21: Plots of observed and GP-predicted suspended sediment (mg/l) for May during validation phase 77
Figure 22: Plots of observed and GP-predicted suspended sediment (mg/l) for June during training phase 78
Figure 23: Plots of observed and GP-predicted suspended sediment (mg/l) for June during validation phase 79
Figure 24: Plots of observed and GP-predicted suspended sediment (mg/l) for July during training phase 80
Figure 25: Plots of observed and GP-predicted suspended sediment (mg/l) for July during validation phase 81
Figure 26: Plots of observed and GP-predicted suspended sediment (mg/l) for August during training phase 82
Figure 27: Plots of observed and GP-predicted suspended sediment (mg/l) for August during validation phase 83
Figure 28: Plots of observed and GP-predicted suspended sediment (mg/l) for September during training phase 84
Figure 29: Plots of observed and GP-predicted suspended sediment (mg/l) for September validation phase 85
Figure 30: Plots of observed and GP-predicted suspended sediment (mg/l) for October during training phase 86
Figure 31: Plots of observed and GP-predicted suspended sediment (mg/l) for October during validation phase 87
Figure 32: Plots of observed and GP-predicted suspended sediment (mg/l) for November during training phase 88
Figure 33: Plots of observed and GP-predicted suspended sediment (mg/l) for November during validation phase 89
Figure 34: Plots of observed and GP-predicted suspended sediment (mg/l) for December during training phase 90
Figure 35: Plots of observed and GP-predicted suspended sediment (mg/l) for December during validation phase 91
Figure 36: Sediment rating curves at Inanda Dam for January 97
Figure 37: Sediment rating curves at Inanda Dam for February 98

Figure 38: Sediment rating curves at Inanda Dam for March 98
Figure 39: Sediment rating curves at Inanda Dam for April 99
Figure 40: Sediment rating curves at Inanda Dam for May 99
Figure 41: Sediment rating curves at Inanda Dam for June 100
Figure 42: Sediment rating curves at Inanda Dam for July 100
Figure 43: Sediment rating curves at Inanda Dam for August 101
Figure 44: Sediment rating curves at Inanda Dam for September 101
Figure 45: Sediment rating curves at Inanda Dam for October 102
Figure 46: Sediment rating curves at Inanda Dam for November 102
Figure 47: Sediment rating curves at Inanda Dam for December 103
Figure 48: Comparison of measured, SRC models and GP models results in the validation phase for January 107
Figure 49: Comparison of measured, SRC models and GP models results in the validation phase for February 107
Figure 50: Comparison of measured, SRC models and GP models results in the validation phase for March 108
Figure 51: Comparison of measured, SRC models and GP models results in the validation phase for April 108
Figure 52: Comparison of measured, SRC models and GP models results in the validation phase for May 109
Figure 53: Comparison of measured, SRC models and GP models results in the validation phase for June 109
Figure 54: Comparison of measured, SRC models and GP models results in the validation phase for July 110
Figure 55 Comparison of measured, SRC models and GP models results in the validation phase for August 110
Figure 56: Comparison of measured, SRC models and GP models results in the validation phase for September 111
Figure 57: Comparison of measured, SRC models and GP models results in the validation phase for October 111
Figure 58 Comparison of measured, SRC models and GP models results in the validation phase for November 112

Figure 59: Comparison of measured, SRC models and GP models results in the validation phase for December 112
Figure 60: Observed and predicted suspended sediment load by the GP and SRC models in the validation phase for January 113
Figure 61: Observed and predicted suspended sediment load by the GP and SRC models in the validation phase for February 114
Figure 62: Observed and predicted suspended sediment load by the GP and SRC models in the validation phase for March 115
Figure 63: Observed and predicted suspended sediment load by the GP and SRC models in the validation phase for April 116
Figure 64: Observed and predicted suspended sediment load by the GP and SRC models in the validation phase for May 117
Figure 65: Observed and predicted suspended sediment load by the GP and SRC models in the validation phase for June 118
Figure 66: Observed and predicted suspended sediment load by the GP and SRC models in the validation phase for July 119
Figure 67: Observed and predicted suspended sediment load by the GP and SRC models in the validation phase for August 120
Figure 68: Observed and predicted suspended sediment load by the GP and SRC models in the validation phase for September 121
Figure 69: Observed and predicted suspended sediment load by the GP and SRC models in the validation phase for October 122
Figure 70: Observed and predicted suspended sediment load by the GP and SRC models in the validation phase for November 123
Figure 71: Observed and predicted suspended sediment load by the GP and SRC models in the validation phase for December 124

LIST OF TABLES

Table 1: GP arithmetic operators and mathematical functions 5
Table 2: Inanda Dam statistics (DWAF 2008) 7
Table 3: Examples of functions and terminal sets used in a GP programme 41
Table 4: Results of correlation analysis between input variables and suspended sediment 61
Table 5: GP parameters and settings used in GPdotNET 64
Table 6: Statistical performance of the best individual monthly models 66
Table 7: The form of the equations and the best fit of the monthly rating relationships developed for Inanda Dam 97
Table 8: RMSE and R^2 values for both GP and SRC 105

SYMBOLS AND ABBREVIATIONS

AGNPS	Agricultural Nonpoint Source Pollution
AI	Artificial intelligence
AIDDM	Artificial intelligence data driven models
ANFIS	Adaptive Neuro-fuzzy Inference System
ANN	Artificial neural network
AVG	Average
CF	Correction factor
CGP	Cartesian Genetic Programming
DDDAS	Dynamic Data Driven Application Systems
DDM	Data Driven Models
DSP	Digital signal processing
DSS	Decision support system
DWAF	Department of Water Affairs
Eqn	Equation
EKF	Extended Kalman Filter
EP	Evolution Programming
ERS	European Remote Sensing
FL	Fuzzy Logic
FRBS	Fuzzy Rule Based Systems
GA	Genetic Algorithms
GE	Grammatical Evolution
GENIE	GENetic Imagery Exploitation
GEP	Gene Expression Programming
GIS	Geographical Information System
GP	Genetic Programming
GSSHA	Gridded Surface Subsurface Hydrologic Analysis
I	Counter
IFR	Interactive fuzzy resolution
ITSFP	Interactive two-stage stochastic fuzzy programming

ITSP	Inexact two-stage stochastic programming
K	Number of neighbours
KINEROS2	Kinematic Runoff and Erosion Model
km	Kilometre
KNN	K-nearest Neighbour
LGP	linear Genetic Programming
LISP	list processing
LM	Levenberg-Marquardt
m	Metres
m^3/s	Cubic metre per second
mg/l	Milligrams per litres
MLP	Multilayer Perceptron
MMF	Morgan-Morgan-Finney
MSE	Mean square error
MSRVR	Multi-scale relevance vector regression
MT	Model trees
n	Number of data point
NMR	Nuclear magnetic resonance
NSE	Nash-Sutcliffe efficiency
PCA	Principal Component Analysis
P	Precipitation
Q	Stream flow
Qs	Suspended sediment discharge
Q_t	Streamflow for a given month in the t^{th} year
Q_{t-1}	Streamflow for the same month in the $(t-1)^{th}$ year
Q_{t-2}	Streamflow for the same month in the $(t-2)^{th}$ year
R	Rainfall
R^2	Coefficient of Determination
RMSE	Root Mean Square Error
RUSLE	Revised Universal Soil Loss Equation
RVR	Relevance vector regression
s^2	Mean square error of the log-transformed regression

SAR	Synthetic Aperture Radar
S_m	Predicted suspended sediment
S_O	Observed suspended sediment
$\overline{S}_m$	Mean predicted suspended sediment
$\overline{S}_O$	Mean observed suspended sediment
SI	Scatter index
SRC	Sediment Rating Curve
SRM	Structural risk minimization
SS	Suspended sediment
SS_{t-1}	Suspended sediment for the same month in the (t-1)th year
SS_{t-2}	Suspended sediment for the same month in the (t-2)th year
SSA	Single Spectrum Analysis
SSC	Suspended sediment concentration
SSL	Suspended sediment load
SVM	Support Vector Machines
SWAT	Soil and Water Assessment Tool
t	Time counter in years
TEMP	Temperature
TSK	Takagi-Sugeno-Kang
TUR	Turbidity
USA	United State of America
USGS	United State Geological Survey
VAF	Variance account for
VC	Vapnik-Chervonenkis
WANN	Wavelet-artificial neural network
WRC	Water Research Commission

CHAPTER 1

GENERAL INTRODUCTION

1.1 INTRODUCTION

Sediment yield is a problem to hydrological structures especially reservoirs and the demand on water resources in South Africa are increasing due to economy expansion and increase in population. Therefore an alternative or improve management of water resources is necessary to improve both the quality and quantity of water from reservoirs and dams. The impact of human activities and the effect of environmental degradation on natural resources are gradually changing the manner in which natural resources are viewed and managed (Brierley and Fryirs 2008). This has led to considering the development of sustainable management policies for these resources a priority and a goal that must be achieved as soon as possible (Chaves and Alipaz 2007). Therefore, more attention should be given to management decisions involving environmental predictions based on the past, present and future condition of the environment (Dessai and Hulme 2007).

The international development and use of reservoirs is an effective way of managing water which is one of life's essential natural resources. This has also helped in the management of rainfall runoff, drought and also in meeting energy demands, especially in developing countries like South Africa (Ran *et al.* 2013). The inflow of water containing sediment into these reservoirs can, over time, result in the reduction of the storage capacity and efficiency of the reservoir. Reservoirs all over the world are losing between 0.5 to 1% of their storage capacity yearly due to sedimentation and if appropriate measures are not implemented, 25% of the world's current water storage capacity will be lost in the next 25-50 years (Dams 2000). The presence of sediment and the resulting turbidity in rivers will greatly affect its usage for water supply, as well as the water quality (Mukundan *et al.* 2013). It is therefore imperative to determine the sediment yield and spatial and temporal patterns to be able to

manage and predict sediment transport processes under both present and future conditions (Mukundan *et al.* 2013). Hence, the regular use of hydrographic surveys and the accurate prediction of sediment yield are very useful in the maintenance of reservoirs including the planning for dredging activities.

The presence of solid matter like suspended sediment, soil and other particles in a river will affect its aquatic life, chemical composition and the turbidity level (Duan *et al.* 2013). According to Turner *et al.* (2006) the presence of sediment in reservoirs during heavy rainstorms can result in the accumulation of sediment leading to the rise of the level of the channel bed by the deposited sediments. This is detrimental to the functionality of the reservoirs as the quality of water and the storage capacity will be affected. It can also result in flooding, thereby losing the channel capacity of the reservoir required to convey the floodwaters (Aytek and Kisi 2008). Therefore, the accurate quantification of sediment flowing into reservoirs is of utmost importance for proper management and dredging in harbour to improve navigation. This task is extremely difficult to achieve because of the temporal variation of both sediment load and streamflow and these measurements are not always available because accurate measurement is an expensive procedure (Partal and Cigizoglu 2008). The use of artificial intelligence models in water resource management over time will be an effective tool for proactive management strategies. It has also helped to test and understand effective methods to extrapolate and interpolate in a significant manner, framing system-specific applications in a broader context (Winz *et al.* 2009).

1.2 PROBLEM STATEMENT

In the design of a reservoir, its design life, maintenance, storage capacity, operations and the quantity of sediment flowing into it are very important factors to be considered (Morris *et al.* 2008). However, according to Heng and Suetsugi (2013) the measurement of sediment yield is deficient in most parts of the world. Several hydrological variables such as bed-form geometry, flow rate, friction factor and discharge have been used to develop different models for predicting sediment yield in rivers (Karim and Kennedy 1990; Lopes and Ffolliott 1993) but this limited historical data is often demanding to measure in a developing country like South

Africa where budget limitations and the fact that it is labour intensive has become a major problem for accurate predictions (Raghunath 2006). The relationship between these variables is often assumed by the time series techniques but this cannot be the case with real hydrological data. This necessitated the use of artificial intelligence models like genetic programming (GP) for more accurate predictions (Kisi *et al.* 2012). This study therefore aims to demonstrate the potential of genetic programming (GP), an evolutionary-driven technique, for the estimation of suspended sediment in a reservoir with limited data.

1.3 RESEARCH OBJECTIVE

The aim of this study is to estimate the monthly quantity of suspended sediment transported into Inanda Dam using genetic programming. The specific objectives are as follows:

- To develop monthly models using genetic programming techniques and sediment rating curve to estimate the quantity of suspended sediment flowing into Inanda Dam.
- To evaluate the accuracy of the developed monthly models by applying standard evaluation criteria to individual models.
- To compare the accuracy of the developed models in order to determine the most accurate model for each month for predicting the quantity of suspended sediment flowing into Inanda Dam.
- To estimate the monthly quantity of suspended sediment flowing into Inanda Dam using sediment rating curve.
- To compare the performance of the most accurate GP models in each month with its corresponding sediment rating curve.

1.4 RESEARCH METHODOLOGY

In this study, GP was employed to predict the monthly suspended sediment flowing into Inanda Dam. Also, sediment rating curve will be used to do a comparative analysis of the GP's performance. Genetic programming, a data-driven modelling method, has been widely used in developing hydrological models, where the

relationships between the relevant variable (inputs and outputs) are poorly understood. Genetic programming starts with an initial population of randomly generated programs. Each individual program in the population was measured in terms of how well it performs in that particular problem environment. The Darwinian principle of reproduction and survival of the fittest, as well as the genetic operators of crossover and mutation was used to create a new offspring population of programs, generation by generation. GP is a robust and dynamic model. It has been widely-applied to solve different kinds of difficult problems (Chiang and Chang 2009). GP has the capability to select the best input from its variables and this makes it possible for the input and output variables to be expressed as a regression equation. GP has been applied to all aspects of life, even to the field of water resources engineering (Akbari-Alashti *et al.* 2014; Datta *et al.* 2014; Fallah-Mehdipour *et al.* 2014; Orouji *et al.* 2014).

In this study, upstream monthly suspended sediment concentration and stream flow data for Inanda Dam over the past 24 years (1990-2014) was collected from Umgeni Water, South Africa and the Department of Water Affairs, South Africa. This data was used as input for training and testing the developed GP Models and developing a sediment rating curve. A GP open source software, known as 'GPdotNET', was used for implementing the GP. GPdotNET was chosen because it supports multi-core parallel processing based on the ParallelFx library and can also handle large quantities of datasets (Wijayaweera and Karunananda 2012).

1.4.1 Genetic programming

In this study several input combinations of suspended sediment and stream flow variables were used to develop various GP models (Aytek and Kişi 2008). The selection of the combination of input variables was based on the strength of their correlation with the observed suspended sediment yield (Oyebode *et al.* 2014). The model for each input combination was developed using five series of arithmetic operators and mathematical functions as stated in Table 1 (Sheikhalipour and Hassanpour 2013):

Table 1: GP arithmetic operators and mathematical functions

Functions series	Arithmetic operators and basic mathematical functions
A	+, -, *, /
B	+, -, *, /, Exp(x), √
C	+, -, *, /, Power(x^2), log, ln(x)
D	+, -, *, /, Exp(x), √, Power(x^2), log, ln(x)
E	+, -, *, /, Exp(x), ln(x), √, Power(x^2), log sin(x), cos(x), tan(x)

The accuracy of the developed models from each input combination for each series of arithmetic operators and mathematical functions (A-E) was evaluated using the Root Mean Square Error (RMSE) and Determination Coefficient (R^2) between the model output and the measured output.

1.4.2 The sediment rating curve

The accuracies of the developed GP Models were also compared with the accuracy of the sediment rating curve. This involved plotting a graph of monthly suspended sediment data against monthly stream flow data. This is one of the most common methods of estimating suspended sediment load in rivers. An equation (model) was developed from the curve (line of best fit) of the graph for each month of the year. This curve is typically defined as an exponential relationship between water flow and sediment discharge ($Qs = a.Q^b$). In which Qs is suspended sediment discharge (mg/l), Q is stream flow (m^3/s), and a, b are coefficients estimated by means of logarithmic linear regression between log Qs and log Q for the Inanda Dam catchment area. The accuracy of each monthly equation (model) to predict suspended sediment yield in Inanda Dam was also evaluated using the Root Mean Square Error (RMSE) and Determination Coefficient (R^2) between the models outputs and the measured outputs. These were ultimately compared with those of the developed GP models to determine the 'best' model for estimating the quantity of suspended sediment flowing into Inanda Dam for each month of the year.

1.5 DESCRIPTION OF THE STUDY AREA - INANDA DAM

Inanda Dam started operations in 1989. It is the lowest impoundment on the Umgeni river (Hudson *et al.* 1990) as indicated in Figure 1.

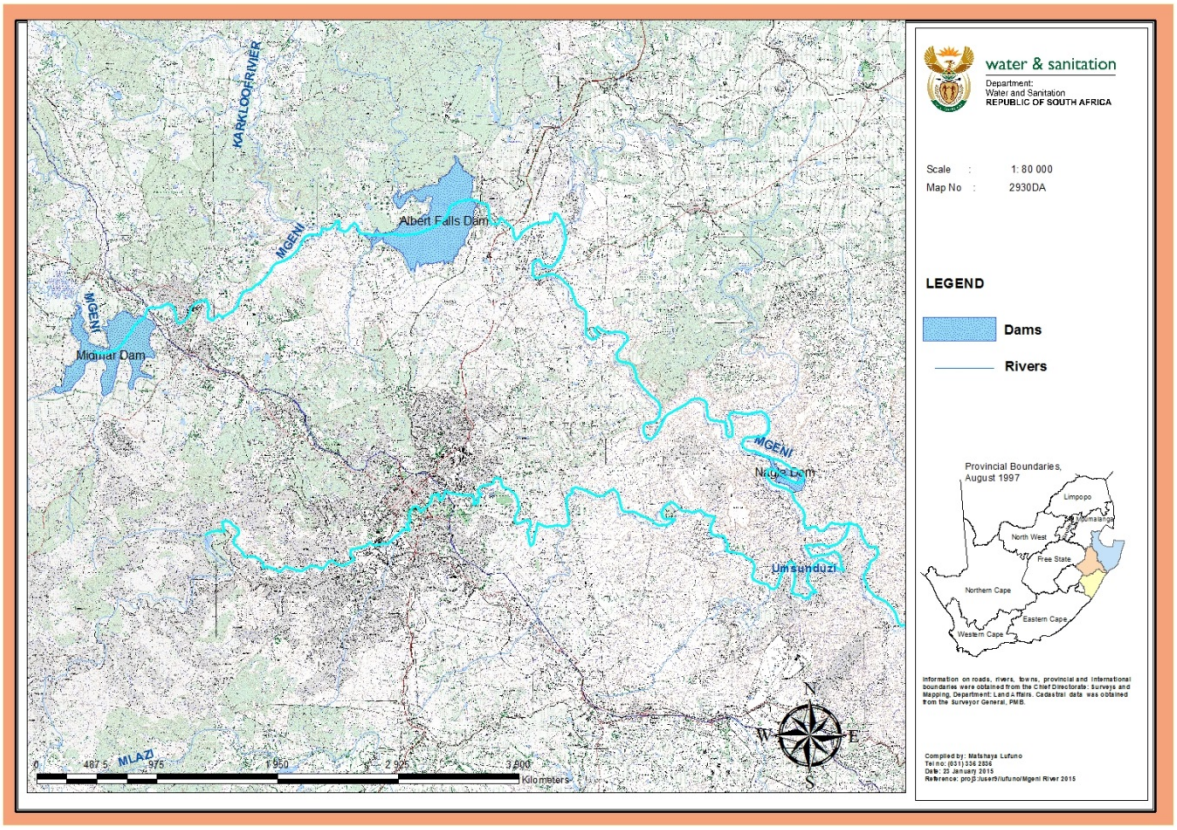


Figure 1: The Umgeni river system

(Source: The Department of Water Affairs, Durban)

The Inanda Dam is situated in the Valley of a Thousand Hills, 42km North of Durban, South Africa. The length of the dam is approximately 23km from the dam wall to the bridge, its widest point is approximately 1.5km and the deepest point is about 50m. Its surface area covers about 1440 hectares with a shoreline at full capacity of approximately 100 km (DWAF 2008). Table 2 shows detailed statistics of the dam.

The Dam is located in a region of high summer rainfall with an annual precipitation ranging from 800 mm to 1125 mm. Summer temperatures range from 25°C to 38°C while winter temperatures range from 9°C to 19°C. Shale, Tillite, Gneiss and Arenite

are the main rock types found in the catchment area. As illustrated in Figure 2 the Dam is fed by a number of rivers flowing in from the surrounding hills. However, the Umgeni River is its main inlet and outlet point. Figure 3 shows Umgeni River flowing into the dam at the inlet of the dam and the sampling point for both suspended sediment load and the streamflow used in this study. The percentage of clay in the soil ranges from 0% to 35% (DWAF 2008).

Table 2: Inanda Dam statistics (DWAF 2008)

Component	**Description**
Maximum height of dam wall	57 m
Length of dam wall	610 m
Length of spillway	140 m
Approximate length of dam	29 km
Approximate width of dam	1.5 km
Spillway discharge at high flood level (1:200yr)	4000 m^3/s
Maximum capacity of outlet works (river outlet)	34.4 m^3/s
Maximum capacity of outlet works (service outlet)	14.2 m^3/s
Maximum spillway capacity (including emergency spillway)	11 100 m^3/s
Gross storage capacity	256 million m^3
Annual yield	96 million m^3/annum
Full drawdown level/FSL	147.00 m
Lowest drawdown level	115.75m
Non-overspill crest level	157 m
Lowest foundation level	92m
Surface area at FSL	1426 hectares
Catchment Area	3949km^2
MAR (mean annual runoff) natural	60.09 million m^3
MAR (mean annual runoff) observed	57.96 million m^3
MAP (mean annual precipitation)	870 mm
Regional maximum flood	8100 m^3/s

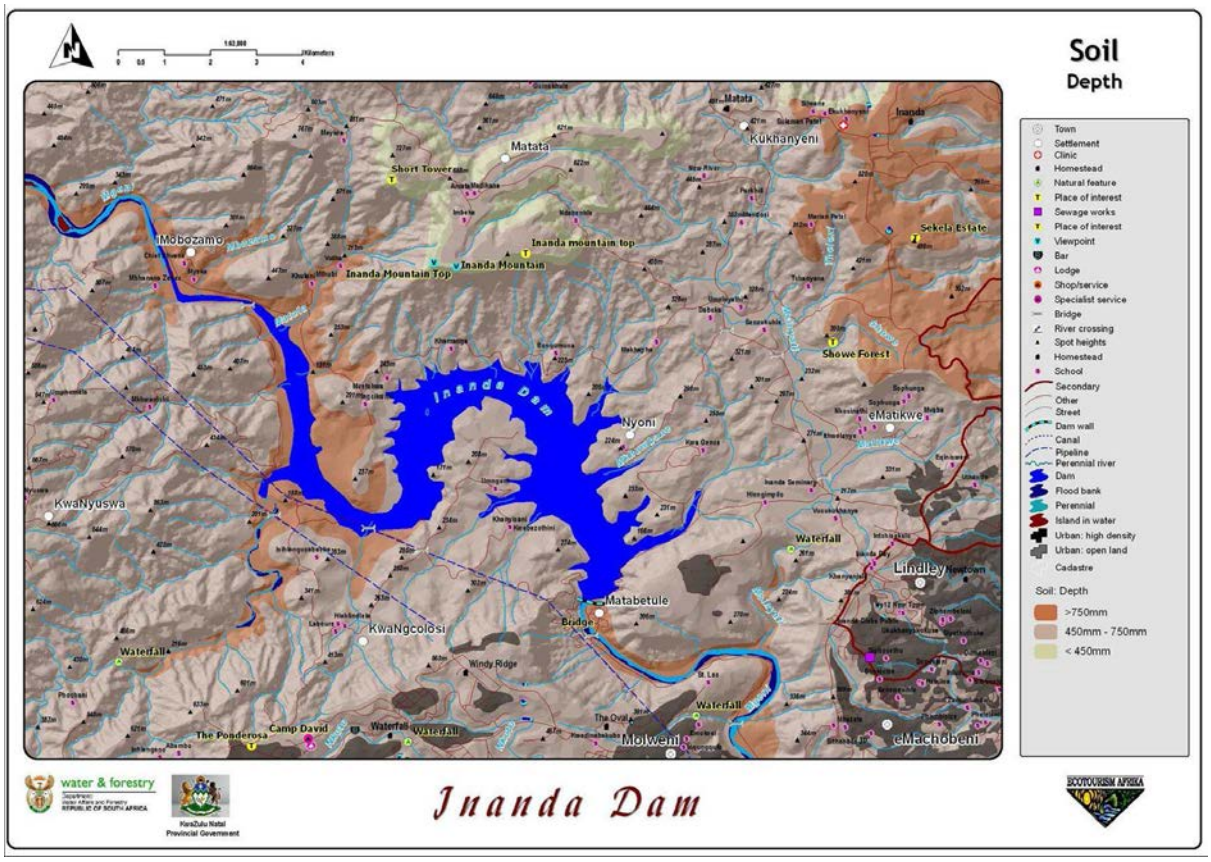


Figure 2: Inanda Dam catchment and its soil depth

(Source: DWAF 2008)

Figure 3: Inlet of Inanda Dam

(Source: The Department of Water Affairs, Durban)

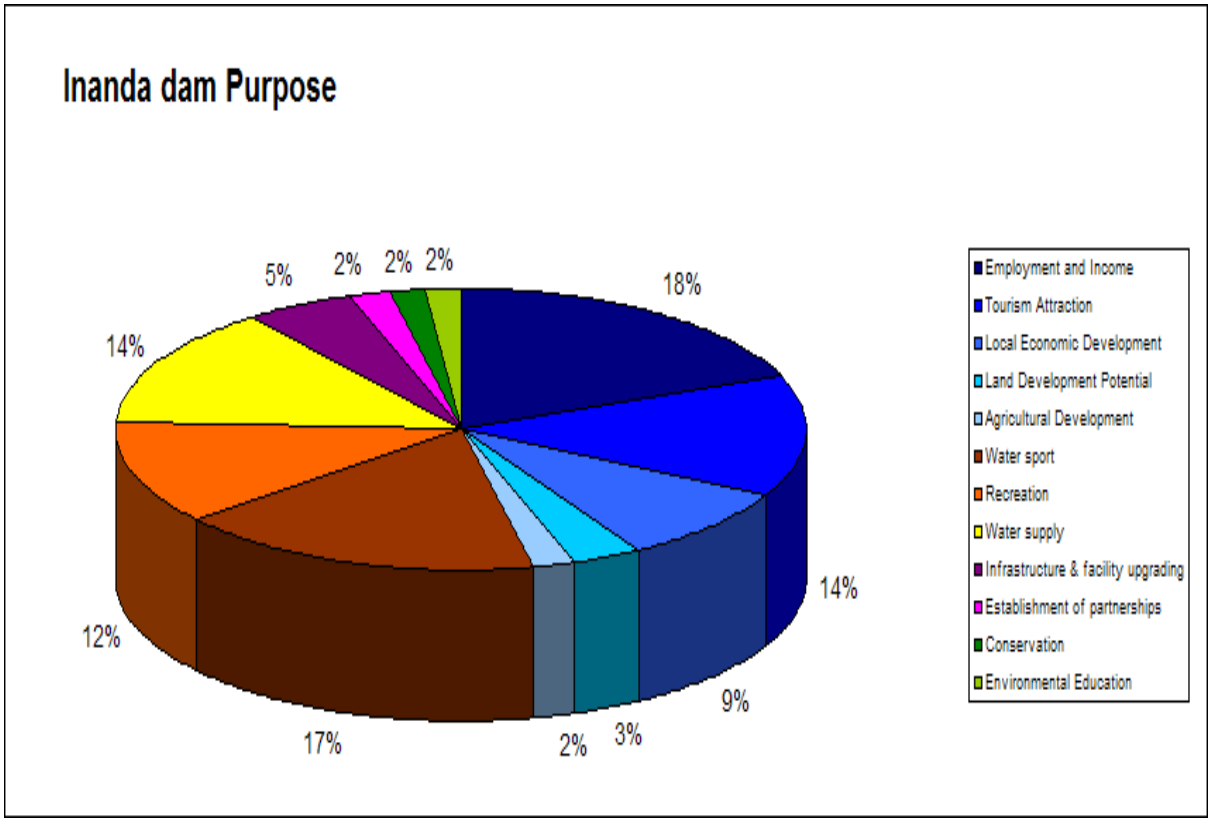


Figure 4: Purpose and significance of Inanda Dam

(Source: DWAF 2008)

The dam is used for the following purposes: Economic (46%), Activity (29%), Consumptive (14%), Social (7%) and the Biophysical Purpose as further illustrated in Figure 4. Therefore, the importance of the functionality of the dam cannot be over emphasized.

1.6 RATIONALE OF THE STUDY

The accurate estimation of the quantity of suspended sediment in Inanda Dam is of great importance due to the following outlined reasons:

1. The dam supplies a large proportion of the South African population with water (Tollow 2004).

2. Sediment controls water quality, riverine hydrology, aquatic ecology and river channel morphology so it is very important for the sustainable development of a water resource system (Melesse *et al.* 2011).
3. According to Hudson *et al.* (1990) the water flowing into Inanda Dam is impaired with impurities (sediment) from upstream activities like industry, agriculture, and urbanization, making it eutrophic and expensive to treat and manage.
4. As a result of sand mining close to the inflow of the dam on the banks of the Mshazi and Umgeni rivers, there is an increase in the quantity of suspended sediment flowing into the dam (DWAF 2008).
5. The presence of suspended sediment in a river plays an important role in the aquatic ecosystems but in excess it may physically damage the aquatic biota and result in changes to the habitat conditions (Dunlop *et al.* 2008).
6. Therefore, the regular use of hydrographic surveys and the accurate prediction of sediment yield is very useful in the maintenance of reservoirs including the planning for dredging activities (Bhattacharya and Solomatine 2006).

1.7 SIGNIFICANCE OF THE STUDY

According to DWAF (2008) the purpose of Inanda Dam is to store water, to act as an economic catalyst and provide recreational value to the Inanda community. The presence of suspended sediment in the dam leads to downstream erosion of these sedimentary depositional environments, and increased sediment build-up in the dam. The dam will eventually develop a reduced water-storage capacity due to the exchange of storage space for sediment. Therefore, the development of a model to estimate the quantity of sediment flowing into the dam on a monthly basis will assist in determining when sediment deposited into the dam can be removed by dry excavation or hydraulic dredging. The model will help in determining the timing and rate of water removal which greatly affect the extent and rate of downstream release and reservoir sediment erosion (East *et al.* 2015). This will ensure that the dam meets its optimal usage potential of domestic and industrial water demand. The recognition of patterns of the quantity of suspended sediment in Umgeni river flowing into

Inanda Dam will also help in understanding the environmental characteristics and fluvial morphodynamics of the river (Park and Latrubesse 2014).

The results of this study will also serve as a source of important baseline information for stakeholders in the water industry such as Umgeni water, the South African Water Research Commission (WRC), and the Department of Water Affairs (DWA) in protecting against the harmful impact of sedimentation of dams as well as protecting the limited water resources and the environment. The results will help them to provide guidance on future management of water resources in such areas as droughts fighting, flood fighting, meeting the increase in energy demands, among others. The accurate determination of suspended sediment flowing into the dam will also contribute to current studies by stakeholders in the water industry on carbon sequestration, reservoir management, and sediment budget especially for a dam like Inanda Dam with large quantities of sediment flowing into the dam and subsequently being trapped within the dam (Ran *et al.* 2013).

1.8 OUTLINE OF THE THESIS

This thesis is divided into five chapters.

Chapter 1: Introduction

In this chapter, the importance of water resources in the environment is discussed. The effect of sediment builds up in reservoirs is also discussed. The study area is briefly described and the purpose of the reservoir (Inanda Dam) is stated. This chapter also contains the rationale for this study and its importance, structure of the thesis and publications from this study.

Chapter 2: Literature review

This chapter extensively reviews artificial intelligence (AI) models and its role in hydrology. This chapter also outlines literature relevant to the theories and applications of evolutionary algorithms and most especially genetic programming (GP). It discusses GP's areas of concern and how to improve its performance coupled with its advantages and disadvantages. This chapter also describes the use of

sediment rating method in determining the relationship between discharge and sediment yield.

Chapter 3: Genetic programming for suspended sediment prediction

This chapter discusses the development of monthly suspended sediment models using upstream monthly streamflow and suspended sediment data for 23 years from 1991 to 2013. This involves the use of four different input combinations of monthly streamflow and suspended sediment data, with five series of arithmetic operators and mathematical functions each, to develop multiple models. The developed models are evaluated using two statistical goodness-of-fit tests, namely: determination of coefficient (R^2) and root mean square error (RMSE) and the result analysed.

Chapter 4: Sediment rating curve for suspended sediment prediction

This chapter examines the relationship between monthly streamflow and suspended sediment using the sediment rating curve. The monthly sediment yield prediction from the sediment rating curve is also evaluated using R^2 and RMSE. Finally in this chapter the results from the GP models are compared with those from the sediment rating curve.

Chapter 5: Conclusions and recommendations

This chapter gives a detailed conclusion based on the results from Chapter 3 and Chapter 4 and a general conclusion on the research. It also discusses areas of concern and recommends future areas of study. This thesis is written in such a way that each chapter is a research paper in its own right, so there may be overlapping contents within the chapters.

1.9 PUBLICATIONS

This study has produced the following publications:

1. Adesoji Tunbosun Jaiyeola and Josiah Adeyemo (2015). Genetic programming for predicting suspended sediment. Prepared manuscript for Water SA.

2. Josiah Adeyemo and Adesoji Jaiyeola (2015). Sediment rating curve for suspended sediment prediction and performance comparison between GP and SRC. Prepared manuscript for South African Institute of Civil Engineering (SAICE) Journal.

3. Adesoji Tunbosun Jaiyeola, Josiah Adeyemo and Fred Otieno (2014). Estimation of suspended sediment yield in Inanda Dam using Genetic Programming. *Water Institute of Southern Africa (WISA) Biennial Conference*, 25-29 May 2014, Nelspruit, South Africa.

4. Adesoji Tunbosun Jaiyeola, Josiah Adeyemo and Fred Otieno (2015). Review of theories and applications of genetic programming in sediment yield modelling. *17th International Conference on Environmental Sciences and Engineering*, 17 – 18, September 2015, Rome, Italy. Accepted.

5. Adesoji Tunbosun Jaiyeola, Josiah Adeyemo and Fred Otieno (2015). Role of Artificial intelligence in Water Resources Modeling. *16 WaterNet/WARFSA/GWP-SA Symposium*, 28 - 30 October 2015, Mauritius. Under review.

CHAPTER 2

LITERATURE REVIEW

2.1 PREDICTIVE MODELLING TECHNIQUES IN SEDIMENT YIELD AND HYDROLOGICAL MODELLING

Sediment yield can be considered to be the total sediment load that leaves a drainage basin. It could be also be defined as the quantity of sediment per unit area removed from a watershed by flowing water during a specific period of time (Griffiths *et al.* 2006). Sediment yield can be divided into 3 categories namely: suspended sediment – these are the particles suspended due to turbulence in the upper portion of a river just below the water surface. It comprises of silt, clay, and sand size; bed load – these are larger particles that move along the bottom of a river, saltation –these are particles that bounce up and down the top and bottom of a river. They are mostly sand size or gravel (Bender *et al.* 2005). Figure 5 shows an illustration of the various categories of particles found in a river.

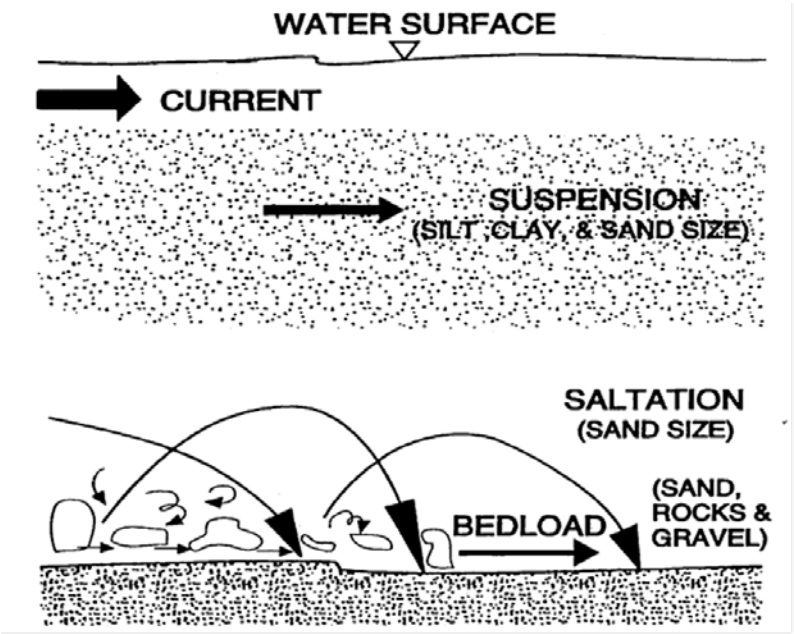


Figure 5: Layers and movement of particles in a river

(Source: Bender *et al.* 2005)

The effect of rainfall splash detachment and entrainment through overland flow also generates sediments. Detachment takes place when locally induced shear stress exceeds the cohesive strength of the soil (Loch and Silburn 1996). Geomorphic characteristics like vegetation cover, land use, precipitation, sediment storage, drainage density, topography, soil erodibility and sediment transport capacity affect the sediment yield of a river basin (Bender *et al.* 2005). This is increased through soil disturbance during land use, unstable geological terrain and/or a high rainfall zone. All rivers contain sediments. When a river is stilled behind a dam, some of the sediments sink to the bottom of the dam. As sediments accumulate, the reservoir gradually loses its ability to store water for the purposes for which it was built. Every reservoir loses storage to sedimentation, although the rate at which this happens varies widely.

Sedimentation is a major technical problem faced by marine industries. Apart from rapidly filling reservoirs, sediment-filled rivers also cause abrasion to turbines and other reservoir components. The knowledge of the quantity of sediment present in a river at a particular time can lead to a better understanding of flood capacity in reservoirs and consequently help control over-bane flooding. The reduction of sediments in a reservoir also has the following advantages: it improves water quality; it makes the water more suitable for man and aquatic life; it allows the design storage to be maintained and it allows for better navigation. The development of hydrological models to forecast the quantity of sediment that will be present in a river at a given time helps planners and managers of water resource systems to understand the system better in terms of its problems and to find alternative ways to address them (Loucks *et al.* 2005). According to Merritt *et al.* (2003) many models are used for simulating and estimating sediment yield transport but they differ in the processes considered and the complexity and data required for calibration and usage. Furthermore it was stated that the most appropriate model depends on the intended use and the features of the catchment in consideration. The following factors also affect the choice of a model: the intended model capabilities and features; the simplicity of the model and output scales; the assumptions made in using the model;

validity and accuracy of the model and the data required by the model (Merritt *et al.* 2003).

2.2 TYPES OF HYDROLOGICAL MODELS

Based on the model algorithms, its data dependence and the physical processes simulated by the model, hydrological models can be categorised into three groups namely: empirical or statistical/metric models, physical based models and conceptual models.

2.2.1 Empirical models

Empirical models are based on the analysis of observations and the characterization of responses from the observed data (Wheater *et al.* 1993). These models require fewer amounts of data and lower computational requirements when compared to other types of models. They have a high level of spatial and temporal aggregation and also incorporate a small number of causal variables (Jakeman *et al.* 1999). The parameter values are obtained from the calibration of experimental sites and are very effective in the identification of sediment sources and nutrient generation (Merritt *et al.* 2003). These models are criticised for using unrealistic assumptions concerning the physic of the catchment while ignoring the catchment heterogeneity inputs and characteristics such as soil types and rainfall (Wheater *et al.* 1993). They also do not respond to events thereby neglecting the impact of rainfall-runoff on the catchment being modelled (Wheater *et al.* 1993). Despite these shortcomings, the more complex and dynamic models in this regard cannot be considered as better when compared to the empirical models because empirical models can be used where there are limited data and parameter inputs. In addition, empirical models can also be used as a first step when identifying sources of sediment and nutrient generation (Merritt *et al.* 2003). Fournier, Dendy and Bolton, and Revised Universal Soil Loss Equation (RUSLE) (Kouli *et al.* 2009; Garg 2011) are examples of empirical models.

2.2.2 Conceptual models

Conceptual models view flow path in catchment as a series of internal storages. They generally consider the description of catchment processes but neglect the specific interaction between the processes (Sorooshian 1991). Therefore, both qualitative and quantitative effects of land use changes are indicated in these models. Parameter values in conceptual models are determined by the calibration against observed data, leading to problems associated with its identification (Abbott *et al.* 1986; Jakeman and Hornberger 1993). Spear (1997), observed that as a result of the calibration techniques used for medium complexity models, many possible 'best' parameter sets can be made available. Thus calibration and identification of additional parameters using *a priori* knowledge of the system can limit the number of parameters to be estimated (Kleissen *et al.* 1990). According to Beck (1987) conceptual models can be used as an intermediary model between empirical models and physics-based models. They also reflect the principles that govern the system to be modelled. Examples of such models include Agricultural Nonpoint Source Pollution (AGNPS) and Morgan-Morgan-Finney (MMF).

2.2.3 Physical based models

These are based on the use of basic physical equations which are solutions that describe discharge and sediment generation. Standard equations such as the equation of conservation of mass for sediment and the equations of conservation of mass and momentum for flow are used in physics-based models (Bennett 1974). In theory, its parameters are measurable but conversely, this is not obtainable in practice due to their large numbers, hence they are calibrated against observed data (Beck *et al.* 1995). The calibration of the model's parameter values when there are occurrences of missing values usually result in uncertainty in the model outcome. Furthermore, uncertainty can also occur when the parameter values cannot be measured. This is due to the likelihood of the occurrence of error during measurements (Kalma and Sivapalan 1995). Examples of widely-used physics-based models include Gridded Surface Subsurface Hydrologic Analysis (GSSHA) (Borah 2011; Pradhan *et al.* 2014), Hydrologic Simulation Program Fortran (HSPF) (Kim *et al.* 2014), Kinematic

Runoff and Erosion Model (KINEROS2) (Sajjan *et al.* 2014), MIKE SHE (Bjørnholt Karlsson *et al.* 2014) and Soil and Water Assessment Tool (SWAT) (Huang *et al.* 2015).

2.2.4 Selecting an appropriate model

The choice of a model depends mainly on its purpose, so a model may not necessarily be used for all modelling situations. From the literature (Letcher *et al.* 1999; Perrin *et al.* 2001; Thorsen *et al.* 2001), the choice of a model largely depends on where the emphasis will be laid, that is, either on the processes of the work or on the expected output that addresses the problem. For instance, in a study carried out by Thorsen *et al.* (2001) both empirical and conceptual models were used to predict nitrate concentrations in groundwater aquifers in a catchment. The models' forecasting abilities were considered because of the semi-empirical nature of the process. Letcher *et al.* (1999), in their technical report, argued that simple empirical and conceptual models can function more effectively when used within a developed framework. Also Perrin *et al.* (2001) examined the relationship between the number of optimized parameters and model performance in 429 catchments. They stated that over-parameterization of rainfall-runoff models can greatly affect the ability of the model to forecast stream flow.

2.3 ARTIFICIAL INTELLIGENCE IN HYDROLOGICAL STUDIES

Water is one of the most important natural resource. Even though its importance to our existence on earth cannot be over emphasized, it is also an important raw material for many industries. Though this resource is abundant and almost 70% of the Earth's surface is covered with water, the amount of freshwater that can be effectively utilized is very limited (Lloret 2013). There is an acute water shortage and people often have to face hardships in getting access to potable water due to depleting water tables, drying up of wells and rivers and irregular rainfall. Due to these shortages, large swathes of land are being rendered barren. The excess water in the form of floods, which causes changes to river courses and rising sea levels have resulted in large scale destruction of life and property. The scarcity of water also has

political implications in many countries around the globe. While some do not favour equitable sharing of river water, others oppose the construction of dams and the diversion of river waters.

From the foregoing, it is clear that proper management of this precious natural resource is very important for the survival of mankind and other forms of life. This calls for research in finding and implementing new methods and techniques for the proper management of limited water resources to ensure a reliable supply of water for fulfilling the needs of the society. Such new methods and techniques can be readily implemented if a good foundation for understanding water resources and the consumption patterns is laid. It is necessary to study the quality and quantity of water available over the years and then match this availability to the demands of various stakeholders. It is necessary to develop models that can predict the water inflow patterns and develop methods and techniques that can ensure proper utilization of water. The field of water resource modeling includes the accurate forecast of rainfall patterns and modeling the flow of the rainfall water. It also involves the applications of various methods and techniques to purify collected water and wastewater for reuse. These processes are complex and thus the development of more accurate and reliable models for these processes is ongoing. There have been numerous attempts to develop these models, many of which have successfully served the purpose in specific situations, for example, an early attempt in forecasting rainfall patterns could be traced back to 1851 when Mulvany used self-registering rain and flood gauges to observe the relationship between flood discharges and rainfall (Mulvaney 1851).

On a broad scale, the different models used in water resource modeling are typically distinguished on the basis of the approach followed for describing the spatial extent of watershed and the hydrologic processes involved. The watershed models are typically classified as lumped or distributed models (Singh and Woolhiser 2002; Anctil *et al.* 2003) whereas the hydrological processes are classified as knowledge-driven models or data-driven models (Solomatine and Ostfeld 2008). Another category of models called mechanistic models use differential equations to describe the processes at the surface and subsurface (Beven *et al.* 1987; Yang *et al.* 1998). A special class of mechanistic models which focus on storage elements are called

conceptual models (Bergström and Singh 1995; Singh 1995). Fewer numbers of parameters are required to describe the system in these models. There are also models that use information from hydro-meteorological data to map the relationship between rainfall and runoff, these models are known as precipitation-runoff data driven models.

2.3.1 Artificial intelligence data driven models

In these models the parameters, or both the parameters and the structure are the unknowns. These unknowns are calculated from the time series analysis. The time series generating process is assumed to be stochastic and time invariant. The other class of models assume predefined structures and these include linear and nonlinear regression models (Salas 1980; Box *et al.* 2013), linear perturbation models (Nash and Barsi 1983), transfer function models (Kachroo 1992) and constrained linear system models (Natale and Todini 1976). Still another class of model is one in which no assumptions are made about the structure as the structure is learnt from the data. Such models are known as artificial intelligence data driven models (AIDDM) and these include K-nearest Neighbour (KNN) models (Karlsson and Yakowitz 1987), chaos theory based models (Solomatine *et al.* 2000), artificial neural network (ANN) based models (Hsu *et al.* 1995; Abrahart *et al.* 2012), Support Vector Machines (SVM) (Liong and Sivapragasam 2002; Kisi and Cimen 2011), Genetic Programming (GP) based models (Liong *et al.* 2002; Poli and Koza 2014) and fuzzy inference systems (Jahanshahi *et al.* 2015).

These models are useful in situations where sufficient data to calibrate the knowledge driven models is unavailable. These models as well as the regression models perform well as shown by numerous studies (Yu *et al.* 2006; Wang *et al.* 2009; Chen *et al.* 2012; Nayak *et al.* 2013; Reed *et al.* 2013; Tayfur 2014). In the next sections of this review, emphasis is laid on this class of model. Firstly, the fundamentals of AIDDM are explained and then a review of literature is presented. Some of the popular models are discussed individually. In each model, the methodology and the data analysis approach used are discussed. The advantages and disadvantages of the respective models are also discussed. Recent studies where

these models have been applied to the modeling of water resources are also highlighted. This study reviews studies in which more than one of these models was used to analyze a particular situation. Such studies can be helpful in establishing the suitability of one model over another in specific situations. Generally, the results from these studies are mixed and no one particular model has been found to consistently out-perform other models. In conclusion, this review briefly mentions the studies in which modified versions of the above models or a mix of models have been used. These studies indicate that such a mix of models can provide the desired levels of accuracy.

2.4 GENERAL APPROACH FOR APPLICATION OF ARTIFICIALLY INTELLIGENT DATA DRIVEN MODELS (AIDDM)

Collection of raw data, pre-processing of the data, selection of an appropriate model, identification of a model and finally, the evaluation of that model are the various stages in a typical AIDDM. Each stage is a complicated process and various methods and techniques are employed to reach a goal as quickly and cheaply as possible. Artificial intelligence techniques are helpful in performing these tasks (Oyebode *et al.* 2014) as they can be used to identify ways to collect data, select an effective pre-processing technique, select an appropriate model and an algorithm to calculate the parameter values for that model. These techniques attempt to minimize the difference between the predicted and target values to arrive at an optimal model. As was mentioned previously, these models do not define the relationship between the predictor and the response but rather only establish a functional relationship by training the model over a data set. Thus the structure of the model is not defined beforehand and the model identified in this way can be nonlinear as well as non-parametric. It is important to note that the training phase during the development of these models is of utmost importance as the improper selection of training data can result in errors in the identification of the model. Also, data driven models do not require a Gaussian distributed data set as is often the case in classical regression models. Thus such models can easily work even if there are irregular seasonal variations in the training data. Finally, the model is evaluated on a test data set and is

accepted only if its performance is satisfactory on this test data set. Often limited data is available to train and test the model and in such cases different combinations of the available data are used to train and test the model.

In the next sections various artificial intelligence models are discussed individually starting with the K-nearest Neighbour method.

2.5 K-NEAREST NEIGHBOUR (KNN) METHOD BASED MODELS

K-nearest Neighbour (KNN) method is a non-parametric model that does not make any assumptions about the data distribution of the time series. The data used for the development of the model can be scalar or vector. The data exists in a 'feature space' or 'metric space'. A class label is associated with each vector in the training data set and its K-nearest Neighbours influence each classification. Thus the input set consisting of K training examples closest in the feature space is used for training, testing and prediction. This algorithm is also called a 'lazy' algorithm as the model does not generalize from the training data set and the generalization is deferred until a new query is encountered. This methodology results in lower training times and consequently a longer test phase because all data points are used for testing. Consequently, this method has higher memory requirements (Yakowitz 1987; Altman 1992). Before utilizing the data in the model, it is normalized to eliminate effects owing to the origination of data from different sources. Additionally, the feature extraction and dimension reduction of data is done before the application of the KNN method.

The KNN method is used for both classification as well as regression. During classification, the KNN method assigns class membership to the input data. This class membership is determined by finding the class to which the greatest numbers of neighbors of an element belong. Similarly, in regression, the output is also determined by averaging the value of its K-nearest neighbors. The distance between the data points is calculated using one of the standard distance measures like Euclidean distance, Mahalanobis distance or Minkowski distance. The 'majority voting' rule is used for classification if the class distribution is skewed (Ashab *et al.*

2014). In such cases, the model is extended and the weight of the neighbors is varied such that weight is inversely proportional to the distance of the neighbor. The training phase involves storing the feature vectors and the class labels of the training data. The number of neighbors, K, is chosen to achieve a balance between the effect of noise and sharpness of boundary of the classes. Choosing a large value of K reduces the effect of noise but at the same time it leads to diffused boundaries. Conversely, a small K results in a distortion of classification due to noise. The important parameters of interest include K and the dimension of the feature vector, d (Lee and Ouarda 2011). The parameters are obtained using global optimization methods. For this purpose, the training data is separated into two groups, one of which is used for constructing the nearest neighbor predictors and the second group is used for calibrating the parameters.

The advantage of the KNN method is that it is robust and can tolerate noise and irrelevant features in the data (Oyebode *et al.* 2014). It can also represent both probabilistic and overlapping features. On the other hand, the disadvantages of the KNN method include its inability to extrapolate, as it doesn't identify a functional relationship between input and output (Benyahya *et al.* 2007). Furthermore, the KNN predicted extremes are confined to the extremes in the training data and the previous test data and do not extrapolate to further extremes even if there is a strong trend indicating the same (Hessenmoller and Elsenhans 2002).

A survey of literature indicates that the KNN method was initially applied to predict rainfall-runoff transformation (Karlsson and Yakowitz 1987) and has since been used to simulate daily precipitation in an area (Rajagopalan and Lall 1999), prediction of river water temperature (St-Hilaire *et al.* 2012), water quality classification (Nowak *et al.* 2010), virtual water (Yakowitz 1987; Ahmadaali *et al.* 2013) and the prediction of stream-flow in rivers (Lee and Ouarda 2011; Modaresi and Araghinejad 2014).

2.6 CHAOS THEORY (CT) BASED MODELS

Chaos theory suggests that even completely deterministic systems can be unpredictable. The unpredictability of future behaviour arises from the fact that even

a very small difference between the initial conditions results in completely different trajectories (Sivakumar and Berndtsson 2010). Such small differences cannot be avoided as they can result from the parametric perturbations, errors in measurement or simply from the limitations of the measuring devices. Such extreme sensitivity of the system to initial conditions and deterministic, bounded and aperiodic behavior are thus characteristics of such chaotic systems. This bounded and aperiodic trajectory of such systems is called an attractor. Chaos has been observed in weather patterns and many other natural and engineering systems (Shepherd 2014). To apply the chaos theory based on theories in water resource management, the researchers assume that the time series data associated with water bodies is a low dimensional, chaotic attractor (Solomatine *et al.* 2000; Solomatine and Ostfeld 2008). Typical methods to identify if a time series is chaotic or not include (a) Correlation dimension, (b) Largest Lyapunov exponent and (c) Kolmogorov entropy (Li *et al.* 2014). Specifically, a system can be considered as chaotic if the correlation dimension's saturation value exists for the correlation integration method or if the maximum exponent by the Lyapunov method or maximum entropy for the Kolmogorov method are greater than 0 (Ghorbani *et al.* 2010). Of these, the correlation dimension method is a better predictor when compared to the largest Lyapunov exponent and the Kolmogorov entropy, as these methods has a tendency to give false positives in cases such as random series. The phase space diagram for the data is found using data series from which various methods like time delay can be used to reconstruct the attractor (Takens 1981; Sauer *et al.* 1991; Laio *et al.* 2003).

Chaos Theory based models have been extensively studied and researchers have shown that rainfall as well as runoff time series are chaotic and the types of nonlinearity and their intensity at different timescales have been studied (Li *et al.* 2014). In their study, the Pingshan hydrometric station was the area used and this study included the analysis of daily runoff time series and the measurement of chaos and nonlinear behavior at three different time scales in one day, 1/3 of a month, and a complete month). Li *et al.* (2014) used six methods of nonlinear dynamics namely, (1) reconstruction of phase space and the estimation of delay time using average mutual information; (2) the estimation of the dimension of sufficient embedding

using an algorithm (false nearest neighbor); (3) method of correlation dimension; (4) method of Lyapunov exponent; (5) 0–1 algorithm for chaos (test algorithm); and (6) the adaptive method of multistep Volterra. The presence of low-dimensional chaotic situation was found at many time scales in the run-off dynamics. Also, the nonlinear behavior intensity was found to be composed of only a limited fraction of time scales (Li *et al.* 2014). Chaos theory has also been used to predict stream-flow (Solomatine and Ostfeld 2008; Ghorbani *et al.* 2010; Sivakumar and Berndtsson 2010). A combination of chaos theory and nonlinear science was proposed to solve the problems of nonlinear complex systems by Song *et al.* (2014). A modified version of models based on chaos theory has also been used for forecasting a city's daily water demand (Bai *et al.* 2014). The multi-scale relevance vector regression (MSRVR) approach decomposes a time series into many different scales using the stationary wavelet transform. This method was used for decomposing the time series of daily water supplies into different scales.

In Bai *et al.* (2014) the relevance vector regression (RVR) method was used for the training of the machine-learning model and the input variables of the RVR model were determined by analyzing chaos features of the water supply series. This method used the wavelet coefficients at each stage and the optimal combination of parameters for the RVR model is found using an algorithm (adaptive chaos particle swarm optimization). A more precise forecast of the daily urban water demand can be done by the MSRVR method in terms of the normalized root-mean-square error, mean absolute percentage error and correlation coefficient (Bai *et al.* 2014). In the coupled chaos optimization-projection pursuit model, dimensions of the decision problem are reduced by adopting projection pursuit and the optimal projection direction is searched using a chaos optimization algorithm. Whether the model is rational and valid was verified by doing a case study in Fuhuan River for the initial water rights (Xiao *et al.* 2011).

In another study, the lake water level time series of three lakes in Sweden was shown to be chaotic based on phase space reconstruction using the KNN method (Tongal and Berndtsson 2014). Unlike the KNN methods, the functional relationship can be deduced in cases of Chaos Theory based models if the strange attractor is known and

thus extrapolation is possible in Chaos Theory based models. Moreover, the functional relationship can be estimated either at global or local level. In the case of global functional relationship, a unique predictor is assumed to exist that is believed to be valid for any of the reconstructed vectors. Instead, if it is assumed that if a definitive number of neighbors can be used to extract functional relationship, then such an approach is of a local level. In this sense, the local approach and KNN method are equivalent and can be used to find a relationship that would exist between strange attractor dimensions and the K of KNN. Other researchers have highlighted the limitations of Chaos Theory based models by pointing out that it is possible that hydrological data might not be lengthy enough or chaotic (Solomatine and Ostfeld 2008).

2.7 ARTIFICIAL NEURAL NETWORKS (ANN) BASED MODELS

Artificial neural network (ANN) based models are inspired by the structure of the human brain, which consists of interconnected neurons that are used for processing information. Similar to the brain, these models consist of interconnected nodes that are used for information processing (Elshorbagy *et al.* 2010). These nodes operate sequentially and the output of one layer is passed to other nodes in another layer in the network for further processing (Yegnanarayana 2009). During the learning phase, the threshold and the optimal weights are found. ANNs can be single-layered or multi-layered. The most common types of processing used in the neural network are feed-forward, self-organizing or recurrent. Commonly used ANNs include Radial Basis Function (RBF) and Multilayer Perceptron (MLP) (Hsu *et al.* 1995).

2.7.1 ANN model development

The various layers in a MLP are an input layer, one or more intermediate layers and an output layer. There may be one or more nodes or neurons in each layer. The nodes in one layer are connected to nodes of other layers but not with the nodes in the same layer. The primary processing unit of ANNs is a node, which generates an output function. The output function is generated using the activation function and the weights assigned to the various inputs (Maier and Dandy 2000). Different activation

functions like linear, threshold, Gaussian, logistic, sigmoid and hyperbolic tangent sigmoid are used depending on the training algorithm, data scaling used and the type of network. A neuron is activated if the input value is more than the threshold value. Upon activation, the inputs are scaled by their weight and the output is calculated from the weighted inputs. An optimum value of threshold and the input weights are obtained by training the ANNs using the training samples. The development of a model is a complex process involving various stages like data gathering, division of data into groups and data preprocessing. Then the model inputs are determined and an ANN is selected. Finally the ANN is trained and evaluated (Maier and Dandy 2000).

The input data sets for ANN simulations are normally collected from gauges placed on site or from remote sensing instruments. Proper functioning of ANNs requires that the inputs are sufficiently large and comprehensive enough to capture all the important characteristics of the relationship between inputs and outputs. Training data sets and validation data sets are created from the available data. When the available data is limited then data points are interchanged between validation and training sets to create sufficient data. If the amount of data is large then a test set is chosen from that data and is used to determine the optimum size of training data to be used for training. This approach prevents issues arising from data over-fitting (Maier and Dandy 2000). Another way to avoid over-fitting is by using cross validation (Oyebode *et al.* 2014). Preprocessing of input data is done using singular spectrum analysis, wavelet transform and discrete wavelet-transform (Sivakumar *et al.* 2001; Partal and Cigizoglu 2008; Kişi 2009). Before passing these inputs to the activation function, scaling is done to bring them within the range of the arguments of the activation functions. Scaling involves standardization and normalization. Inputs of the models are selected based on the *a priori* knowledge of the system, trial and error, correlation analysis, Principal Component Analysis (PCA), knowledge method or a combination of these (Bowden *et al.* 2005). If the number of inputs is large then PCA is used (Al-Alawi *et al.* 2008). Sensitivity analysis is used to identify the model characteristics from the trained ANN. If the relationship between inputs

and outputs is unknown then correlation techniques like cross-correlation, auto-correlation and partial auto-correlation are used.

In some cases nonlinear correlation techniques like partial mutual information and average mutual information are also used. Sometimes optimal global optimization methods like genetic algorithm are used for the identification of inputs (Piotrowski and Napiorkowski 2011; Dhamge *et al.* 2012). Another important area of interest is the architectural design, which deals with the geometry of the layer and finding the number of nodes in the intermediate layers. Various algorithms like constructive and stopped training, regularization and trial and error methods are used for this purpose (Sharma and Chandra 2010). Training of ANNs is required to bring the network output close to the target output. Local training algorithms like back propagation as well as global training algorithms like GP, Shuffled Complex Evolution, particle swarm optimization, conjugate gradient algorithm and Levenberg-Marquardt (LM) algorithm are used for training (Piotrowski and Napiorkowski 2011; Dhamge *et al.* 2012). The local algorithms may result in finding local minima instead of a global one but their advantage lies in the fact that they are less computationally intensive. Some newer approaches combine the local and global methods to harness the advantages of both (Muttil and Chau 2006; Wu 2010; Wu and Chau 2011).

Water resource modeling has used the ANNs based methods extensively (Govindaraju and Rao 2010), though researchers' opinions are mixed on their use. While reviewing the use of ANNs in river systems, Maier and Dandy (2000) observed that the methods used for determining inputs of the model, appropriate data subsets and best model structure are generally determined in an *ad-hoc* manner (Maier *et al.* 2010). Use of MLPs for model architecture and gradient-based methods for calibration of the models are the popular choices in river systems. ANN was used in the estimation of groundwater levels in coastal aquifer systems (Nourani *et al.* 2011; Taormina *et al.* 2012). It was observed that spatio-temporal water table predictions could be improved by using ANNs alone or in conjunction with other geo-statistical models. Other areas of application of ANNs in water resource modeling include finding a relationship between rainfall and river pollution (Mohsenifar *et al.* 2011), estimation of suspended sediment load (Melesse *et al.*

2011), and nitrogen content in streams (He *et al.* 2011), water quality measurements and improvement (Bieroza *et al.* 2011; Khalil *et al.* 2011), desalination (Khayet *et al.* 2011), salinity of groundwater (Banerjee *et al.* 2011) and water demand estimation (Odan and Reis 2012). ANNs do not assume a functional relationship between the inputs and outputs and yet they are capable of modeling complex processes. ANNs are evolving and self-organizing and the computational costs of ANN based models are also optimal. On the downside, these models are not easily accepted in the real world as they are somewhat opaque in their functioning. The extrapolation of these models is not possible as the models evolve using the training data and thus their parameters are dependent on their training. ANNs have many local minima and finding an optimal structure takes time and the solutions obtained depend on the training set. Lastly, these models function on their own in the sense that information available from the known physical laws cannot be incorporated in the model (Giustolisi and Laucelli 2005).

2.8 FUZZY RULE BASED SYSTEMS (FRBS)

Fuzzy Rule Based Systems (FRBS) describe the relationship between the inputs and outputs of a model using fuzzy if-then statements (Adriaenssens *et al.* 2004). These systems divide the input space into various overlapping fuzzy sets. This leads to the division of inputs into fuzzily defined intervals. This process of dividing the inputs into fuzzy sets is known as fuzzification. Other than these fuzzy sets, FRBS consist of membership functions, domain partitions and if-then inference rules. A membership function assigns membership values to the inputs of a fuzzy set. This membership value reflects how much the input belongs to a particular set. There are consequents given by if-then inferences for every membership set. The consequents can be fuzzy in which case they are known as the Mamdani model (Mamdani 1974) or linear model, in which case the model is a Takagi-Sugeno-Kang (TSK) model (Takagi and Sugeno 1985; Sugeno and Kang 1988).

2.8.1 FRBS model evaluation and applications

An operator is used in the inference process in the Mamdani model and the consequents give the likelihood of an input being mapped to the output sets. 'Minimum' or 'product' functions are used to find the intersection. 'Minimum' is used for independent input sets and 'product' is used if they are dependent. 'Sum' and 'maximum' operators are used to combine the results in the 'aggregation' phase. In the defuzzization stage, the fuzzy output is converted to a single value. Methods like bisector or center of gravity are used in the defuzzization stage. On the other hand, in the case of TSK systems, the rule consequents are explicit functions of input variables. Linear local models are used and the interactions between the rules result in the nonlinear response of the overall system. Defuzzization is not required in these systems and the output is obtained by aggregating the output of individual inputs.

The selection of input variables is of utmost importance. The number of rules rises exponentially with the increase in the number of input variables or membership functions. Grid partitioning or fuzzy clustering is used for partitioning the input space. In grid partitioning, the relationship between the input variables is not taken into account in defining the membership function of each input variable. Thus, the number of rules rises exponentially with an increase in the number of inputs in this case. Prior experience is then used for defining the membership functions. In fuzzy clustering, fuzzy clusters are used for calculating the antecedent parameters. It uses Gustafson Kessel clustering and subtractive clustering. Euclidean distance is used for defining the membership functions. The disadvantage of fuzzy clustering is that its performance is degraded in the presence of outliers. In both cases, the use of a particular operator influences the inference mechanisms.

FRBS have been used in water resource allocation (Lu *et al.* 2010; Sadegh and Kerachian 2011; Wang and Huang 2012), reservoir systems (Rani and Moreira 2010), water distribution (Chen and Chang 2010; Fu and Kapelan 2011) and in the selection of watershed plans (Chen *et al.* 2011). Variants of FRBS have been developed like the Fuzzy VIKTOR method that ranks and selects from a set of alternatives in the presence of conflicting criteria and then proposes one or more

compromising solutions. Weights, as well as the criteria, can be fuzzy in this model. This model has been used for finding the optimal size of the reservoir system for storage of surface flows of the Mlava river (Opricovic 2011). Trade-offs between economic efficiency and constraint-violation risk were found using an interactive two-stage stochastic fuzzy programming (ITSFP). ITSFP incorporates interactive fuzzy resolution (IFR) method within an inexact two-stage stochastic programming (ITSP) framework. It also found a compromise between the feasibility of the constraints and the satisfaction degree of the goal (Wang and Huang 2012). A functional relationship between the inputs and outputs emerges in the FRBS models and this permits the incorporation of subjective knowledge (Jacquin and Shamseldin 2006). In this sense they feel more 'real' as opposed to the ANN based models. The disadvantage of these models is that they are not successful if the functional relationship is complex or there are many input variables as the complexity of FRBS increases exponentially with the number of inputs (Li *et al.* 2010).

2.9 SUPPORT VECTOR MACHINES (SVM) BASED MODELS

SVMs focus on minimizing the bound on generalization error or what is known as structural risk minimization (SRM) which is different from the previous models which focus on minimizing the error between observed and predicted values or empirical risk minimization (Elshorbagy *et al.* 2010). SVM based models employ the concepts from statistical learning theory. In such models, a mapping function maps the inputs to the outputs and the parameters of this mapping function are discovered by using the training data (Jordaan 2002). SVMs can be thought of as ANNs with two layers in which the first layer has nonlinear weight and the second has linear weight - the differences resulting from the way in which they are trained. So, the methods of finding the model inputs and the data preprocessing techniques are the same in both cases (Oyebode 2014). Similar to the ANNs, *a priori* knowledge, sensitivity analysis, statistical analysis and trial and error are used for model identification. In addition, data preprocessing techniques like Single Spectrum Analysis (SSA), rescaling and decomposition are also used in SVMs and linear,

polynomial, sigmoid, radial basis function and hyperbolic tangent as the kernel functions.

The training stage is crucial and involves finding parameter values that result in good generalization or, in technical terms, the minimization of the Vapnik-Chervonenkis (VC) dimension. The optimization problem is solved using a dual set of Lagrange Multipliers (LM). The parameters of the first layer are chosen as the training input vectors for the second layer and then global evolutionary techniques are used for solving the optimization problem (Hearst *et al.* 1998; Liong and Sivapragasam 2002; Steinwart and Christmann 2008; Wu 2010; Kisi and Cimen 2011; Song *et al.* 2014). The optimal structure of SVM models is uniquely definite and the optimized LMs are used for finding the relevance and importance of various inputs.

Unlike ANNs in which the number of parameters increases exponentially if the inputs are multidimensional, the number of parameters does not increase in the case of SVMs. SVM based models can also be applied to non-Gaussian noise in data and are far more robust against outliers. Additionally, these are not as opaque in their functionality as the ANNs and some information about the relationship between the inputs and outputs can be obtained. The main limitation of these models is that their computational complexity increases exponentially with the size of the training data (Lin *et al.* 2006). So, essentially, the complexity shifts from the number of input parameters to the size of the training data. Many modern water resource modeling techniques use SVMs for their obvious advantages and these have been employed in forecasting floods (Liong *et al.* 2002), modeling runoff (Behzad *et al.* 2009), statistical downscaling of daily precipitation (Chen *et al.* 2010), measuring water quality variations (Nikoo and Mahjouri 2013), water quality (Singh *et al.* 2011), stream flow prediction (Noori *et al.* 2011) and the prediction of suspended sediment load (Kakaei Lafdani *et al.* 2013).

2.10 GENETIC PROGRAMMING (GP) MODEL DEVELOPMENT

In GP, the inputs are mapped to outputs using a randomly generated mapping function. This mapping function may undergo crossover, mutation, duplication, and

deletion during iteration. Parse trees are operated by approaches such as genetic programming instead of bit strings (Koza 1992) using methods like symbolic regression (Babovic 2000). The parse trees are constructed by making the first element and functions root nodes and interior nodes respectively. The variables and constants, which are operated upon by the functions, become the leaf nodes of those function nodes. The interior nodes can represent functions like arithmetic operators, logical operators, Boolean operators, loops or any other mathematical function. So, many trees of various shapes and lengths are created. Then the fitness of the trees to the objective function is evaluated and the programs are selected by looking at the ones which fit the data set best. These programs undergo crossover and mutations and the process is repeated many times in order to identify symbolic expression best fitting the data. These GP based approaches do not identify a model and no *a priori* knowledge is needed. These models optimize the structure of the model as well as its parameters. A structure relating the inputs and outputs is developed but no functional form is assumed.

2.10.1 GP model evaluation and applications

The disadvantage of genetic programming based models is that they have a limited applicability to find constants and complex results are produced by these models (Koza 1992; Wu 2010). GPs have been used for predicting the dispersion in streams (Azamathulla and Ghani 2011; Nasseri *et al.* 2011), evaluating the structure of hydrological models (Selle and Muttil 2011), modeling daily pan evaporation (Guven and Kişi 2011a), suspended sediment yield in natural rivers (Guven and Kişi 2011b), monthly groundwater levels (Fallah-Mehdipour *et al.* 2013a), constructing semi-variogram models (Sivapragasam *et al.* 2011), modelling pipe failures in water distribution systems (Xu *et al.* 2011), developing decision support system (DSS) (Zeng *et al.* 2012) and reservoir operational decision rules (Fallah-Mehdipour *et al.* 2012, 2013b).

2.11 COMPARISON OF ARTIFICIAL INTELLIGENCE (AI) MODELS

This section reviews the recent literature on the comparison of different Artificial Intelligence (AI) models used in water resource modeling. Firstly, the important characteristics of all the models are briefly stated. K-nearest Neighbors algorithm (KNN) is one of the simplest techniques and models based on KNNs are very robust against noise and irrelevant attributes in the data. Their disadvantage is that they are unsuitable for real time forecasting and cannot be extrapolated. Chaos Theory based models are accurate for long-term predictions but these models are applicable only if the data series is chaotic. Similar to KNNs, the extrapolation abilities of these models are also poor. ANN based models have the advantage that their training is quite fast and they map the complex relationships easily and are also good for long term forecasting. These models are also robust against incomplete data, noise and outliers. The disadvantage of these models is that they are not transparent and easily interpretable and are difficult to generalize. Further, these models cannot incorporate the knowledge about the systems within the model. FRBS can incorporate the structured knowledge within the model. These models are relatively transparent and provide some information about the rules used for mapping the inputs to outputs. These are very robust to noise and very useful if the accuracy of sensors is low. Limitations of these models include a slow convergence rate and an exponential increase in the number of rules with the increase in the number of input variables. SVMs are easy to generalize and there is no increase in the number of parameters on giving multidimensional inputs. These models perform well even for small data sets. Their main disadvantage is the requirement of a large computational capacity.

Further, there is an exponential increase in training time when increasing the number of samples. GPs have an easily understandable model structure but they are also computationally intensive and hardly provide any physical insights about the underlying relationships (Salas 1980; Singh 1995; Yegnanarayana 2009; Govindaraju and Rao 2010; Sivakumar and Berndtsson 2010; Wu 2010; Box *et al.* 2013; Poli and Koza 2014; Tayfur 2014). Each model reviewed in this study has its advantages and disadvantages and understanding the problems at hand and the goals to be achieved determines which model is best suited for the task. Moreover, the

models are complementary so a hybrid approach is often better than using a single model. The next section reviews some recent studies that shed light on the usefulness of these models in specific situations.

2.12 RECENT APPLICATIONS OF ARTIFICIAL INTELLIGENCE IN WATER RESOURCES MODELING

SVMs, nonparametric KNN, radial basis function NN were applied in the study of virtual water content of crops (Ahmadaali *et al.* 2013) and KNN was found to best serve the purpose. SVM performed better than the radial basis function Neutral Network (NN). In another study, SVM was found to the best alternative out of SVM, probabilistic NN and KNN for the classification of water quality. KNN was the worst performer with the maximum number as well as value of errors (Modaresi and Araghinejad 2014). GP and Adaptive Neuro-fuzzy Inference System (ANFIS) techniques were used for forecasting ground water levels by Shiri and Kişi (2011). Based on different combinations of depth values of the water table from two stations, 5 GP and ANFIS models were developed, and forecasts were made for one, two and three-days ahead for the water table depths. Root mean square errors (*RMSE*), scatter index (*SI*), variance account for (*VAF*) and coefficient of determination (R^2) statistics were used for measuring the performance. Both models successfully forecasted water table depth fluctuations but GP performed better than the other model, giving explicit expressions for the problem (Shiri and Kişi 2011). In yet another study, hybrid wavelet-artificial neural network (WANN) and linear GP (LGP) techniques were proposed for forecasting monthly stream flow. RMSE and Nash-Sutcliffe efficiency (NSE) measures were used for comparing the performance. The ANN method was used as the primary reference model to model six different monthly stream flow scenarios based on the records of two successive gauging stations and the main time series of input(s) and output records were decomposed into sub-time series using the wavelet transform. These sub-time series of each model were imposed to ANN to develop WANN models as optimized versions of the reference ANN models. LGP performed better than WANN in all the reference models (Danandeh Mehr *et al.* 2013). GP and ANN based methods were used as potential surrogate models for

coastal aquifer management to determine the optimal rate of groundwater extraction. A 3-D simulation model for coupled flow and transport simulation together with an optimization algorithm in a linked simulation-optimization framework was used for the comparison (Sreekanth and Datta 2011). GP again performed best against the other three models in estimating everyday suspended sediment load (Kisi *et al.* 2012).

2.13 RECENT APPROACHES AND FUTURE DIRECTIONS

Recently a combination of different AI models has been used for finding optimal solutions. For example, Nourani *et al.* (2011) used a wavelet analysis and genetic programming method to construct a hybrid model for optimizing rainfall-runoff time process modeling (Nourani *et al.* 2011). The hybrid model that linked wavelet analysis to GP used sensitivity analysis for identifying input variables of an artificial neural network rainfall-runoff model. Further, the time series of both the variables – rainfall and runoff, were decomposed into many multi-frequency time series using the wavelet transform and these were imposed to the GP as input data to optimize the structure of ANN modeling. The results could be compared favorably to both GP and ANN models. The same methodology also worked in predicting the suspended sediment load in rivers (Rajaee *et al.* 2010). The introduction of wavelet coefficient inputs from wavelet genetic programming and wavelet neuro-fuzzy, for forecasting daily precipitation on the basis of previously recorded results, resulted in an improvement of the forecasting results. Further enhancement of the model by merging the inputs from both (best single model and hybrid models) and using these as the model inputs, enabled the new hybrid wavelet GP models to successfully predict the daily precipitation, although the neuro-fuzzy models still performed badly (Kisi and Shiri 2011).

A combination of Extended Kalman Filter (EKF) with GP was used by Nasseri *et al.* (2011) to forecast the water demand in urbanized areas. The latent variables were inferred from the EKF. Five models were presented by them where they first used five-three lags of observed water requirement as the independent and most probable inputs. The effect of observation precision on water demand prediction was clearly

evident in their findings. A genetic algorithm combined to the linear programming method used for solving problems related to the designing of water distribution systems, where a genetic algorithm is used to decompose looped networks into a number of branched networks which was further optimized by the LP method, was also proposed (Cisty 2010). In another work, it was proposed to use adaptive insensitive factor for improving the performance of SVM. In this case, wavelet de-noise was used as a method to reduce/eliminate the noise in the time series of a streamflow and phase-space reconstruction theory to avoid the subjective arbitrariness of artificial judgment. The model was able to process data series of complex hydrological data in a better way and showed greater prediction accuracy and generalizability than the conventional methods of ANNs and SVMs (Guo *et al.* 2011).

2.14 SOFT COMPUTING AND ENGINEERING MODELLING

The desire to better understand life on earth has been in existence from the very beginning and this has been inspired by the efforts made by man to find ways to produce life-oriented behaviours (Adami 1998). The field of artificial life has, year on year, produced a series of new methodologies incorporating the use of biologically-inspired computation methods to solve problems. These 'bio-inspired' models are related to and include the study of probabilistic reasoning, machine learning, emergence of novelty, complex adaptive systems, social behaviour, intelligence, sustainability and survival (Banzhaf *et al.* 2013). The past two decades have witnessed great improvement in the application of artificial intelligence techniques to solve problems where conventional approaches have failed or performed poorly (Mellit and Kalogirou 2008). A data-driven approach which utilizes soft computing or Artificial Intelligence (AI), namely, Artificial Neural Network (ANN), Fuzzy Logic (FL), and Genetic Algorithms (GA) has been used by researchers to address some of these issues (Tokar and Johnson 1999; Vas 1999; Lin and Lewis 2003). AIs have been used successfully over the years in hydrology and water resources applications like sediment yield estimation models (Aytek and Kisi 2008), rainfall-runoff modelling (Smith and Eli 1995) and water quality prediction

models (Ömer Faruk 2010). In this study, emphasis is laid on past and recent applications of genetic programming a family of evolutionary algorithms which adopts a biologically-inspired computation model to solve problems.

2.14.1 History of evolutionary algorithms

Evolutionary algorithms belong to the category of evolutionary computation models which are based on the principle of natural evolution. The idea of using evolution to solve problems and to exhibit intelligent behaviours was initiated by Alan Turing (Castelfranchi 2013). In his essay entitled 'Computing Machinery and Intelligence' in 1950, he noted that digital computers would only be able to solve problems deterministically but would not be able to produce intelligent behaviour (Turing 1950). Holland (1962) noted that, "The study of adaptation involves the study of both the adaptive system and its environment." He pictured an adaptive system as a 'population of programs' and encouraged the idea of seeing adaptation from the viewpoint of population rather than individuals. He also noted strength in the evolved solutions when the population of solutions are applied and tested in parallel in a given problem (Holland 1962). Other major developments with respect to evolutionary algorithms include evolutionary programming (Fogel *et al.* 1966) and evolution strategies (Rechenberg 1994). For a detailed review on early works on evolutionary computing, readers are referred to Fogel (1998). John Holland is renowned for genetic algorithms found in his 1962 paper where his discussion centred on computer programmes.

However, the idea of evolving computer code or programmes through random mutation and recombination to produce better programs was not yet fully understood. Thus the idea was not accepted immediately as it was believed to be difficult to improve computer code when recruited randomly during evolution (Banzhaf *et al.* 2013). Smith (1980), suggested a variable-length program representation which allows the search space to link rules into rule-based programs that can be used to solve problems when defined by a fitness function. Forsyth (1981), developed a logical rule system with parameters where the programs would logically and numerically evaluate data samples by training the programs. By 1980 the idea of

early Genetic Programming (GP) was fully introduced. Two evolutionary programming systems were developed using different simple languages (Cramer 1985). The standard programming language, list processing (LISP), was also applied by Hicklin (1986) and Fujiko and Dickinson (1987) to particular applications such as automatic program generation and solving prisoners' dilemmas. In 1989 Koza documented a method that uses a universal language that could be applied to different problems. GP was also documented officially when he published a book in 1992, in which he recognized the power and generality of this method when applied to different areas of life (Koza 1992).

2.15 OVERVIEW OF GENETIC PROGRAMMING (GP)

2.15.1 Introduction

Genetic programming (GP), (Koza 1992), which derives from genetic algorithms is a systematic, domain-independent method that generates computer programs to solve problems automatically giving it a high level of what is expected from it (Poli *et al.* 2008). GP involves a repeated random search for solutions from an existing pool of computer programs – which are potential solutions – by applying the principle of natural evolution such as crossover and mutation to form a new population. This process continues until the best solution is obtained. These programs are expressed in the form of a syntax tree where the nodes represent the instructions – called the functions, and the leaves, which are the terminals, represent the independent variables and random constants. Five preliminary steps are necessary before the operation of GP. These include the determination of (i) the terminal set; (ii) the functional set; (iii) the fitness measure; (iv) the parameters for controlling the run; and (v) the termination criterion and method of designating the result of the run (Burke and Kendall 2005). The steps involved in the implementation of GP are explained in detail in the subsequent sections.

According to Banzhaf *et al.* (1998) the major advantages of genetic programming over other soft computing techniques includes the following: it is used when there is a large amount of data in computer readable form that needs to be examined, classified and integrated; it is used in situations where small performance

improvements are easily and routinely measured; when the interrelationships between the variables are poorly understood; when limited dataset is available; when the ultimate solution to the problem is difficult to find; and finally, when conventional mathematical models cannot provide the required analytical solution.

2.15.2 Initialization of genetic programming

As mentioned earlier GP programs are expressed in a syntax tree form (see Figure 6) made up of constants, variables and arithmetic operators. The constants and variables are called the terminal and they are similar to the leaves of a tree while the arithmetic operators are called functions and are considered as the internal nodes of the tree. Table 3 shows the various forms of function that could be used in GP. The allowable set of these parameters in a GP system is the primitive set of the system.

In the GP program illustrated in Figure 6, there are two functions (+ and *) and three terminals (p, v, and z). The selection of this primitive set is a function of the user and this will affect the performance of the developed models. This was proven in this study.

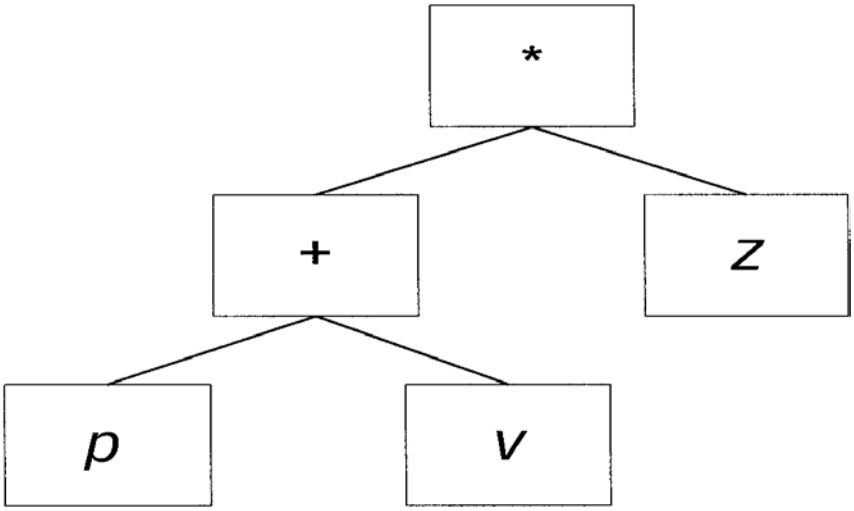


Figure 6: GP syntax tree

(Source: Babovic and Keijzer 2000)

Table 3: Examples of functions and terminal sets used in a GP programme

Functional sets	
Functions	**Examples**
arithmetic operations	+, −, ∗, ÷
Standard programming operations	conditional, iterative
Standard mathematical functions	sin, cos, exp, log
Boolean iterations	AND, OR, NOT
Looping	For, repeat
Terminal set	
Types	**Examples**
Constants values	2, 0.25
Variables	x, y
O-arity funtions	go_left, rand

Source: Poli *et al.* 2008

The formation of the preliminary random population is formulated by randomly searching blindly through the GP space for solutions (programs) (Harris *et al.* 2003). According to John Koza, there are three methods of generating the initial population in GP. These methods are the grow method, full method and ramped-half-and-half method. In the grow method, the programs in the population are created individually with different depths thereby producing different structures of programs. This method usually produces programs with one (terminal) node while the programs (trees) produced in the full method are full and of specified depths (Luke and Panait 2001) and they have more than one node. The ramped half-and-half method is the most preferred method (Walker 2001). It is a combination of grow and full methods. Here the produced programs have different sizes and shapes forming a good population. However, according to Poli *et al.* (2008), this method is easy to use but it makes the control of some statistical properties like sizes and shapes difficult.

2.15.3 Fitness evaluation and selection

Once the initial population is developed for a specific problem, the programs (solutions) in the search space (population) are evaluated in a similar manner to the use of Evolutionary algorithms in optimisation (EA). An ‘objective function’ is applied on individual programs to evaluate the difference in error between the

observed variable in the problem and the predicted variable from individual programs. According to Babovic and Keijzer (2000) the measured error can be measured in terms of its mean square error (MSE), root mean-square error (RMSE), etc. or in terms of the numbers of correctly predicted points (hits) tolerated within an accuracy interval. The better programs which are believed to have more offspring are then selected to form a new generation. The population then undergoes a selection process to make sure the best solution is still within the search space and also to make sure the search space only contains best probable solutions.

The most commonly used selection processes, according to Poli *et al.* (2008), are the fitness proportional selection, the truncation selection and the tournament selection. In the fitness proportional selection, the programs are assigned a probability value based on their 'objective function' value. The next generations of programs are based on these probabilities. In the truncation selection process, only the best programs are kept for the next generation. The tournament selection process is the most widely used selection process in GP (Oyebode 2014) and it involves the random comparison of best programs from the search space which then form the next generation of programs. This process rescales fitness because it only compares programs and not the magnitude of their difference. This gives an opportunity of selection to average programs preventing the loss of diversity (Poli *et al.* 2008).

2.15.4 Genetic operators

The selected best programs based on their fitness functions are transformed to a new generation of programs mostly by the use of two operators, namely, crossover and mutation. The initial generation of programs is destroyed once an equal number of programs are produced by these operators. Crossover is similar to the reproduction process that occurs in living organisms. This involves two parent programs randomly interchanging their corresponding subtrees to produce two new offspring (solutions) of higher accuracy for the new generation. This process is illustrated in Figure 7 and it enables the evolutionary process to tend towards a promising solution and is therefore applied to 90% of the population (Walker 2001).

According to Fogel (1992) mutation can be described as any manipulation of an object. While in GP, the mutation process is the process of randomly replacing the sub-tree of a program with the sub-tree of another program to generate a new offspring (program). This process is demonstrated in Figure 8.

Generally, the genetic operators are applied on consecutive generations until a symbolic expression describing the data and representing the ultimate solution is established in either a mathematical equation form or computer program form (Deo 2009). Before this can be achieved however, the GP system must be fed with selected input variables and a number of control parameters must be set as mentioned in Section 2.15.4 above. Some of these control parameters include: (1) Population size, (2) Maximum number of generations, (3) Probability of crossover (usually between 0% and 90%) and (4) Probability of mutation (Walker 2001).

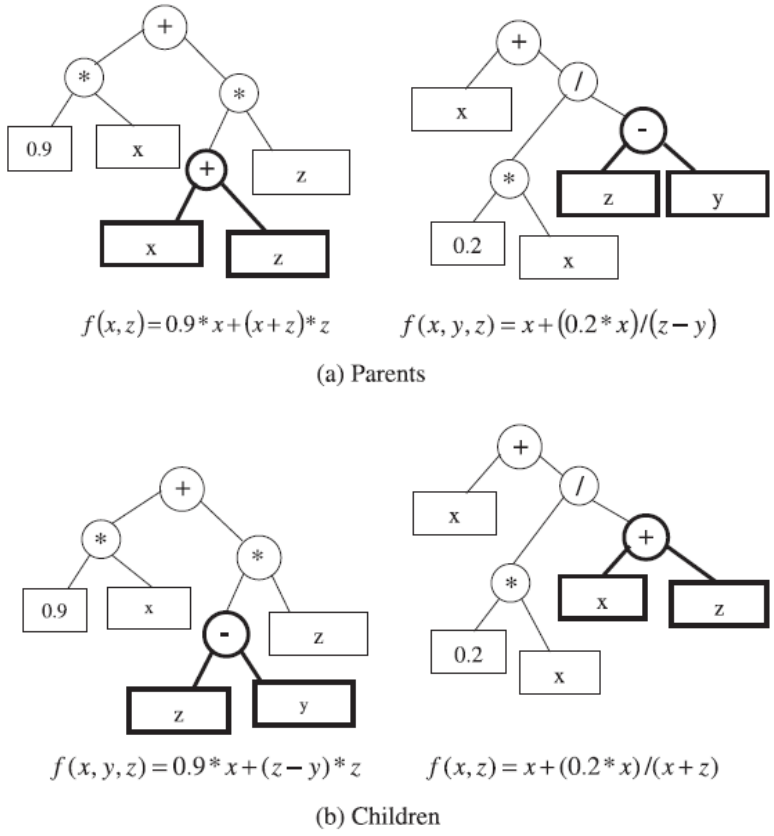


Figure 7: Illustration of the crossover process in GP

(Source: Selle and Muttil 2011)

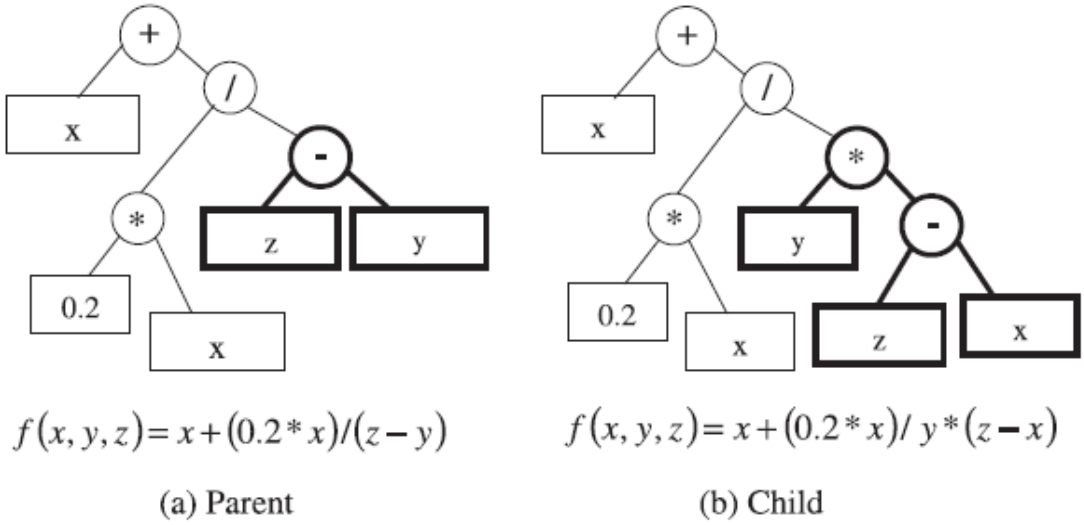


Figure 8: Illustration of the mutation process in GP

(Source: Selle and Muttil 2011)

2.15.5 Variants of genetic programming.

There are two extensions or variants of GP that are widely used in solving complex problems. They include Gene Expression Programming (GEP) and Linear Genetic Programming (LGP). GEP is similar to GP as it also evolves computer programs. These programs are encrypted in the form of fixed length linear strings (chromosomes) and are later transformed to nonlinear individuals of different sizes and shapes in the form of an expression tree. According to Ferreira (2001) , the advantages of GEP includes (i) the chromosomes are smaller and easy to influence genetically through replicating, mutating, recombining etc.; (ii) the expression trees are solely made up of individual expressions of each chromosome. LGP is a variant of GP that expresses its programs in a linear form, using imperative language such as C or C++ instead of LISP as the functional programming language (Ozgur and Jalal 2012).

2.16 GENERAL APPLICATIONS OF GENETIC PROGRAMMING

GP is a robust and dynamic model. It has been widely applied to solve different kinds of difficult problems (Chiang and Chang 2009). GP has the capability to select the best input from its variables making it possible for the input and output variables to be expressed as a regression equation. GP has been applied to all aspects of life such as in the field of water resources engineering, photogrammetry , medicine , biology , electrical engineering, science, civil engineering, industrial engineering, electrical power and mechanical engineering (Zhang *et al.* 2005; Mitra *et al.* 2006; Nunkesser *et al.* 2007; Makkeasorn *et al.* 2009).

The review of the applications of GP in all these fields is beyond the scope of this study but a brief overview of the application of GP in selected fields is provided, with emphasis on sediment modelling. In the field of water resources management, Savic *et al.* (1999) applied GP to develop a rainfall-runoff model for predicting runoff. A new formulation for bed concentration of suspended sediment was expressed by converting suspended sediment data into an equation for a better understanding of its generation process using GP techniques and the experimental flume data utilized by Zyserman and Fredsøe (1994) by mining data from sediment transportation near a riverbed. The results were compared with those from human experts and it was found to be very promising for mining of knowledge acquired data. GP was used to formulate GP sedimentary particle settling velocity equations by Babovic *et al.* (2001). GP was applied by Liong *et al.* (2002) to predict rainfall-runoff in a catchment area in Singapore. The intensities and durations of six different storms were used to train and test the model. From the results obtained in the study a consistent relationship between rainfall and runoff was identified. This implies that the application of the GP technique to forecast rainfall-runoff is a better alternative to other traditional models.

The prediction of velocity of flow on wetlands and vegetated areas using the GP technique was explored by Harris *et al.* (2003). They discovered a symbolic expression from laboratory data that showed a better understanding of the effect of vegetation on velocity and discovery processes. GP and Artificial Neural Network

(ANN) techniques were used to predict and model the rainfall-runoff relationship of a typical urban basin by Dorado *et al.* (2003). Sivapragasam *et al.* (2008), examined the relationship between storage and discharge in the Walla Walla river in the United State of America (USA). They discovered that this relationship is insufficient for routing flood hydrographs on natural channels. Therefore a GP model was developed for routing flood hydrographs. The developed model was very effective for routing complex flood hydrographs and it was able to express the route in a simple and mathematical expression. Giustolisi (2004) used GP to determine the coefficient of Chezy resistance in corrugated channels by using three corrugated plastic pipes to measure hydraulic parameters. His work produced two GP equations for Chezy resistance coefficients which represent the experiment data. Johari *et al.* (2006) used GP to predict soil characteristic curve by conducting pressure plate tests on silty clay, clay, loam and sandy loam using Soil Vision software. The test results were used for training and testing the GP model and the resultant model was compared with experimental results and other models, and was found to be superior to them.

Rabunal *et al.* (2007) used GP to predict the unit hydrograph of a typical urban basin in conjunction with ANN. The two models were combined to establish an accurate relationship between rainfall and runoff in that basin. GP was also used to predict short-term and long-term river flows and the result was found to be more accurate compared with that from ANFIS techniques (Kisi and Shiri 2010). In another study, Shiri and Kişi (2011) predicted groundwater table depth fluctuations using GEP and ANFIS. The results showed that both models can be used to successively predict the fluctuation but that GEP models were found to be more accurate. Shiri *et al.* (2012) used limited climatic variables to model daily reference evapotranspiration, and also found that GEP performed better than ANFIS. In all the situations where the GP technique was applied, it proved itself to be accurate and superior to other techniques.

2.16.1 Sediment modelling using GP approach.

The GP approach has also been used successfully and intensively as a hydrological modelling tool especially for estimating sediment yield. Kizhisseri *et al.* (2005)

employed GP to develop a better sediment-temporal pattern of fluid field relationship, using numerical model results and Sandy Duck field data. The data sets of suspended sediment and discharge from two stations on Tongue River in Montana were used by Aytek and Kisi (2008) to develop an explicit relationship between daily suspended sediment and discharge using GP. Their results suggested that GP is a better technique than the sediment rating curve and multi-linear regression techniques and that GP is more practicable to use.

Garg (2011) explored the ability of GP to estimate sediment yield in the Arno River basin in Italy which is susceptible to flooding. Five variables – river length, drainage density, yearly average rainfall, erodible area and watershed area, were used as input variables in this study and the results showed that GP is an efficient and reliable technique for estimating sediment yield even when the data set is limited. In a study carried out by Ozgur and Jalal (2012), three soft computing techniques namely, Gene Expression Programming (GEP), Adaptive Neuro-Fuzzy Inference System (ANFIS) and Artificial Neural Networks (ANNs) were used to estimate daily suspended sediment load in the Eel River near Dos Rios, in California, USA. Suspended sediment load, stream flow and daily rainfall data were used as input for developing the models. The results, when compared, show that the GEP model is superior to the other developed models in predicting daily suspended load in the river. LGP, GEP and ANN techniques were used in the estimation of daily suspended sediment in the Tongue River in Montana, USA. Discharge and suspended sediment data from two stations on the river were used as inputs. GEP performed better than ANN but LGP models were found to be superior to the GEP models (Guven and Kişi 2011b).

GP model trees, (MT) and ANN, which are data driven models, were used for the estimation of the quantity of sediment deposited in Gobindsagar reservoir (Garg and Jothiprakash 2013). It was found that both GP and artificial neural networks (ANN), which are nonlinear models, captured the trend of sediment deposition into the reservoir better than linear model trees (MT) (Garg and Jothiprakash 2013). GP models were also compared with support vector machine (SVM), Adaptive Neuro-Fuzzy Inference System (ANFIS) and ANN models by Kisi *et al.* (2012). They used daily discharge and sediment yield data from1972 to 1989 from two stations on the

Cumberland River in the United States of America to test and train the models. The predicted outputs from the developed GP models were compared with those from the support vector machine (SVM), Adaptive Neuro-Fuzzy Inference System (ANFIS) and ANN models. The results showed that the GP models are superior to the other three models in predicting sediment yield. The data set of discharge and suspended sediment yield from Rio Valenciano and Quebrada Blanca Stations operated by the US Geological Survey (USGS) were also used as training and testing data by Kisi and Guven (2010). They compared the performance of the developed LGP model, which is an extension of GP, with those of ANFIS and ANN models using standard model evaluation criteria. Furthermore, it was discovered that the LGP model is superior and more accurate than both the ANFIS and ANN models.

Genetic programming models have been found to exhibit exceptional performance when used as regression models in the majority of the case studies mentioned, especially for pattern recognition and complex non-linear estimations. It was also found that GP is less prone to over-fitting during training with observation datasets (Guven and Kişi 2011b). In all, the application of GP may serve as a decisive factor in the planning, construction, operation, management and maintenance of water resources projects.

2.17 APPLICATIONS OF GENETIC PROGRAMMING IN OTHER ENGINEERING APPLICATIONS

GP is also widely used in other fields of Engineering. Applications in engineering include areas such as system modelling, control, optimisation and scheduling, design and signal processing. GP has been applied to solve problems relating to systems modelling. Oakley (1994), applied Koza's LISP code to predict time series of chaotic systems. In the area of control, it was also used to develop hierarchical fuzzy logic controllers by Chaker and Hampel (1996). For optimisation and scheduling, GP was used by Grimes (1995) to prolong the lifespan of a railway track by developing the maintenance plan for the track. GP is also used in design. Porter *et al.* (1996), used GP techniques to design new polymeric materials. In order for GP to predict the desired polymer properties, the structural optimisation of monomers was carried out.

GP has also been applied to signal processing, for instance, the parameters and structure of adaptive digital signal processing (DSP) algorithms were developed using the GP technique by Alcázar and Sharman (1996).

2.17.1 Genetic programming in photogrammetry

Ice-flow ridges were detected by Daida *et al.* (1996) using the GP technique, from Synthetic Aperture Radar (SAR) imagery. A two-stage GP technique was used by Howard *et al.* (1999) to identify ships automatically from SAR images taken by the European Remote Sensing (ERS) satellite. GP was also used by Rauss *et al.* (2000) to categorize hyperspectral imagery. To cope with hyperspectral image classification, the GENetic Imagery Exploitation (GENIE) system was developed by Harvey *et al.* (2002), using a combination of conventional classifier algorithms and LGP. To properly differentiate minerals such as alunite, buddingtonite and kaolinite from hyperspectral images, mathematical and Boolean expressions was formulated using the GP technique by Ross *et al.* (2005). To aid in the classification of riparian zones, a two-stage GP technique was used by Makkeasorn *et al.* (2009) to find the best vegetation index, taking into consideration soil moisture variation in a semi-arid landscape.

2.17.2 Genetic programming in medicine, biology and bioinformatics

GP was applied in the field of medicine and biology by Handley and Forrest (1993) for the automatic learning of a detector for alpha-helices in protein. Koza and Andre (1996), used GP to forecast the properties and behaviour of proteins and biological systems. A medical doctor called Howard Oakley, who has a great interest in frost bite, used the GP technique to predict blood flow in toes (Kinnear *et al.* 1999). Brain tumour samples containing nuclear magnetic resonance (NMR) spectral data were examined using the GP technique by Gray *et al.* (1998). GP technique was also used to detect a group of protein clusters by the analyses of mass spectrometry data. This helped to detect leptomeningeal metastasis in 106 breast cancer patients (Dekker *et al.* (2005). In another breast cancer study (Nunkesser *et al.* 2007), 63 single

nucleotide polymorphis's (SNP) were analysed using GP techniques resulting in the formulation of a Boolean rule as a result of the interaction between two SNP.

2.18 EVALUATIONS OF GP THEORIES AND PRINCIPLES

2.18.1 Benchmarking of GP

Irrespective of the wide acceptance and application of GP in solving complex problems and its application in diverse areas of life, there appears to be an eminent need to critically evaluate its theories and principles. According to McDermott *et al.* (2012) comparison between studies is very difficult because of the absence of standardization among the studies. They claim the development of standard benchmarks is a very important step necessary for the maturation of the GP field. They noted that most of the studies reviewed applied GP to solve domain-specific and nontrivial problems. They further stated that very simple benchmarks were used for analysis and comparison studies. This is detrimental to the advancement of the GP approach. O'Neill *et al.* (2010), also stated that the development of a good benchmark suite is an issue to be considered in the next 10 years of GP. According to McDermott *et al.* (2012), a good benchmark should have the following qualities: relevance; speed; be varied and accommodating to implementers; representation-independent; easy to interpret and compare; current; and precisely defined. Examples of current benchmarks include: predictive modelling; classification; binary functions; and symbolic regression (McDermott *et al.* 2012). It was further suggested by McDermott *et al.* (2012) that a candidate benchmark suite needs to be deliberated upon by the GP community.

Another issue that needs serious attention is the generalization capability of GP solutions – GP should be able to produce the same generalization performance from training data set for unseen data. This ability is affected by bloat and over-fitting. Naik and Dabhi (2013) reviewed, surveyed and classified the various methods used in controlling bloat. Four bloat control techniques, namely, double tournament method; lexicographic parsimony pressure with ratio bucketing; lexicographic parsimony pressure with direct bucketing; and tarpeian method, were applied on six

different problems and the outcomes were analysed against each other. Based on this, tarpeian method and double tournament method were combined and used on the six problems. The study stated that the combination of these two methods performed better than the individual methods, except on a multi-valued regression problem without a constant. It is believed that GP has a simple algorithm but the process of obtaining a sound theoretical model and precise mathematical results has been difficult to obtain spanning many years after the origin of the GP technique (Poli 2001). According to Poli *et al.* (2010), the delay for this was as a result of the different versions of GP requiring different theoretical models. Also the different representations of GP such as tree-based, graph-based and linear differ in dynamics and require different theoretical tools. Theoreticians are facing a lot of challenges due to the non-linearities, randomness, and numerous degrees of freedom present in a typical GP system as well as in Grammatical Evolution (GE) (Ryan *et al.* 1998), Evolution Programming (EP) (Fogel *et al.* 1966) and Cartesian GP (CGP) (Miller 1999). Poli *et al.* (2010), also highlighted some fundamental questions that need the attention of the GP community. Questions such as: What goes wrong when GP cannot solve a problem? Is the biasness of its genetic operators fully understood? Can the properties of both GP systems and GP problems be expressed in a common language? How should current GP theory be adapted to suit dynamic environments? These questions and many more need to be addressed and are meant to stimulate researchers into improving the theories behind GP.

2.18.2 Open issues in genetic programming

Some very important open issues in GP were discussed in O'Neill *et al.* (2010) which started at a panel discussion at the EuroGP series of conferences which began in 1998. Some of the issues include: representing GP appropriately; determining the level of difficulty of a problem for GP; comparing the performance of GP in both static and dynamic environments; determining the level of natural evolution. For complete and comprehensive details of these open issues readers are referred to O'Neill *et al.* (2010). These issues are meant to help researchers improve the techniques of GP and to stimulate future research to provide greater knowledge and strengthen the GP algorithm.

2.18.3 Conclusions – Advancement of GP

The applications of artificial intelligence models and techniques to real life situations have brought greater understanding to evolutionary computing. As extensively highlighted in section 2.16 and 2.17, GP approaches have been used sucessfully to provide solutions to many complex problems in the fields of hydrology, medicine, biology, photogrammetry, telecommunications and network, finance, gaming and engineering. Certain areas of the application of GP were also highlighted, especially the applications of GP to sediment yield modeling. The community of GP operators and researchers needs to be optimistic because as this approach matures, important advances in the theory of GP are being achieved. Although progress has been made and opportunites have opened up in this field over the past decades, there are still some open issues that need the attention of the GP communities. Although the review in this study may not be exhaustive, it is meant to give GP theoreticians and the general community of GP sufficient research direction and open issues needing attention going forward. These open issues and the sucesses made in the application of GP to solve complex and nonlinear problems should be discussed in the literature so that it can spread into the science and research communities. This will generate more intellectual and productive discussion that will eventually lead to the advancement of GP technique.

2.19 SEDIMENT RATING CURVE

The continuous flow of water in any river system similar to the Umgeni river system, constantly carries erosional substances into the system and utimately into oceans at lower sea level (Kondolf 1997). The need to meet the demands of an increasing population has lead to the construction of dams and the rechanneling of watercourses that has affected the natural flow of water (Syvitski *et al.* 2005). The construction of dams has interrupted the quantities of sediment transported in rivers causing changes to the sediment management and flow (Hu *et al.* 2011). Therefore, the importance of quantifying the amount of sediment flowing into a dam cannot be over emphasized. Variables like soil moisture content, soil properties, land use patterns, rainfall, etc, are used to develop models for predicting sediment load. Even though there are

sophisticated models that use both hydrometeorological and hydrological variables to predict sediment load, it is preferable and more economical to develop such models using sediment and streamflow records (Kişi 2007). According to Quilbé *et al.* (2006) the choice of using sediment load and streamflow records in a rating curve should depend on the correlations between the variables. They suggest that a sediment rating curve should be used if the coefficient of determination (r^2) is greater than 0.5. However, Harrington and Harrington (2013) futher suggest that the choice of a rating curve should not be based only on a goodness-of-fit indicator but the loads predicted from rating curves should be compared.

2.19.1 Representation of the sediment rating curve

The sediment rating curve can be defined as the statistical relationship between sediment load (Qs) or suspended sediment concentration (S) and stream discharge (Q) (Fan *et al.* 2012). According to Syvitski *et al.* (2000) this relationship is usually expressed as a power or logarithmic function:

$$C = aQ^b \quad (1)$$

$$\text{Or } \log C = \log a + b \log Q \quad (2)$$

The sediment rating curve is a 'black box' type of model and its coefficient a and b have no meaning physically (Asselman 2000) but Morgan (2009) stated that a-coefficient is an indicator of erosion severity and a high value representing easily transported and intensively weathered materials. He also stated that the b-coefficient represented the erosive power of a river, with high values indicating an increase in erosive power due to a small increase in the river's discharge. The relationship between these variables (a and b) and some river characteristics like grain size of sediment, river channel morphology, erodibility and the streampower of the river basin were also examined by other researchers (Morehead *et al.* 2003; Yang *et al.* 2007; Wang *et al.* 2008).

2.19.2 Application of sediment rating curve in sediment yield modeling

The use of a rating curve is common among engineers and scientists for various scientific and engineering purposes (Wang *et al.* 2008). Sedimenet rating curves are

especially used by engineers and hydrologists to estimate the life expentancy of a dam while scientists use it to study the depositional and erosional environments (Syvitski *et al.* 2000). It can also be used to predict past records of sediment transported in a river that does not have such records (Asselman 2000). It is also useful for the calculation of sediment deficits during heavy floods (Xu *et al.* 2005). In China, Wang *et al.* (2008) used Hankou station's monthly suspended sediment and streamflow records from 1954 to 2005 to develop a sediment rating curve to determine its parameters for different periods and the effect of human activity on those sediment rating parameters was also examined. It was also used by Yang *et al.* (2002) to predict sediment rating parameters at Changjiang's hydrological stations in China. They used daily suspended sediment yield and water discharge records from the 1950s to the 1980s and also looked at the effect of the predicted sediment rating parameters on human activities and hydrological processes.

2.19.3 Areas of concern

The predicted suspended sediment load from sediment rating curve techniques seems to be either under-estimatedf or over-estimated when compared with the corresponding observed suspended sediment load. This may be due to the conversion of its results from log-space to arithmetic-space and also due to the scatter-nature of its regression line (Walling and Webb 1988; Asselman 2000). More studies on sediment rating have been carried out globally on small river basins compared to large rivers. Large river basins may have different geologic or climatic characteristics which may complicate the relationship between suspended sediment load and streamflow in those river (Hudson 2003).

2.19.4 Performance improvement

Various forms of modification have been developed to compensate for the deficiences in the sediment rating curve technique in predicting suspended sediment load. Some of the modifications which have brought improvement includes using non-linear regression equations, dividing its data into hydrological or seasonal groups and using correction factors (Phillips *et al.* 1999; Holtschlag 2001).

Holtschlag (2001) also suggested that with or without ‘smearing’ correction very good results with less than 15% error could be achieved by using either linear or second-order polynomial sediment rating curves.

2.19.5 Advantages and disadvantages

The use of a suspended sediment rating curve is relatively less expensive than the use of turbidometers or automatic samplers for the long-time frame measurement of sediment. It is also relatively more accurate in the prediction of long-term suspended sediment load, which would complement the scarce historical sediment database. The sediment rating curve could be linear, concave or convex giving an indication of the hydrological processes (Horowitz 2003).

2.20 CONCLUSION

Artificial Intelligence data driven models are very helpful in situations where the functional relationship between the inputs and outputs is either unknown or there is uncertainty in the measurement. Futuristic approaches are now combining various types of data models and use various pre-processing methodologies for enhancing the efficiency, accuracy and effectiveness of the models for long term prediction. Various studies indicate that the use of wavelet transform with SVMs and GPs with other models can further improve the performance of DDM AI models. The last few years have seen the emergence of Dynamic Data Driven Application Systems (DDDAS) capable of integrating the observations from real-time sensors on different temporal and spatial scales (Song *et al.* 2014). These systems are highly efficient due to their ability to properly integrate data measurement and assimilation and enable model selection using dynamic optimization for the control and steering of the water systems. Another direction in which present researchers are moving is multi-model simulations to determine how to accurately acquire real-time data at required times and their integration with Geographical Information System (GIS). Efforts are ongoing to incorporate the user-generated content in the futuristic models to improve the accuracy and efficiency of water resource modeling systems.

CHAPTER 3

GENETIC PROGRAMMING FOR PREDICTING SUSPENDED SEDIMENT

3.1 INTRODUCTION

Reservoirs are very important for the existence of mankind. Many other functions they perform are to supply water to industries, municipalities, for irrigation purposes and also for hydropower generation. However, the presence of reservoirs may interfere with the natural flow of water in a river thereby disturbing the aquatic life, the quality of water, flooding and sediment accumulation. The increase in population in the catchment where a reservoir is located brings about an increase in human activities which greatly increases the quantity of sediment flowing into the reservoir. Therefore, the determinations of the volume, rate and the pattern of sediment flowing into a reservoir have been considered to be very important to the operation and management thereof (Merritt *et al.* 2003; De Vente and Poesen 2005).

The process of reservoir sedimentation is a complex hydrological process which is influenced by factors like precipitation, wind, temperature, atmospheric particle count, humidity, atmospheric pressure and human activity (Bhattacharya and Solomatine 2006). Therefore, the regular use of hydrographic surveys and the accurate prediction of sediment yield is very useful in the maintenance of reservoirs including the planning for dredging activities (Bhattacharya and Solomatine 2006). The use of Data Driven Models (DDM) such as genetic programming (GP), Artificial neural network (ANN), Model trees (MT), K-nearest Neighbour (KNN) and Support vector machines (SVM) have been used for more than two decades for the accurate prediction of processes where there is insufficient knowledge about the physics of the process (Londhe and Charhate 2010) and also for the estimation of sediment yield taking into consideration the parameters that influence it (Bhattacharya *et al.* 2005; Elshorbagy *et al.* 2010; Garg and Jothiprakash 2013). A detailed description of

some data driven models and their application in water resources was highlighted in Chapter 2 of this thesis.

GP, a data driven model, has been successfully and widely used in water resources modelling (Kizhisseri *et al.* 2005; Aytek and Kisi 2008; Aytac and Kisi 2010; Guven and Kişi 2011b; Garg *et al.* 2014; Petke *et al.* 2014; Westerberg and Levine 2014; Zhao *et al.* 2014) hence it is being used in this study to estimate the quantity of suspended sediment flowing into Inanda Dam, KwaZulu-Natal, South Africa. Twelve models, one for each month of the year were developed using monthly in-streamflow datasets as inputs. The developed monthly models were compared among themselves using standard models evaluation criteria which are coefficient of determination (R^2) and root mean square error (RMSE). In chapter 4 of this thesis, they are also compared with the corresponding developed monthly SRC models.

3.2 METHODOLOGY

3.2.1 GPdotNET version 4.0 software

An evolutionary algorithm which is GPdotNET was used to determine the relationship between suspended sediment and streamflow flowing into Inanda Dam in the form of symbolic expression. The software is a free, artificial intelligence tool developed from Microsoft.NET technologies in 2009. It is a tree based algorithm which can be applied in Genetic Programming, Genetic Algorithm, Time Series modelling and in other forms of modelling and optimization for the graphical and visual presentation of outputs. The software supports multi-core parallel processing based on the ParallelFx library and can also handle large quantities of datasets (Wijayaweera and Karunananda 2012). It can run on both Linux and Windows based operating systems and is written in C programming language. The input datasets are usually stored in CSV format. A full description of the software and a quick tour of its operation are available in the literature (Hrnjica 2015).

3.3 STUDY AREA AND DATASETS

The Inanda Dam is located on the Umgeni River near Hillcrest, KwaZulu-Natal, South Africa. The dam is fed by water from Mooi Midmar Dam, Albert Dam, Nagle Dam, uMizimduzi River and the Umgeni River. The detailed description and purpose of the dam and its suitability for this study is highlighted in section 1.6 of this thesis.

The datasets used in this study were collected from Umgeni Water and the Department of Water Affairs (DWA). The Department of Water Affairs (DWA) supplied the historical up-streamflow dataset from 1990 to 2013 from the gauging station USH055 with geographical coordinates $29^0$42’55.42” S, $30^0$52’8.06’’ E while the suspended sediment dataset was supplied by Umgeni Water from the gauging station RMG 022 (Umgeni weir down stream of Inanda) USH055 with geographical coordinates 29°42'56.1" S, 30°52'09.7" E. The locations of these stations are indicated in Figures 3 and 9.

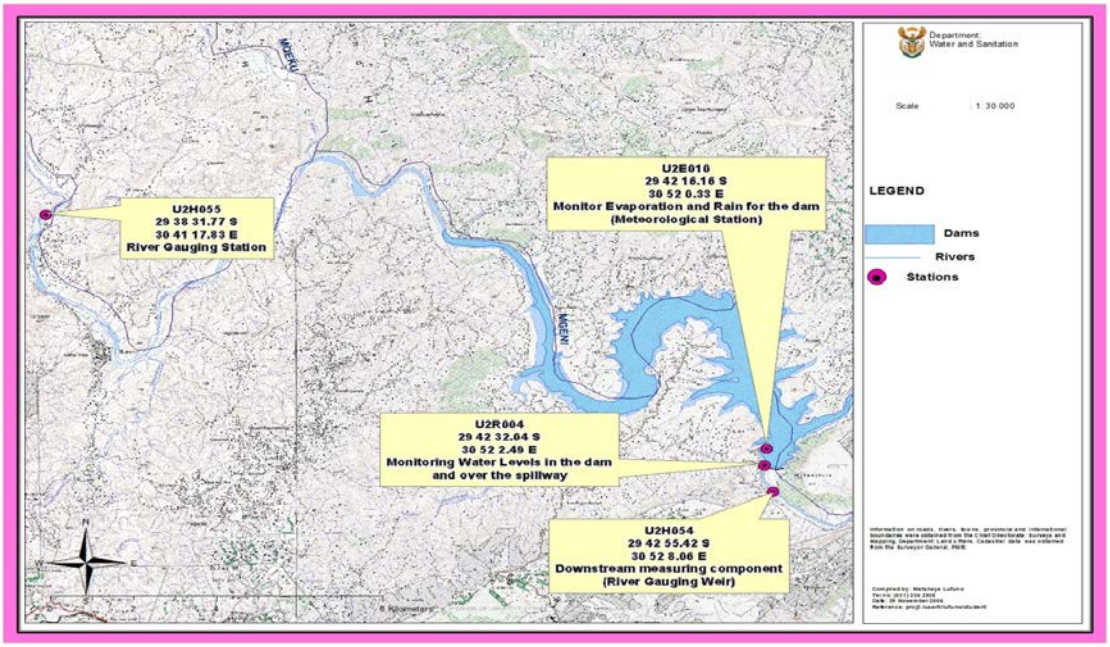


Figure 9: Gauging stations in Inanda catchment area

(Source: The Department of Water Affairs, Durban)

It was observed from plotting the average monthly streamflow dataset (Figure 10) and plotting the average monthly suspended sediment dataset (Figure11) for the

study area that the distribution patterns of both plots were similar. It was also observed that the peak stream flows and largest concentration of suspended sediment occurs between January and March. Likewise, the least stream flow and least concentration of suspended sediment occurred between June and September. This shows a high correlation between streamflow datasets and the suspended sediment datasets.

3.4 MODEL DEVELOPMENT

3.4.1 Input variable section

The selection of an appropriate input variable from the available variables plays a very important role in developing forecasting and prediction models (May *et al.* 2008). This is due to the fact that the performance of these models, especially data driven models like genetic programming, are highly influenced by the input data. The inclusion of unnecessary input variables could increase the size of the model making it computationally complex requiring a larger computational memory and resources. The enlarged search space makes the model calibration more difficult due to the increased presence of local optimal (Back and Trappenberg 1999). According to Lachtermacher and Fuller (1994) and Jayawardena *et al.* (2005) DDM (GP) has the ability to select the appropriate input data so no prior knowledge of the relationship between the variables is needed but in some modelling cases with multiple input variables, the selection of appropriate input variables is generally based on prior knowledge of the variables. However, principal component analysis or cross correlation analysis can be used to determine the appropriate input variable if the relationship between the variables is not well understood (Lachtermacher and Fuller 1994; Sudheer *et al.* 2002; Aqil *et al.* 2007). Cross correlation analysis was used in this study.

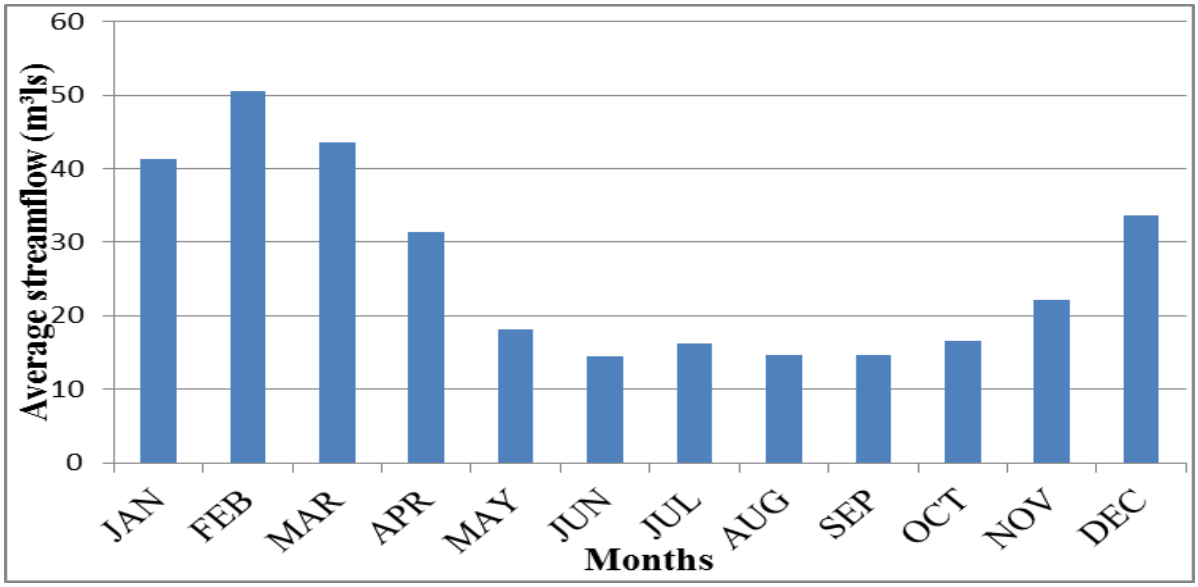


Figure 10: Plot of average monthly streamflow dataset

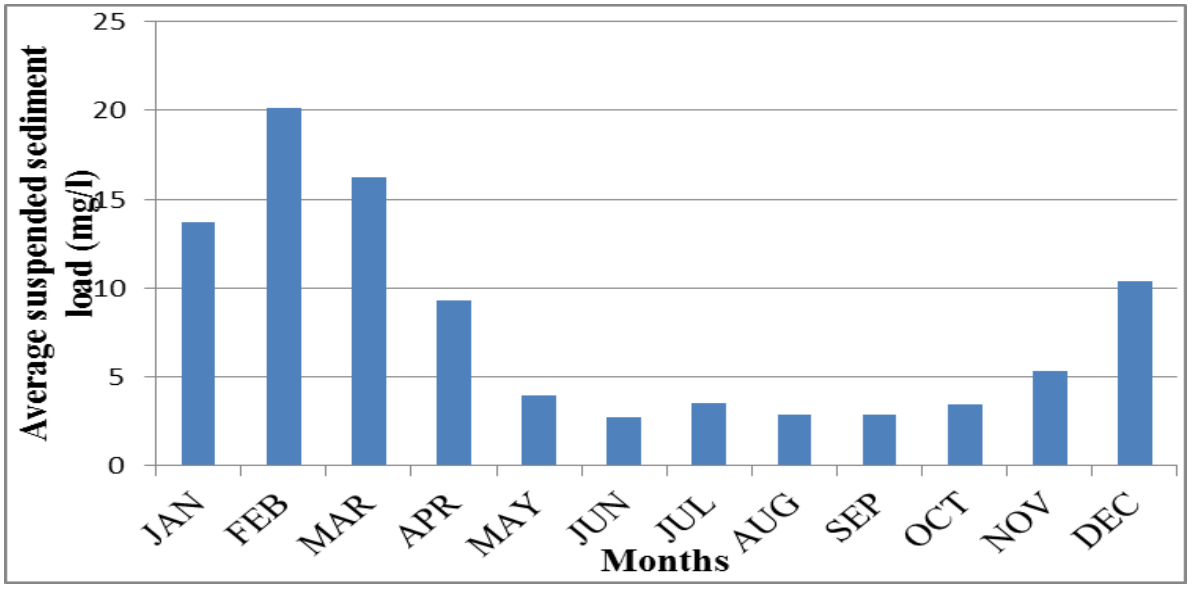


Figure 11: Plot of average monthly suspended sediment dataset

Table 4: Results of correlation analysis between input variables and suspended sediment

Input Parameters	Target Output–suspended sediment (SS)
Turbidity (Tur)	0.2017
Precipitation (P)	0.2055
Temperature (Temp)	0.1867
Streamflow (Q)	0.9943
Rainfall (R)	0.4313

There are several variables that influence the transportation of sediment into a reservoir, some of which are turbidity, precipitation, temperature, streamflow, rainfall, flow depth, particle density, hydraulic radius, sediment size, mean flow velocity, acceleration due to gravity, shear stress, kinematic viscosity, density of water, volumetric concentration of sediment, cross-section geometry, bed roughness, friction factor with sediment, bed slope etc (Ghani 1993). In this study, correlation analysis between suspended sediment and some of these variables (turbidity, precipitation, temperature, streamflow and rainfall) was performed and the results show low correlation values for most of the variables except streamflow which shows a high correlation. This is evidence in Table 4 that shows the correlation analysis result between suspended sediment and those variables. The results showed high correlation between the historical average monthly suspended sediment values and it corresponding streamflow values for 20 years. These results agrees with studies by Bhattacharya and Solomatine (2006) and Kisi and Shiri (2012) which produces similar experience. This may be due to their complex relationship with suspended sediment which cannot be represented as a simple linear relationship. As a result, streamflow and previous suspended sediment data in various time tags were used as the input variables for the development of simple GP models in this study. Also, the unavailability, limitation or inconsistence of other historical variables and the need to develop models with good generalization capabilities also influenced the choice of only streamflow dataset for this study (Aytek and Kisi 2008).

In this study, five input variables were used in the GP input space. These include up-streamflow values for a given month of the year and the last two years' streamflow

values of the same month (Q_t, Q_{t-1}, Q_{t-2}) and the corresponding monthly suspended sediment values for the last two years (SS_{t-1}, SS_{t-2}). The target output is the suspended sediment value (SS_t) for the given month of the year, where the subscript 't' represents the given time period (month). Four input combinations of these five variables were used to develop the suspended sediment model for each month of the year. According to Sivapragasam *et al.* (2011) the development of one model for each month of year gives a better prediction than a single model for the entire year. Therefore twelve models, one for each month of the year, were developed to predict the suspended sediment load for each month of that year.

3.4.2 Data splitting

Once the GP vector space is created by selecting the appropriate input variables and developing the required input combinations, the next phase is to split the available monthly historical dataset into two, namely validation dataset and testing dataset. In this study three quarters of the available monthly streamflow and suspended sediment measured values flowing into Inanda Dam (about 75% of the total dataset) (Muttil and Lee 2005) were used as the validation dataset and the remaining one quarter were used as the testing dataset in the GPdotNET programming software. The validation datasets were used to train and develop the models while the testing datasets were used to validate and check the predictive capabilities of the models (Oyebode 2014).

3.4.3 Fitness Function and Model Fitness

To determine the symbolic expression which shows the relationship between the observed (inputs) variables (streamflow and suspended sediment) and the predicted (output) variables (suspended sediment), GP software, GPdotNET, discussed in section 3.2.1 was applied. The fitness function which was also used as the objective function in this study was the Root Mean Square Error (RMSE). It was used to evaluate individual programs in the search space (population) and the programs with the least error between the predicted values and the measured values are selected.

The fitness function can be represented mathematically as (Olyaie *et al.* 2015; Waters and Curran 2015):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(S_M - S_O)^2} \quad (3)$$

In the equation, S_M and S_O represent predicted and observed suspended sediment values respectively, n represents number of data points, and i is a counter which is from 1 to the number of data points.

3.4.4 GP algorithm setup

The instructions consisting of the selected input variables, the functional set and randomly generated constant were used for running the GP algorithm in GPdotNET. In this study the functional set is made up of the following basic arithmetic operators and standard mathematical functions (+, -, *, /, Exp(x), ln(x), √, sin(x), cos(x), tan(x)) while the terminal set includes several input combinations of variables (streamflow and suspended sediment values) and the generated constant 6 which is usually between 0 and 10 contants. For the run, the number of programs in the population per generation was set at 500 to prevent bloating (increase in the size of the program without a corresponding increase in its accuracy); elitism is 1 (the selection of the best program from every generation). During the algorithm run, the following GP operators of crossover, mutation and reproduction were set at different probabilities: crossover with a probability of 0.9, mutation with a probability of 0.05 and reproduction with a probability of 0.2 as advised by Poli *et al.* (2008). These GP operators were discussed in detail in section 2.15.4 of this thesis. The complete list of the GP parameters used for the runs is found in Table 5. The initial population was generated using the ramped half-and-half method (see section 2.15.2) to produce programs of different sizes and shapes (Koza 1992). The tournament selection process was used to create the mating pool that is the temporary population for the GP operators. The stop condition for the GP algorithm run was the maximum generation of 500 which is to prevent bloating as explained.

Table 5: GP parameters and settings used in GPdotNET

Process Configuration	
Population Size	500
Fitness	RMSE
Initialization	Half-and-half method
Selection	
Elitism	1
Method	Tournament selection
Tournament size	2
Maximum Tree Depth	
Initializing depth	5
Operation depth	6
Probability of GP operators	
Crossover	0.9
Mutation	0.05
Reproduction	0.2
Random constants	
Counts between 0 and 10	6
Processor	Multi core

3.4.5 Evaluation of the performance of the models

In the development of any model it is very important to develop and analyse algorithms by evaluating the accuracy of the developed models. The most common form of performance evaluation used is to perform statistical comparisons of the accuracy of the developed models. This involves the use of mathematical functions to estimate the error or differences between the observed values and the predicted values. In this study the accuracy of the developed monthly models was evaluated using the Coefficient of Determination (R^2) and the Root Mean Square Error (RMSE). The Coefficient of Determination (R^2) can be expressed mathematically as (Schindler *et al.* 2015; Wang *et al.* 2015):

$$R^2 = \left[\frac{\sum_{i=1}^{n} (S_0 - \bar{S}_0)\,(S_M - \bar{S}_M)}{\sqrt{\sum_{i=1}^{n} (S_0 - \bar{S}_0)^2 \sum_{i=0}^{n} (S_M - \bar{S}_M)^2}} \right]^2 \qquad (4)$$

where S_O is the suspended sediment amount observed (recorded) at the ith time step, S_M is the corresponding simulated suspended sediment, n is number of time steps, $\overline{S}_O$ is mean of observational values and $\bar{S}_M$ is mean value of the simulations. The Coefficient of Determination (R^2) ranges between 0 and 1, with higher values indicating the better performance of the model and 1 being the perfect value. The R^2 values are an indication of the linear relationship between the observed and the corresponding predicted values. They are also sensitive to outliers so the Coefficient of Determination (R^2) is not used in isolation to evaluate the developed models (Legates and McCabe 1999). Therefore, the developed monthly models were also evaluated using the Root Mean Square Error (RMSE). The combined use of these measures provided a good insight into the performance of the developed models.

3.5 RESULTS AND DISCUSSION

As explained in section 1.4.1, GPdotNET programming software was used to develop 240 models from 240 computer simulations (20 models for each month of the year), each model consisting of both a training phase and a validation phase. The best models with minimum error (RMSE) and corresponding highest R^2 values in each month were selected to represent the best model to predict most accurately the quantity of suspended sediment flowing into Inanda Dam. The statistical performance of the individual monthly models at both training phase and validation phase are presented in Table 6.

From Table 6, it can be observed that all the monthly models produced very low RMSE and very high R^2 values both during the training and validation phases indicating the accuracy of GP. This is in agreement with the study of Londhe and Charhate (2010) that concludes that GP models can predict accurately both normal and extreme events. During the training phase the developed monthly models produced R^2 values of between 0.9941 and 0.9999. This shows a strong and positive correlation between the measured and predicted suspended sediment during the training phase. The models in the months of March and April have the lowest error of 0.99417 and 0.99644 respectively while the models in the months of January, May and August have the highest errors of 0.999906, 0.999998 and 0.999979

respectively, while the RMS errors range from 0.0266 to 5.9987. The lowest errors were in the months of May, June, August and October with values of 0.0576, 0.0613, 0.0266 and 0.0615 respectively and the highest errors occurred with the models in the months of March and April with values of 4.8973 and 5.9987 respectively.

Table 6: Statistical performance of the best individual monthly models

MONTH	TRAINING PHASE		VALIDATION PHASE	
	R^2	RMSE	R^2	RMSE
JAN	0.999906	0.232603	0.999970	0.214396
FEB	0.999633	0.669156	0.996507	2.382141
MAR	0.994170	4.897335	0.999962	0.548428
APR	0.996440	5.998664	0.999741	0.201124
MAY	0.999998	0.057579	0.999999	0.032855
JUN	0.999785	0.061325	0.999997	0.064152
JUL	0.999867	1.479767	0.999978	0.066710
AUG	0.999979	0.026600	0.999990	0.039383
SEP	0.999734	0.416884	0.999992	0.148439
OCT	0.999625	0.061456	0.999965	0.150845
NOV	0.999320	0.248159	0.999967	0.173291
DEC	0.999887	0.285832	0.999965	0.257714
AVG	0.999029	1.202947	0.999669	0.356623

In the validation phase the monthly models produced R^2 values of between 0.996507 and 0.999999. The lowest R^2 values were produced by the models in the months of February and April with values of 0.996507 and 0.999741 respectively and the highest R^2 values were produced by the models in the months of May and June with the values of 0.999999 and 0.999997 respectively while the RMS errors ranged from 0.032855 to 2.382141. The lowest errors occurred in the models of the months of May and August with errors of 0.032855 and 0.039383 respectively. The models in the months of February and March produced the highest errors of 2.382141and 0.548428 respectively. Also, the models in the months of May, June, July and August with the lowest streamflow values have the lowest errors and the models in the months of December, January, February and March with the highest streamflow values have the highest errors. According to Mugumo (2012) who got similar result in his study, this pattern may be due to the absence of many high streamflow values

in the training dataset. However the high performance of all the models during the validation phase shows strong positive correlations between observed and predicted suspended sediment values (Wang *et al.* 2009).

Generally the error estimates in terms of R^2 and RMSE values in Table 6 show that most of the models converge better during the validation phase than during the training phase. The exceptions were the models in the months of February, March and April where increment occurred. This agrees with the studies of Mugumo (2012) that found that the slight differences in errors between the training and validation phases show high similarities in their dataset patterns. These results generally show the accurate performance and consistency in prediction of the developed GP monthly models (Danandeh Mehr *et al.* 2013). The observed and GP-predicted suspended sediment load (mg/l) for the twelve months during the validation and training phases are presented in Figures 12 to 35. These figures are the graphical and visual presentation of the outputs (suspended sediment) from the GPdotNET programme used in this study for both the training and validation phase. They were exported directly so this affects the quality of the figures. These figures show the degree of correlation between the observed suspended sediment values (historical data) flowing into Inanda Dam and the predicted suspended sediment values from GP. It also shows the pattern of suspended sediment in terms of quantity flowing into Inanda Dam and the ability of GP to recognise that pattern.

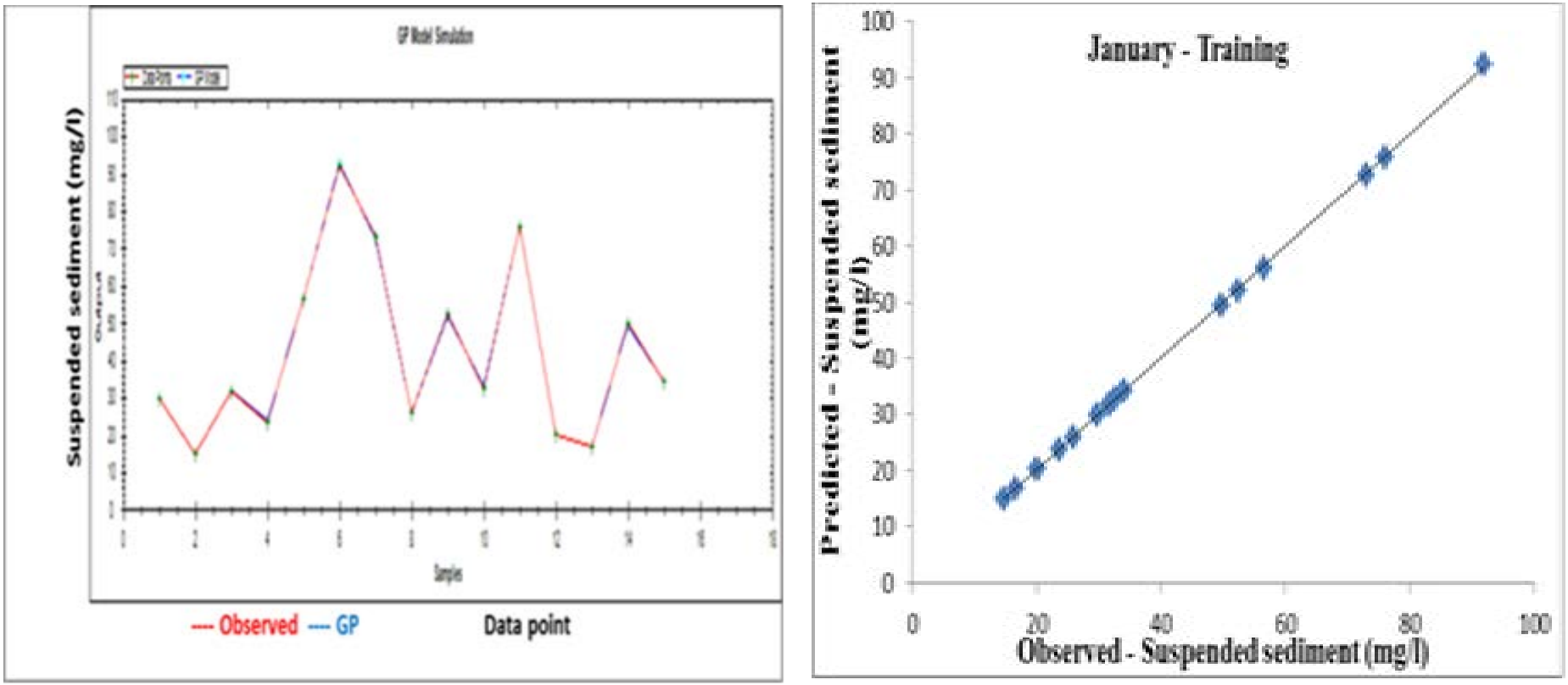


Figure 12: Plots of observed and GP-predicted suspended sediment (mg/l) for January during training phase

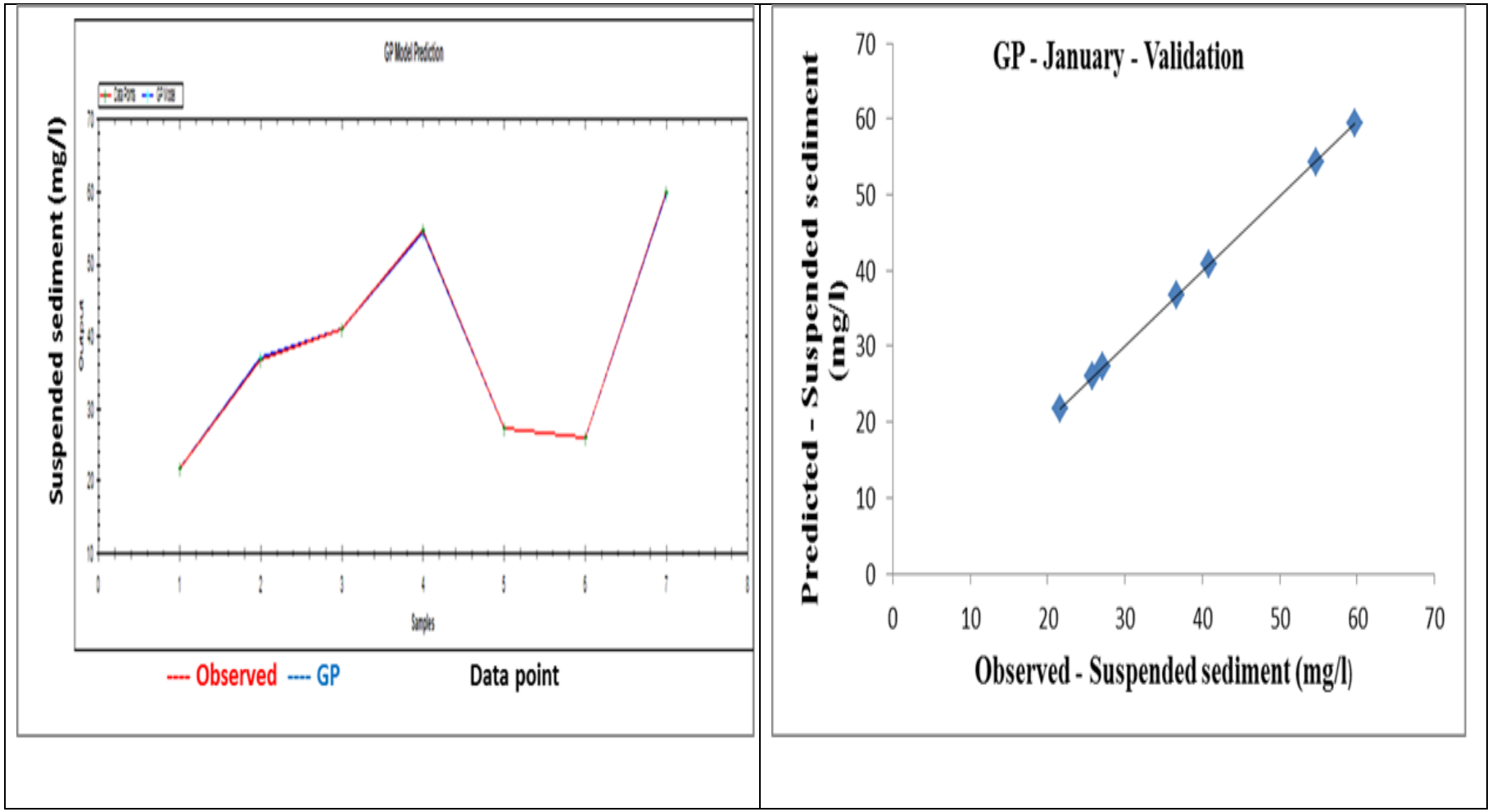
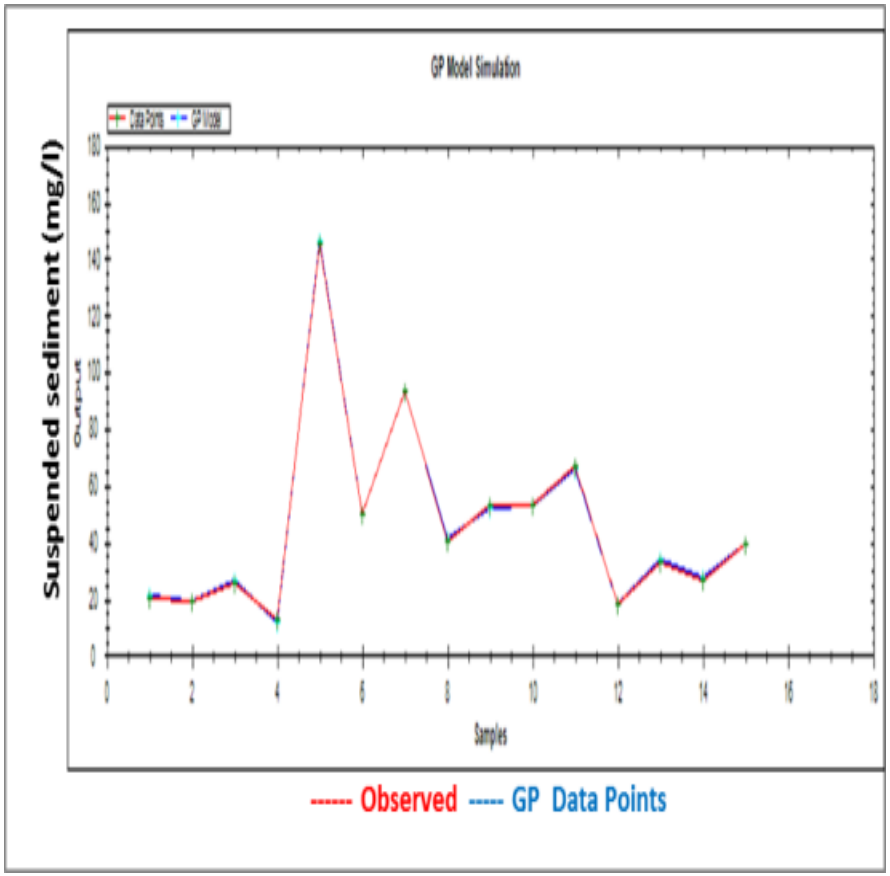


Figure 13: Plots of observed and GP-predicted suspended sediment (mg/l) for January during validation phase



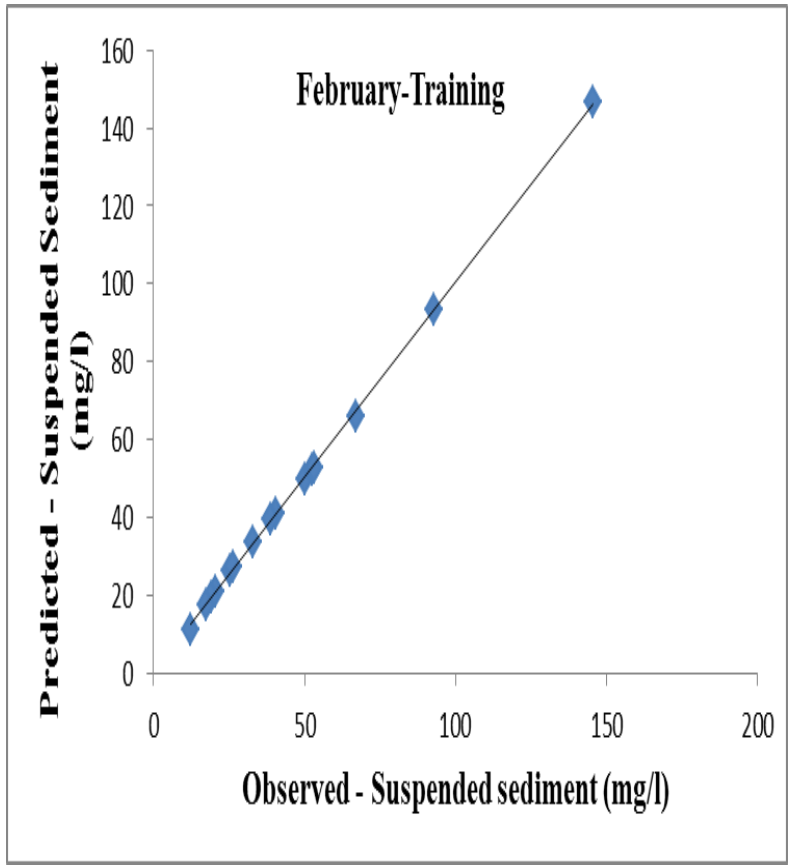


Figure 14: Plots of observed and GP-predicted suspended sediment (mg/l) for February during training phase

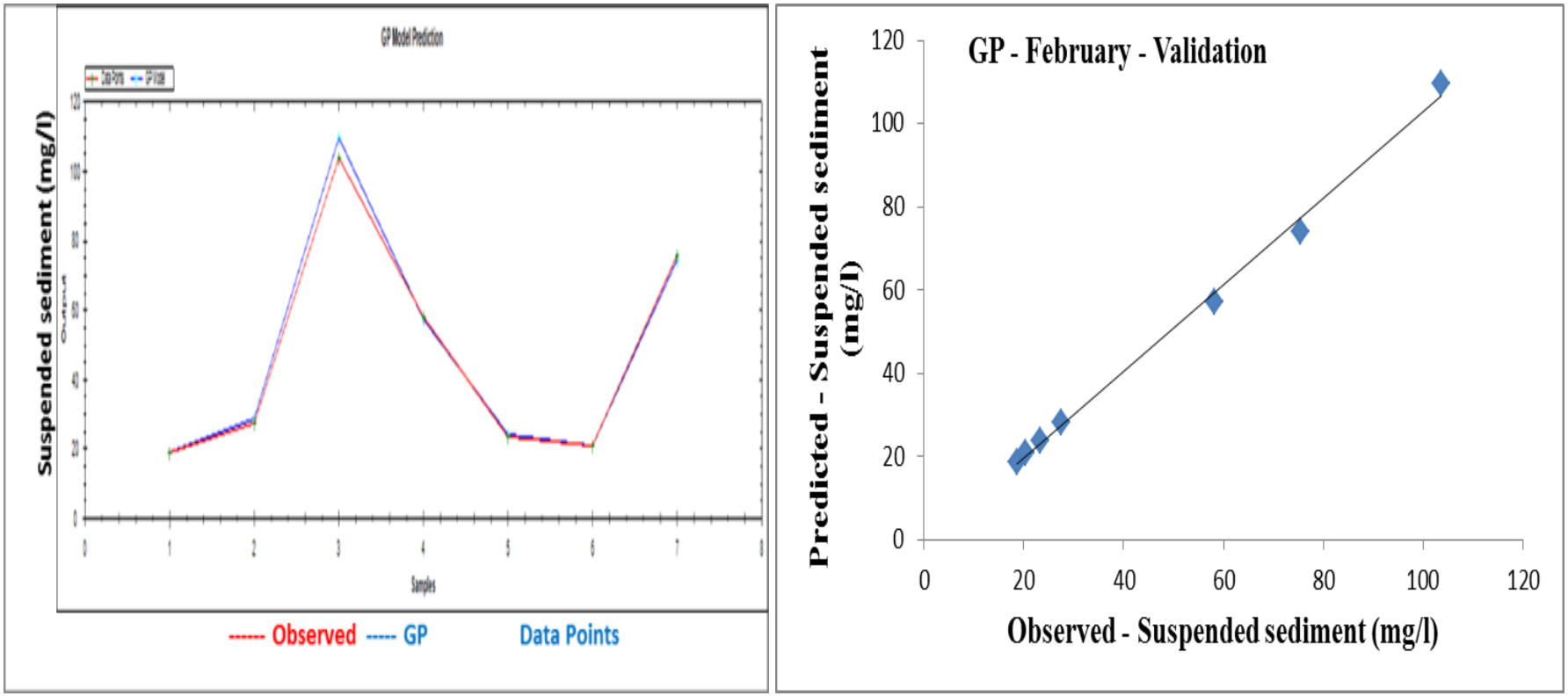


Figure 15: Plots of observed and GP-predicted suspended sediment (mg/l) for February during validation phase

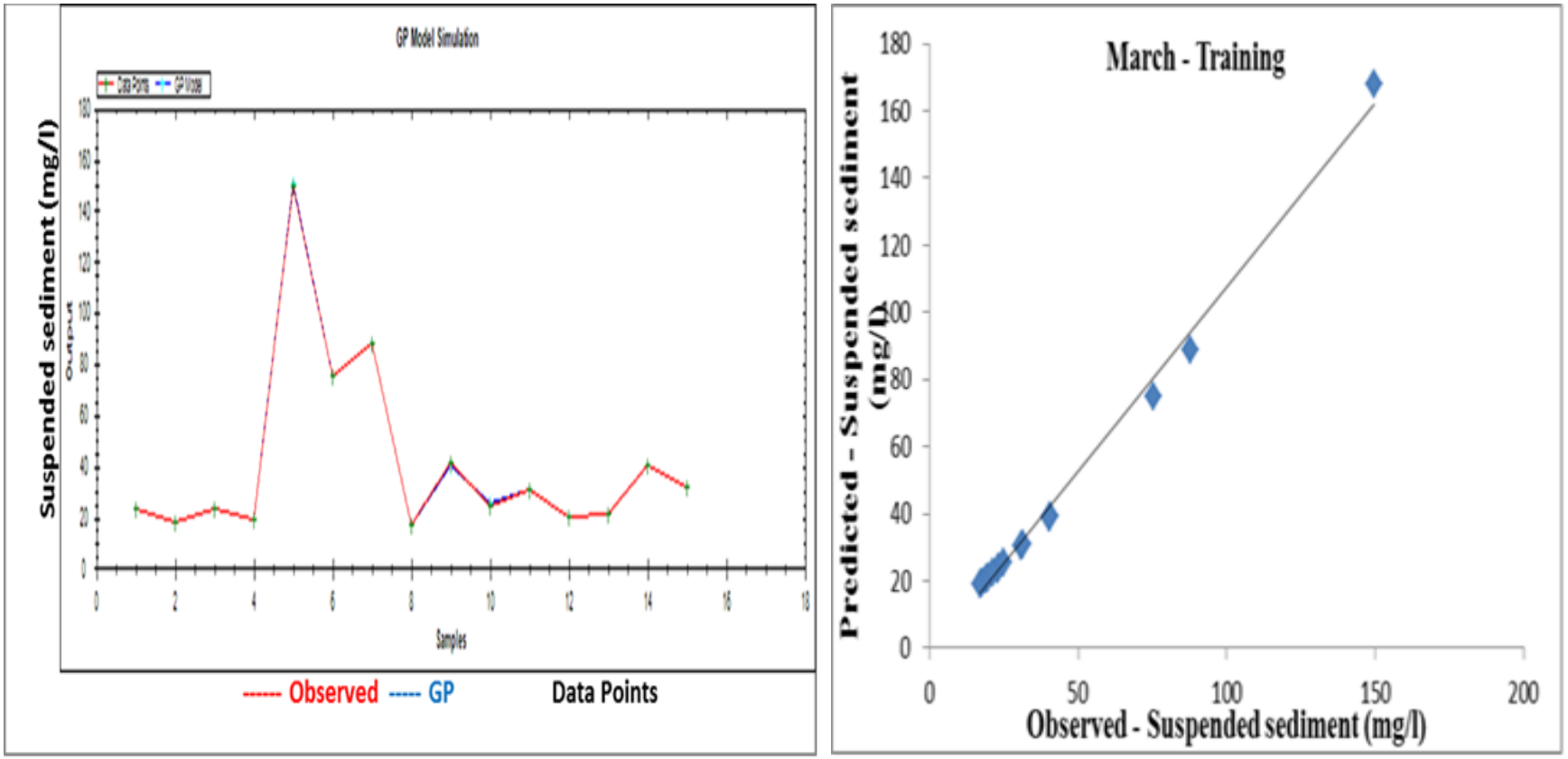


Figure 16: Plots of observed and GP-predicted suspended sediment (mg/l) for March during training phase

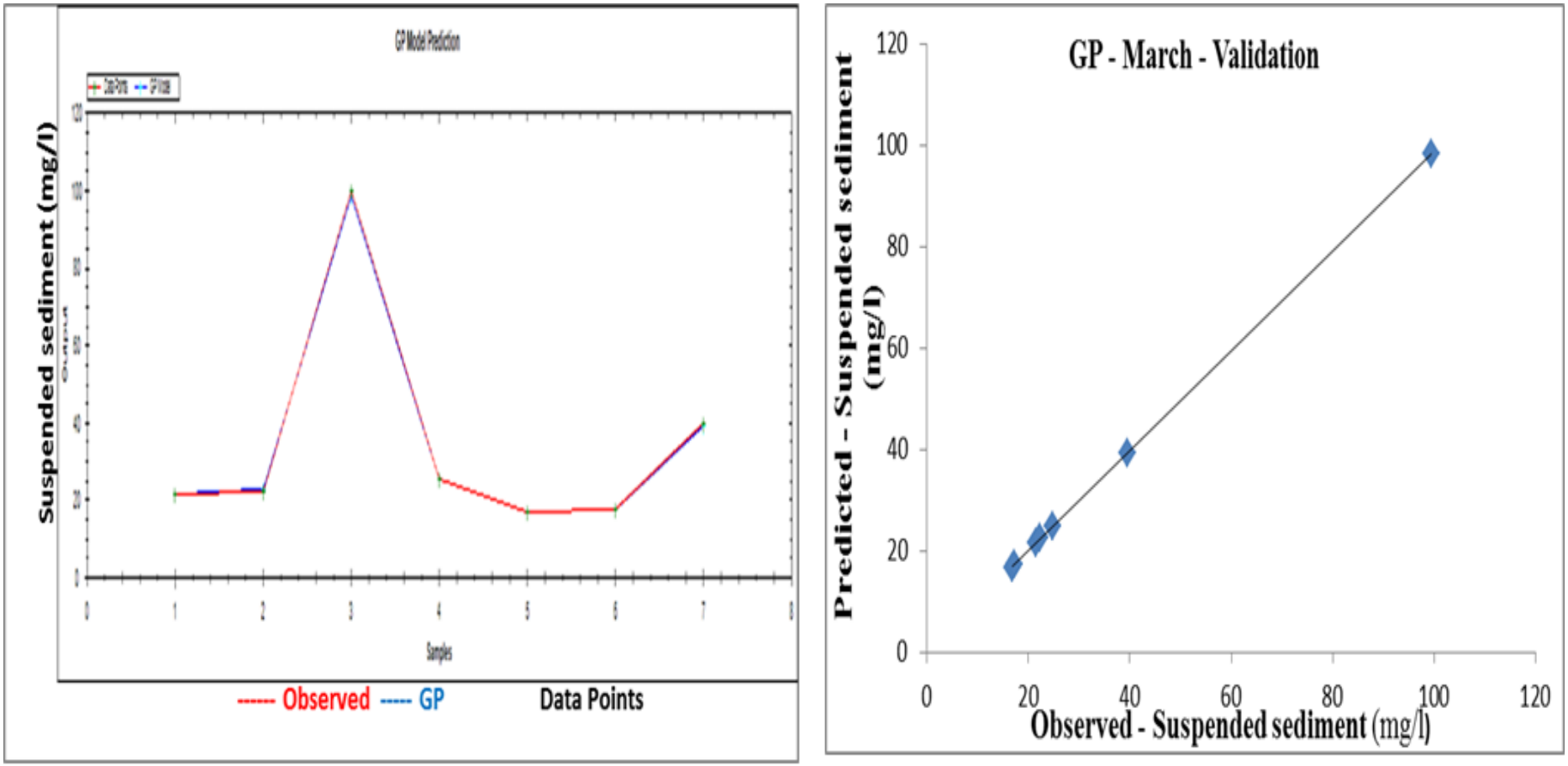


Figure 17: Plots of observed and GP-predicted suspended sediment (mg/l) for March during validation phase

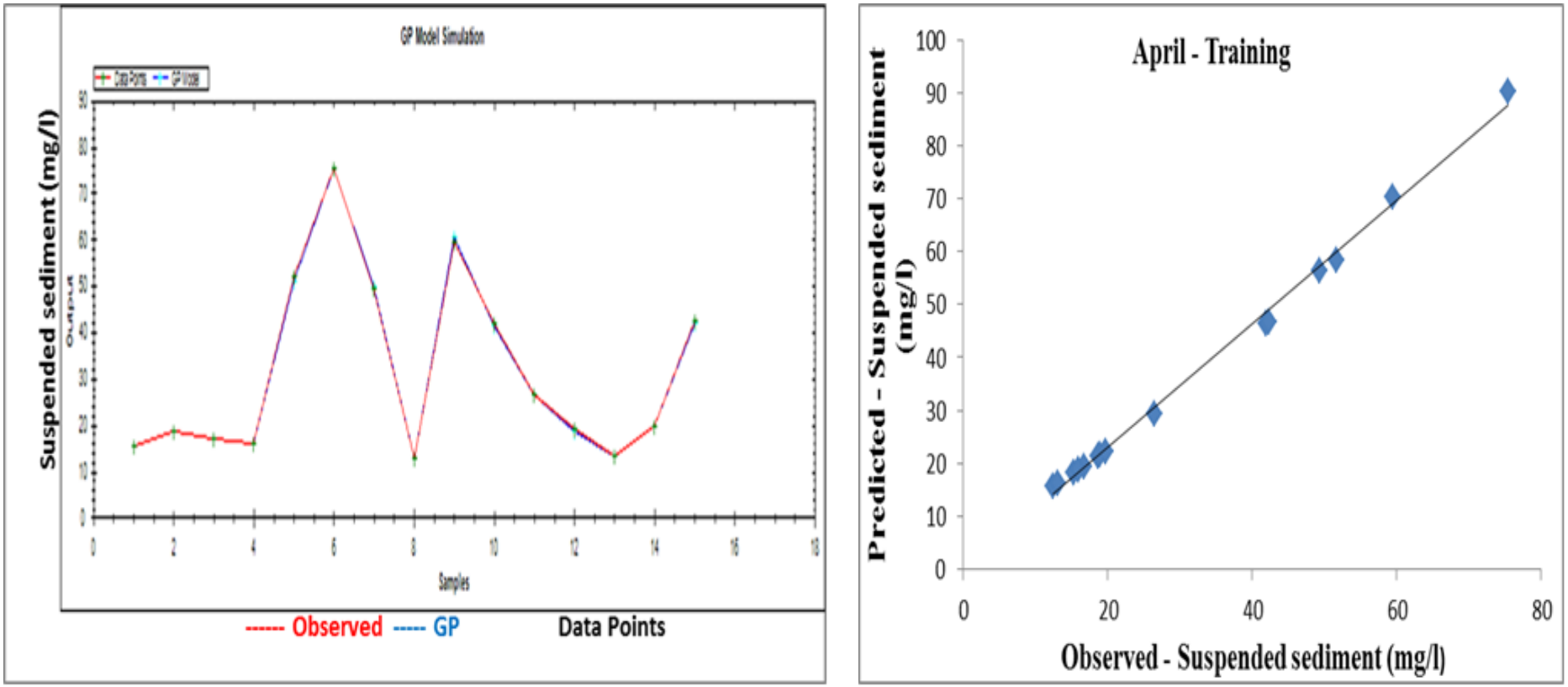


Figure 18: Plots of observed and GP-predicted suspended sediment (mg/l) for April during training phase

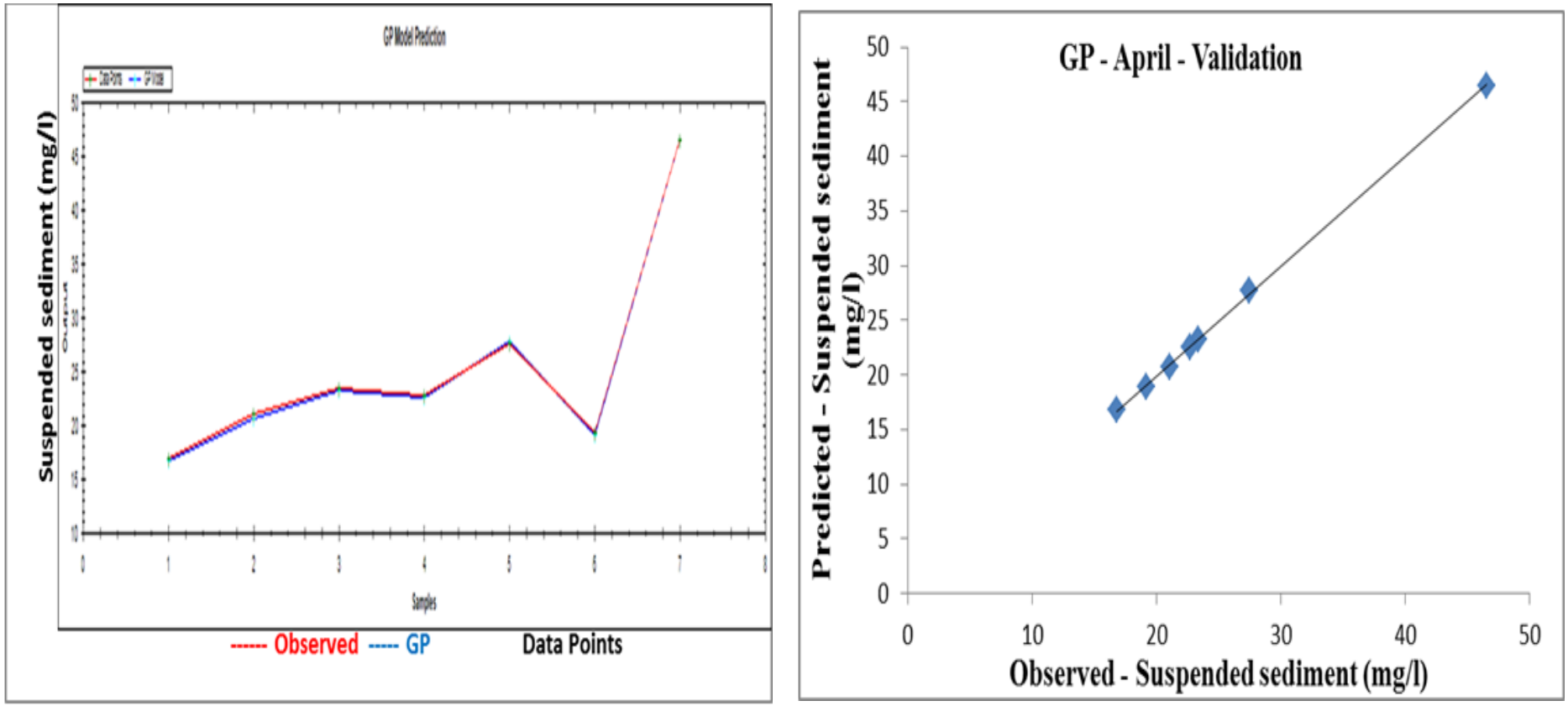


Figure 19: Plots of observed and GP-predicted suspended sediment (mg/l) for April during validation phase

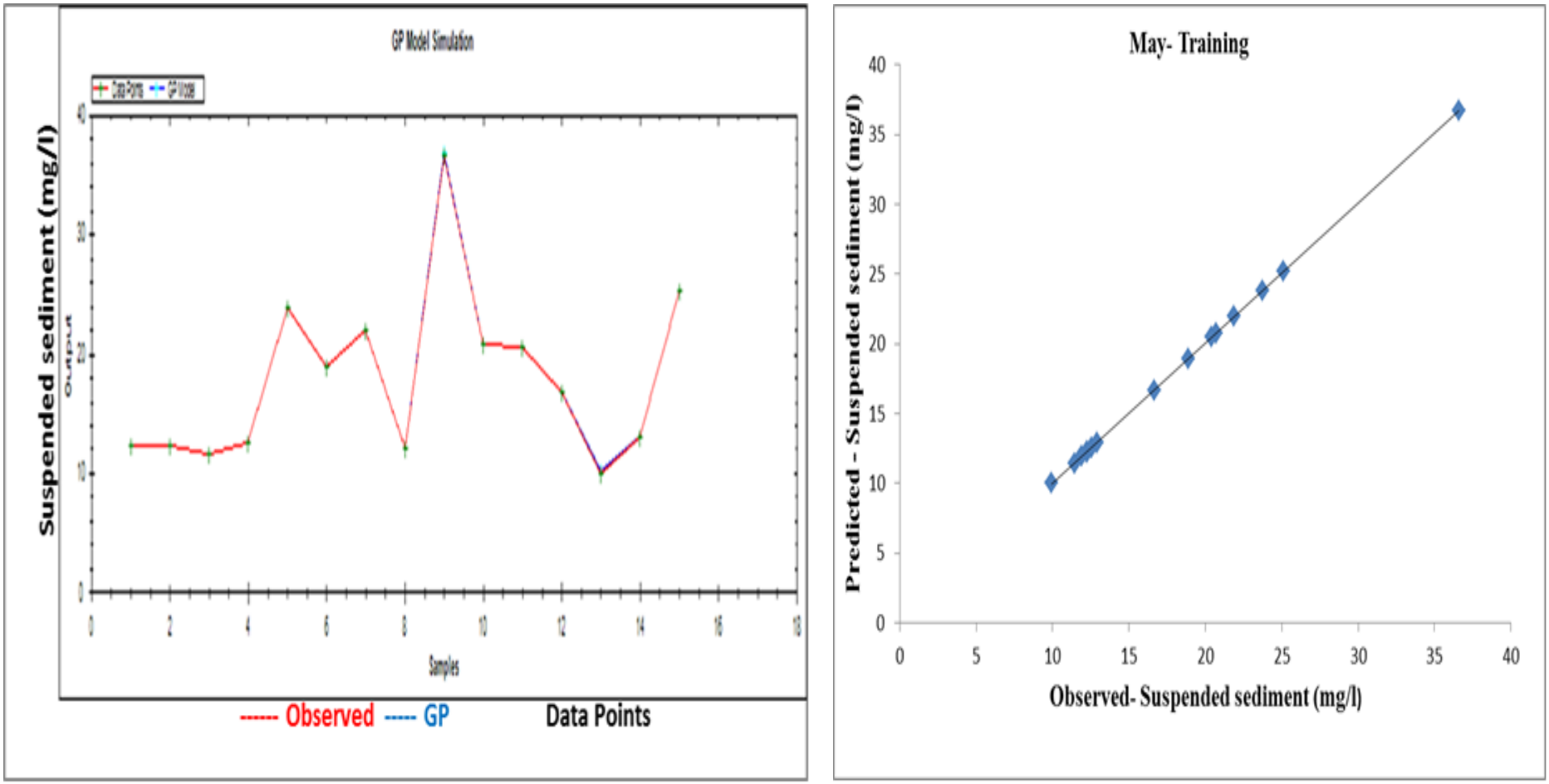


Figure 20: Plots of observed and GP-predicted suspended sediment (mg/l) for May during training phase

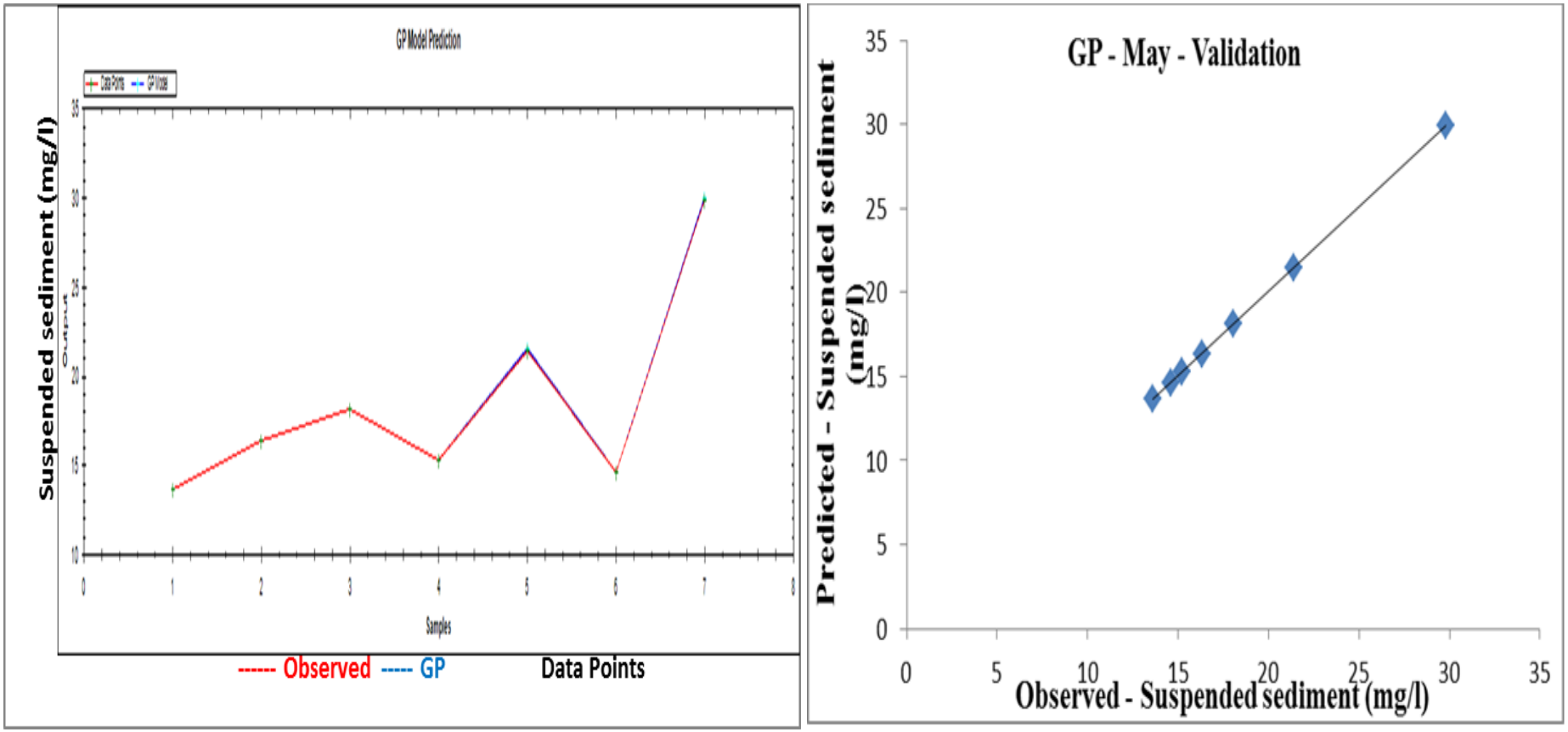


Figure 21: Plots of observed and GP-predicted suspended sediment (mg/l) for May during validation phase

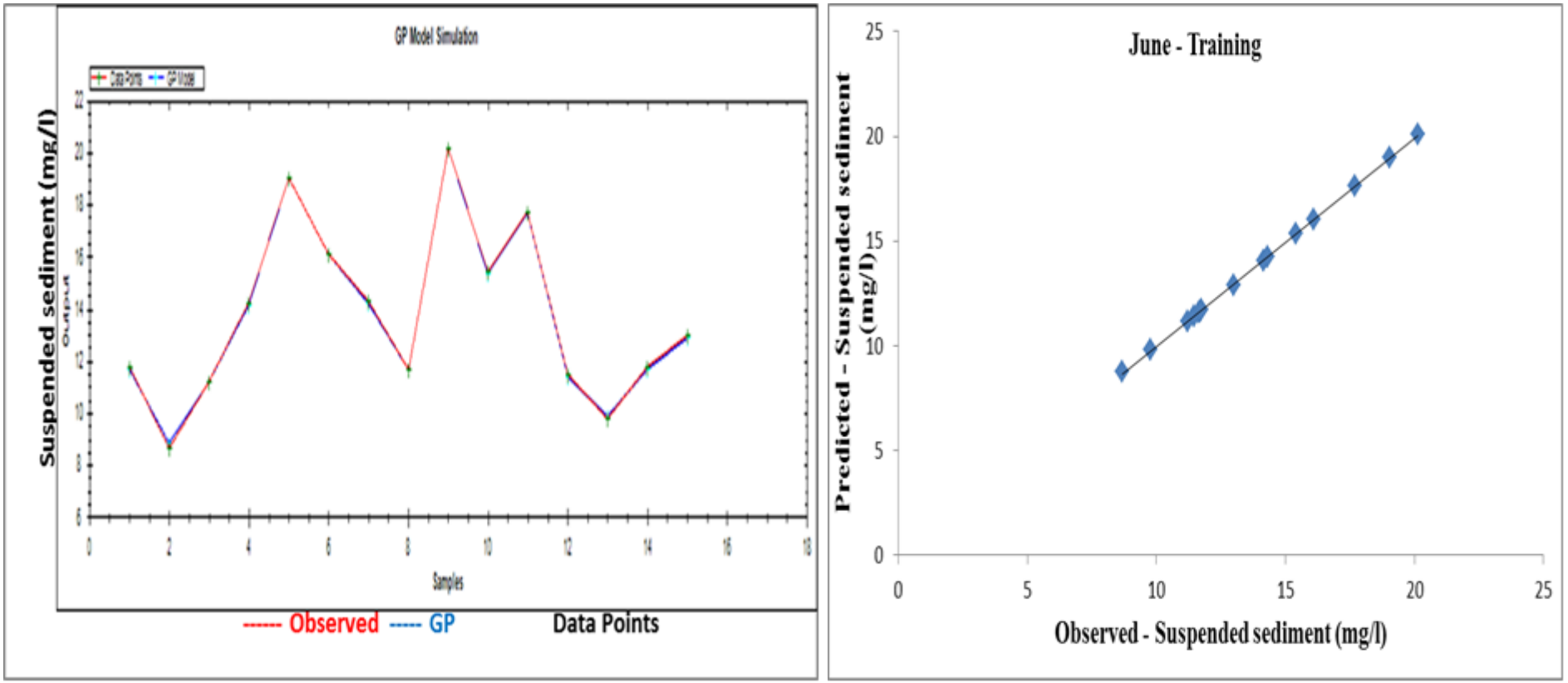
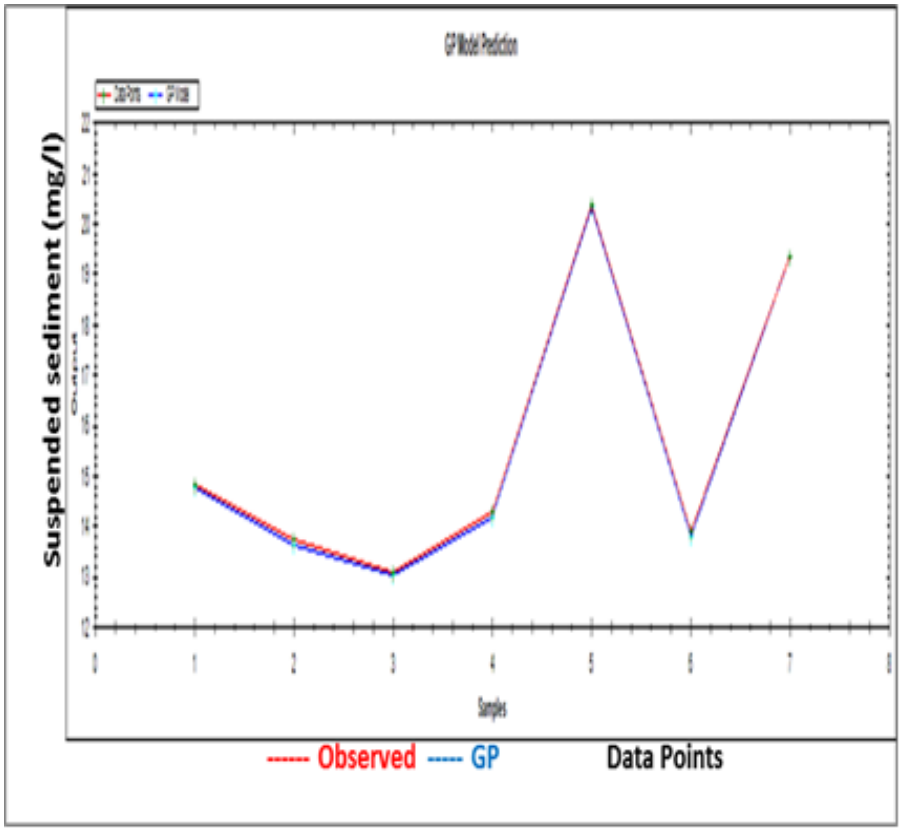


Figure 22: Plots of observed and GP-predicted suspended sediment (mg/l) for June during training phase



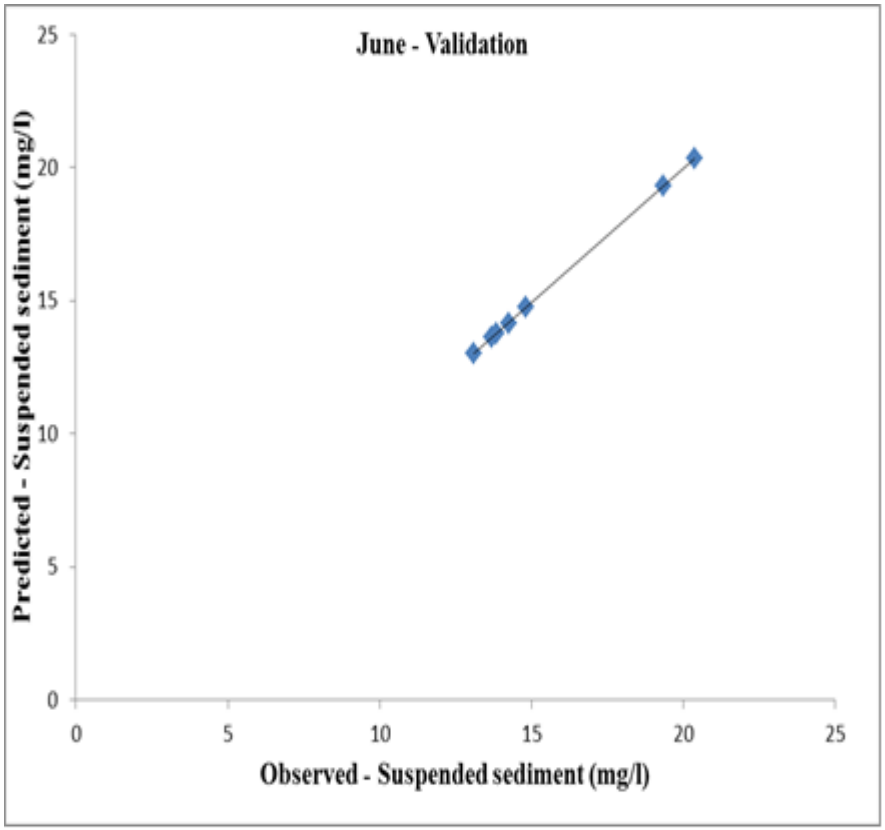


Figure 23: Plots of observed and GP-predicted suspended sediment (mg/l) for June during validation phase

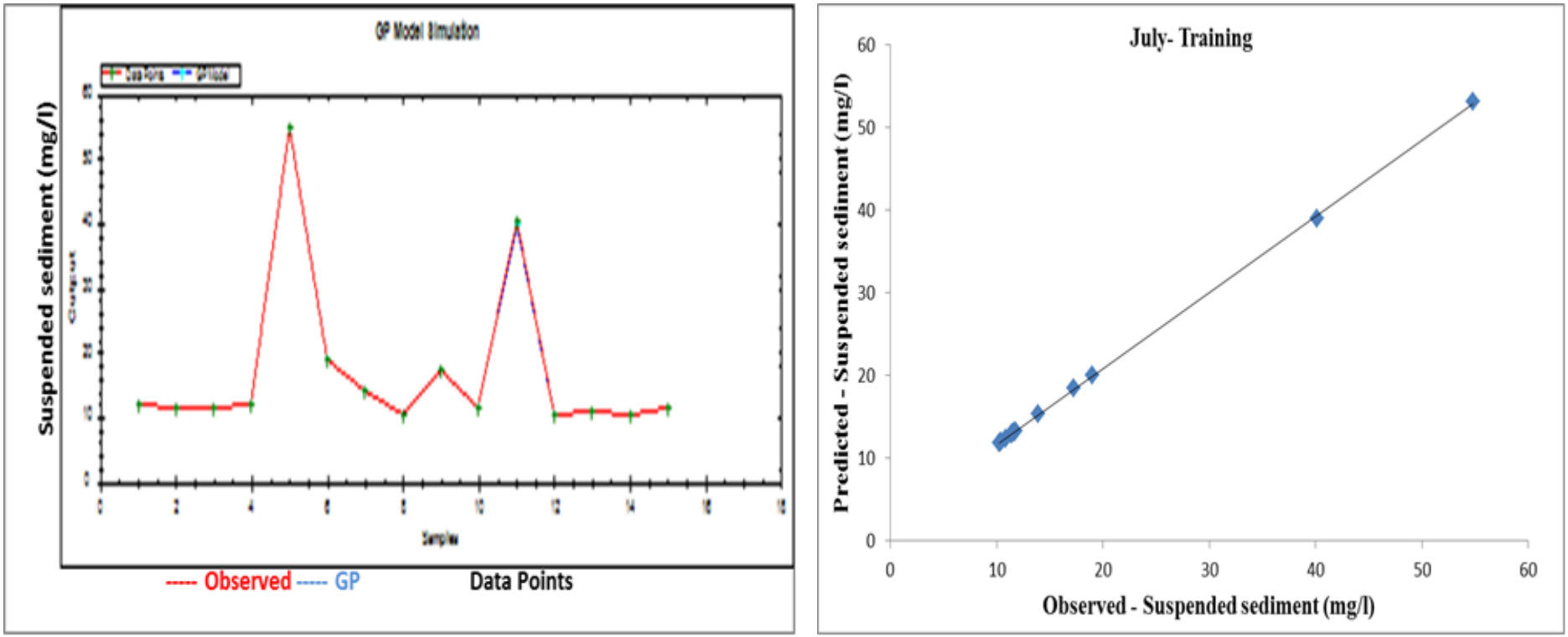


Figure 24: Plots of observed and GP-predicted suspended sediment (mg/l) for July during training phase

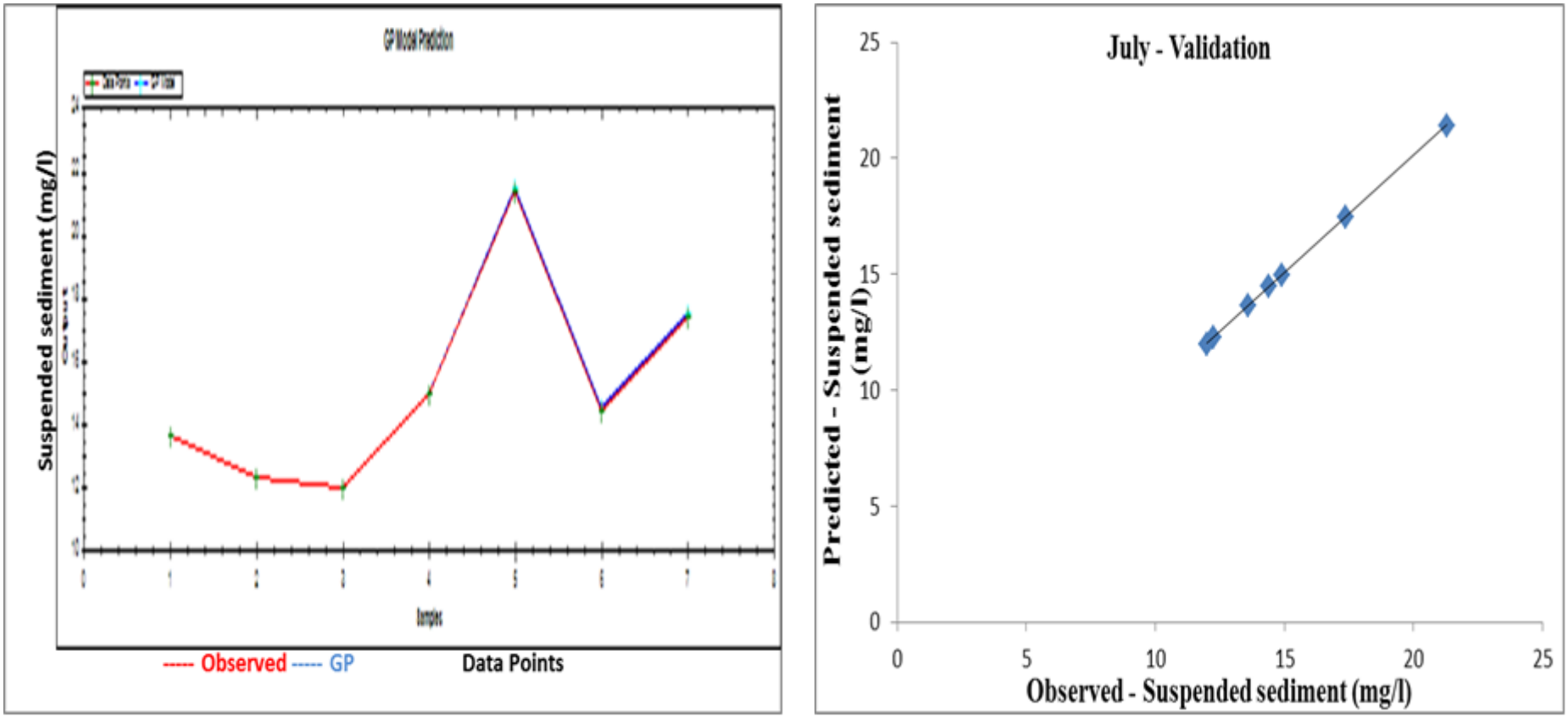


Figure 25: Plots of observed and GP-predicted suspended sediment (mg/l) for July during validation phase

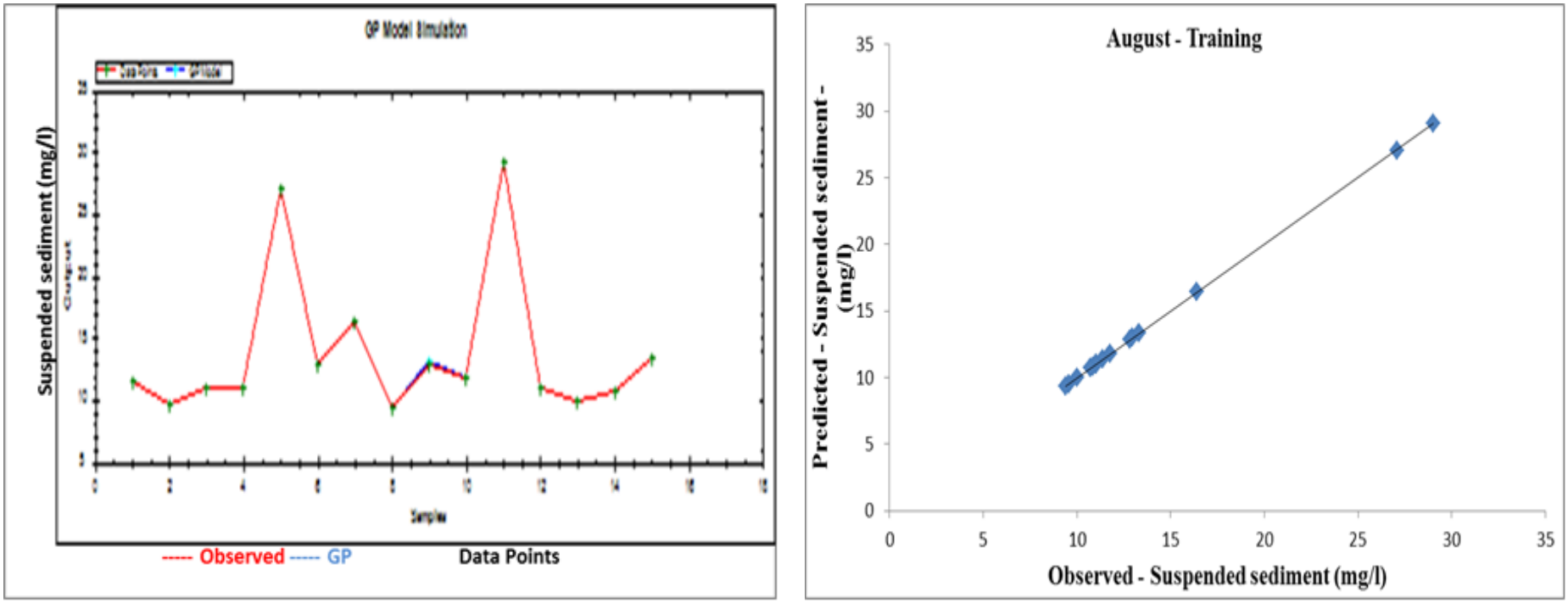


Figure 26: Plots of observed and GP-predicted suspended sediment (mg/l) for August during training phase

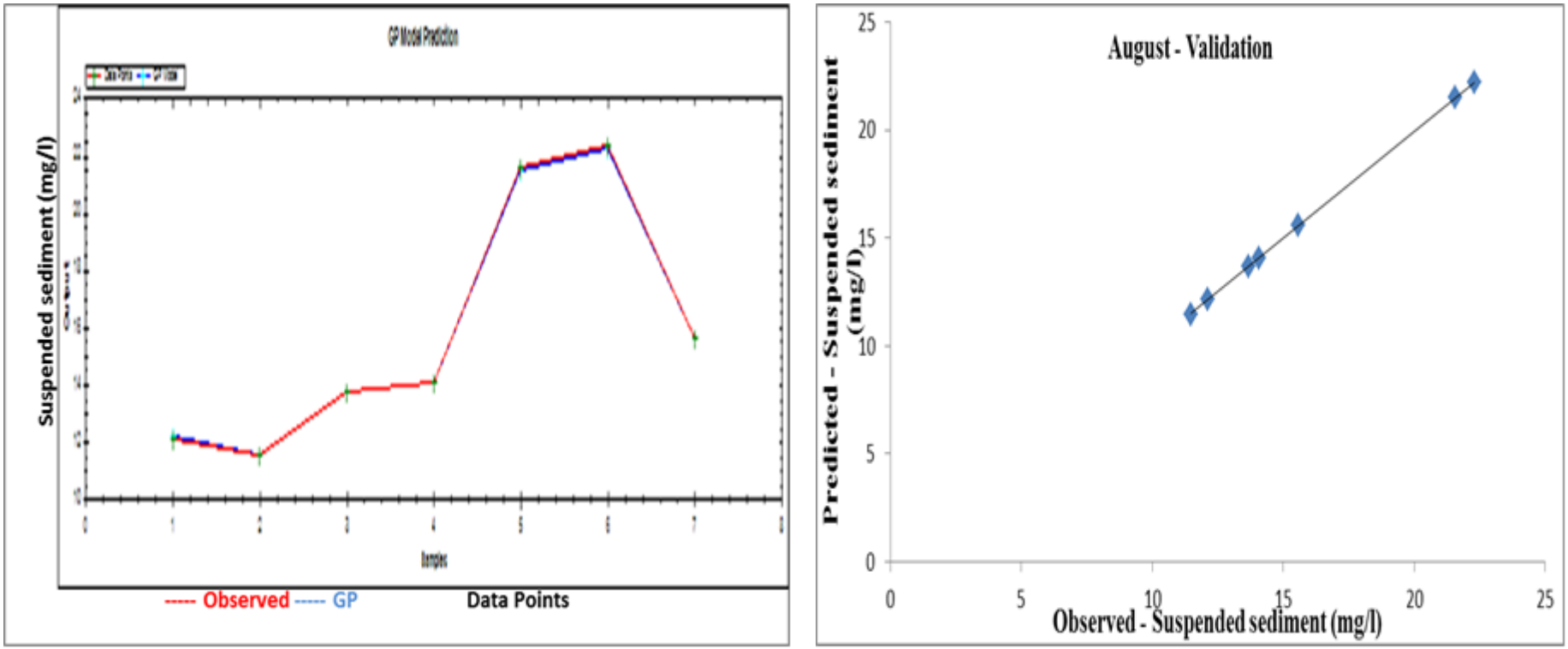


Figure 27: Plots of observed and GP-predicted suspended sediment (mg/l) for August during validation phase

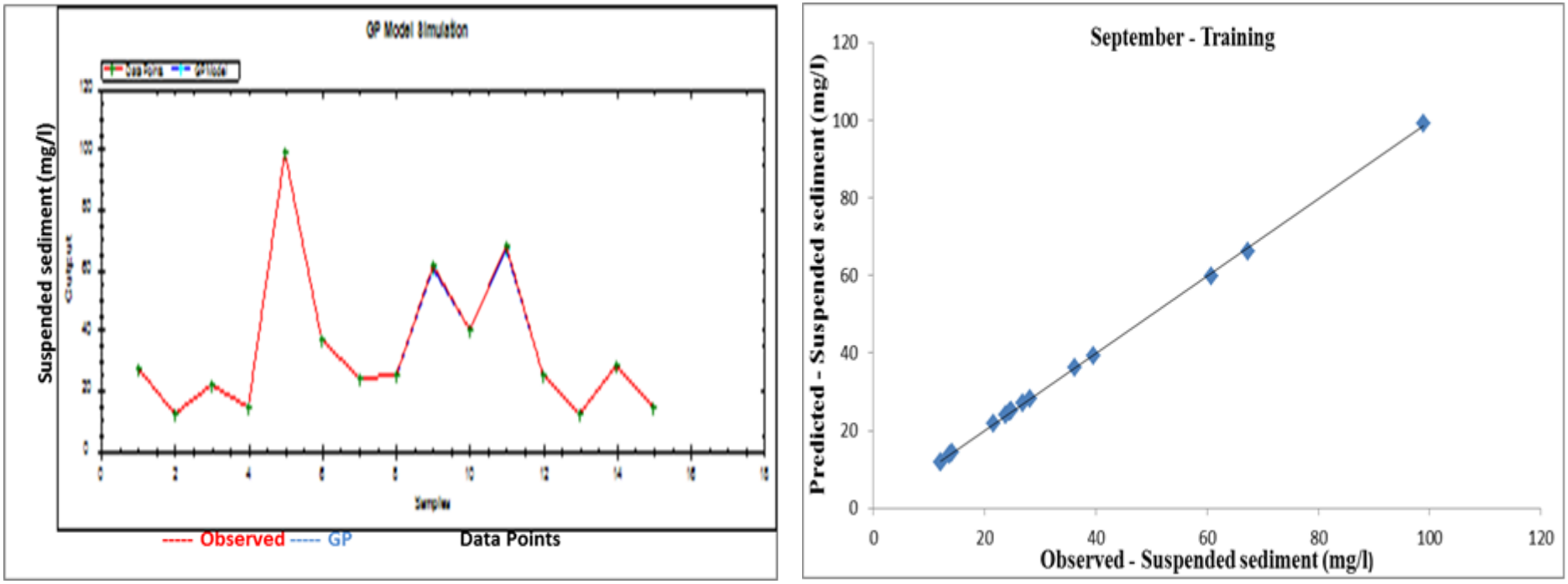
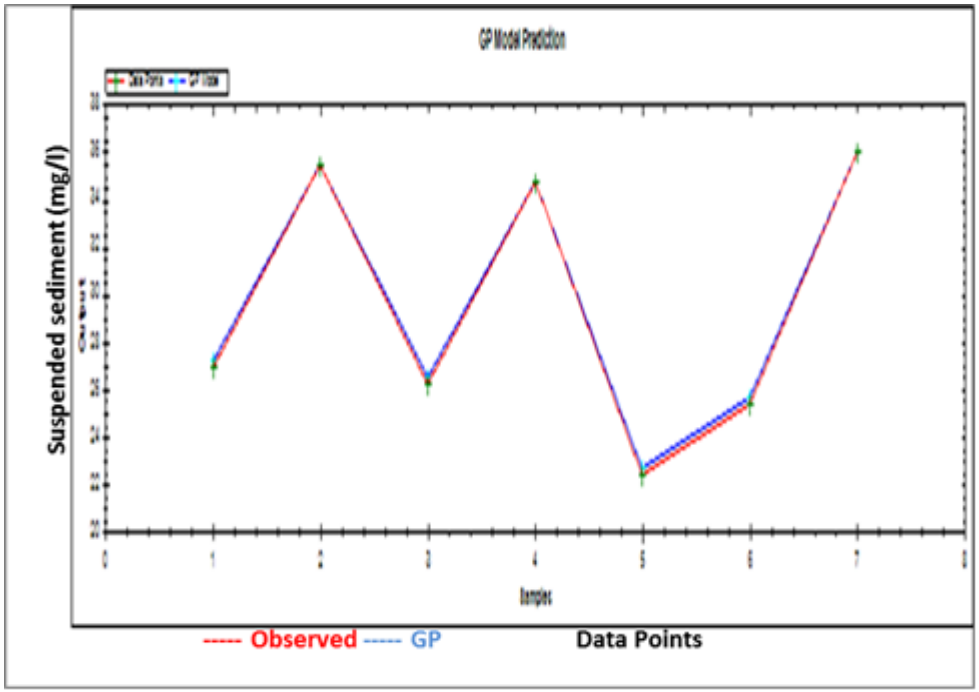


Figure 28: Plots of observed and GP-predicted suspended sediment (mg/l) for September during training phase



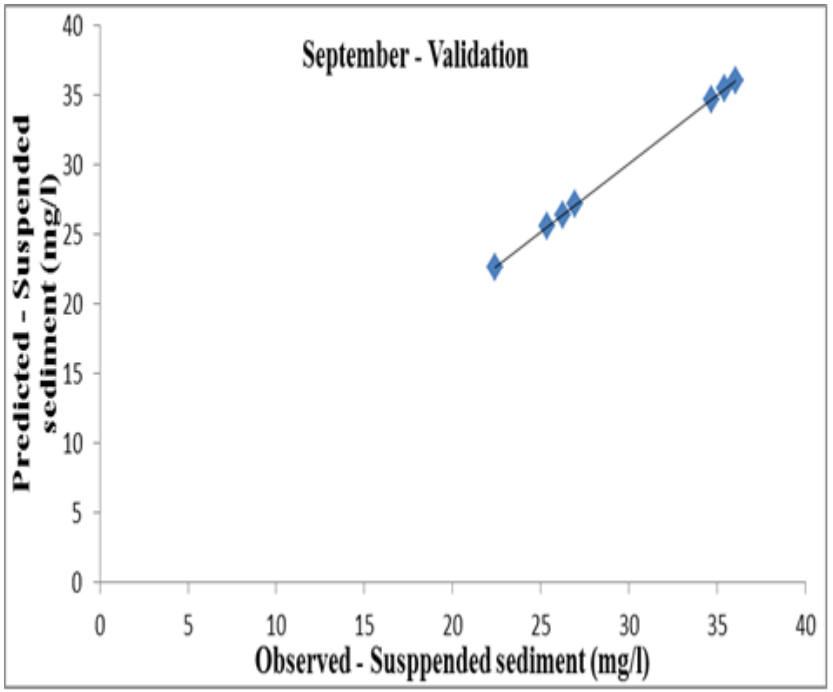


Figure 29: Plots of observed and GP-predicted suspended sediment (mg/l) for September validation phase

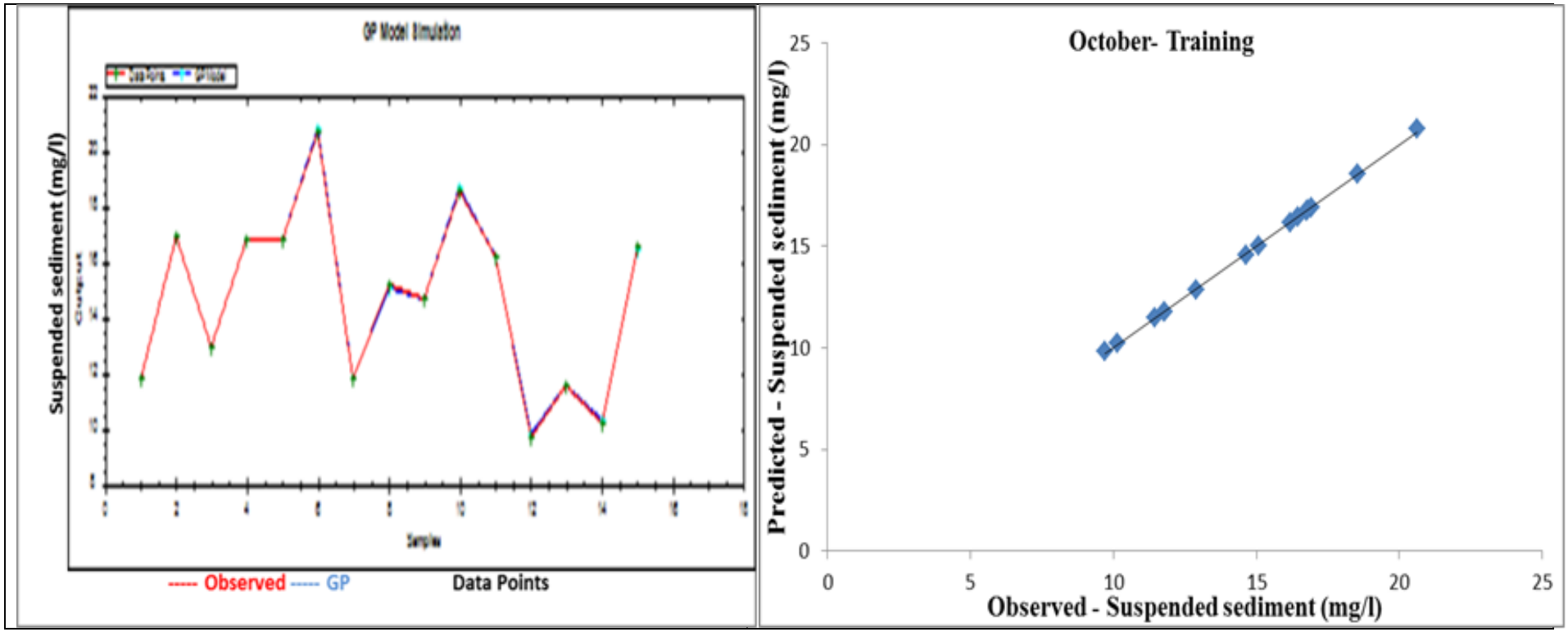
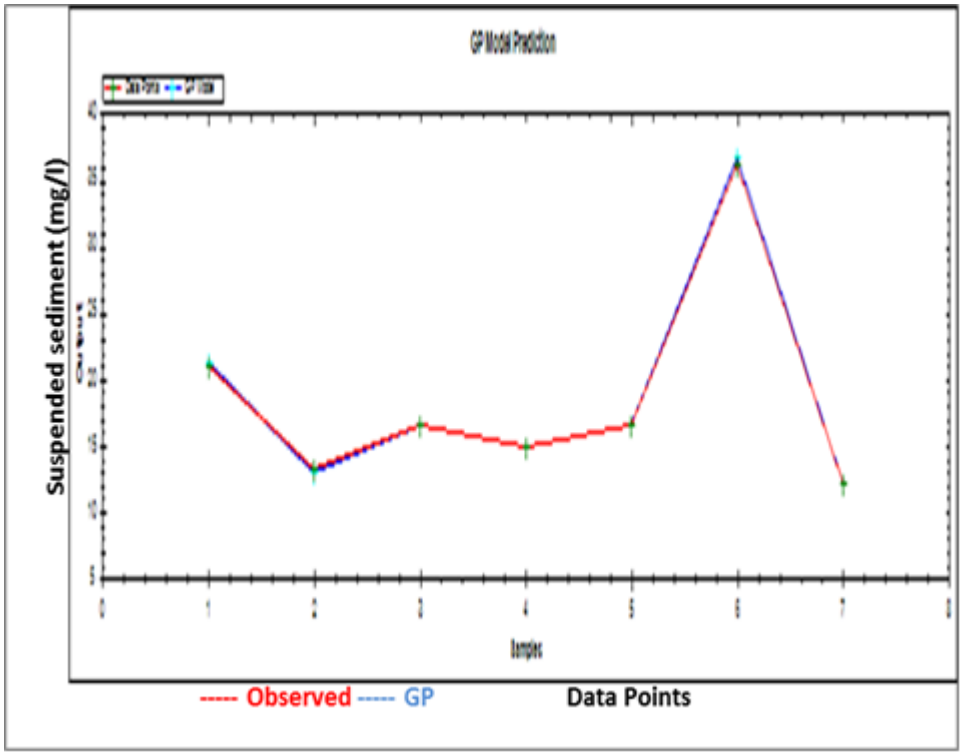


Figure 30: Plots of observed and GP-predicted suspended sediment (mg/l) for October during training phase



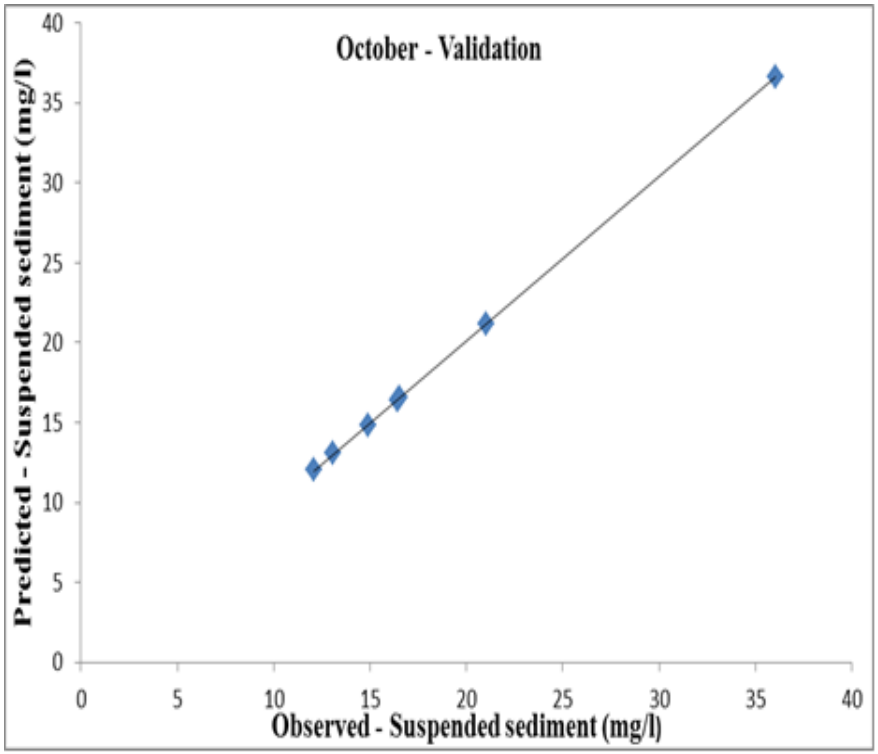


Figure 31: Plots of observed and GP-predicted suspended sediment (mg/l) for October during validation phase

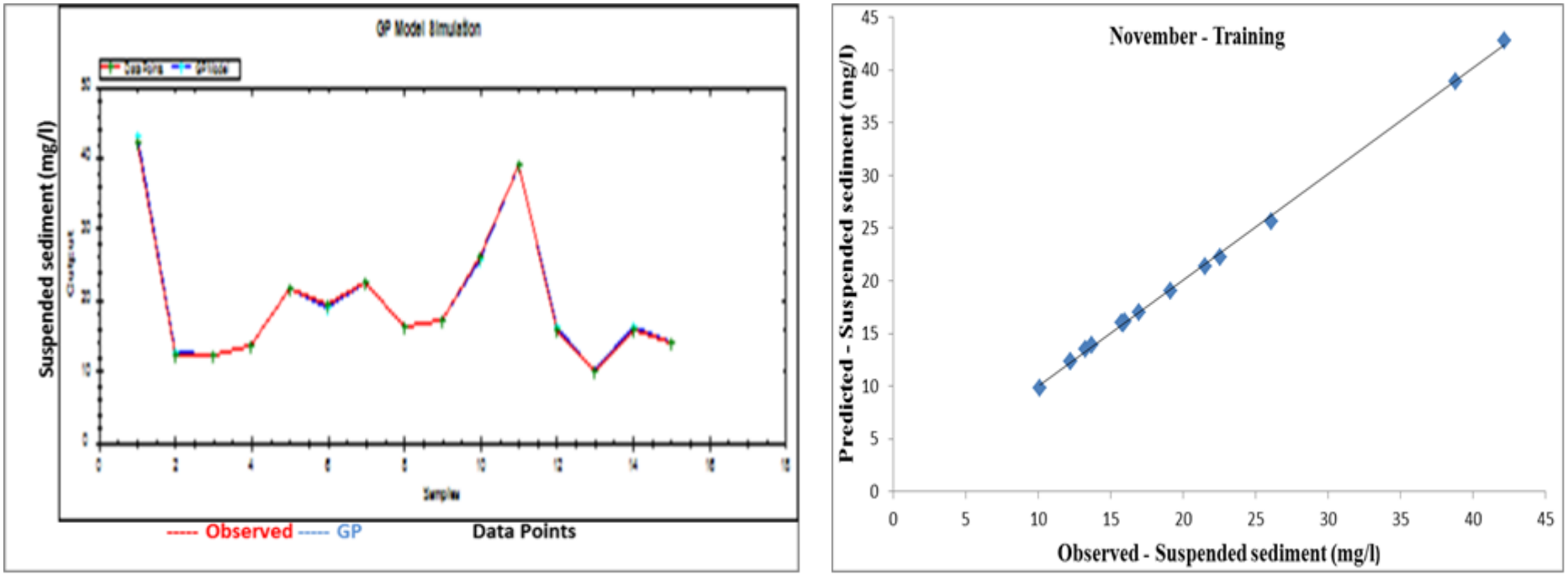


Figure 32: Plots of observed and GP-predicted suspended sediment (mg/l) for November during training phase

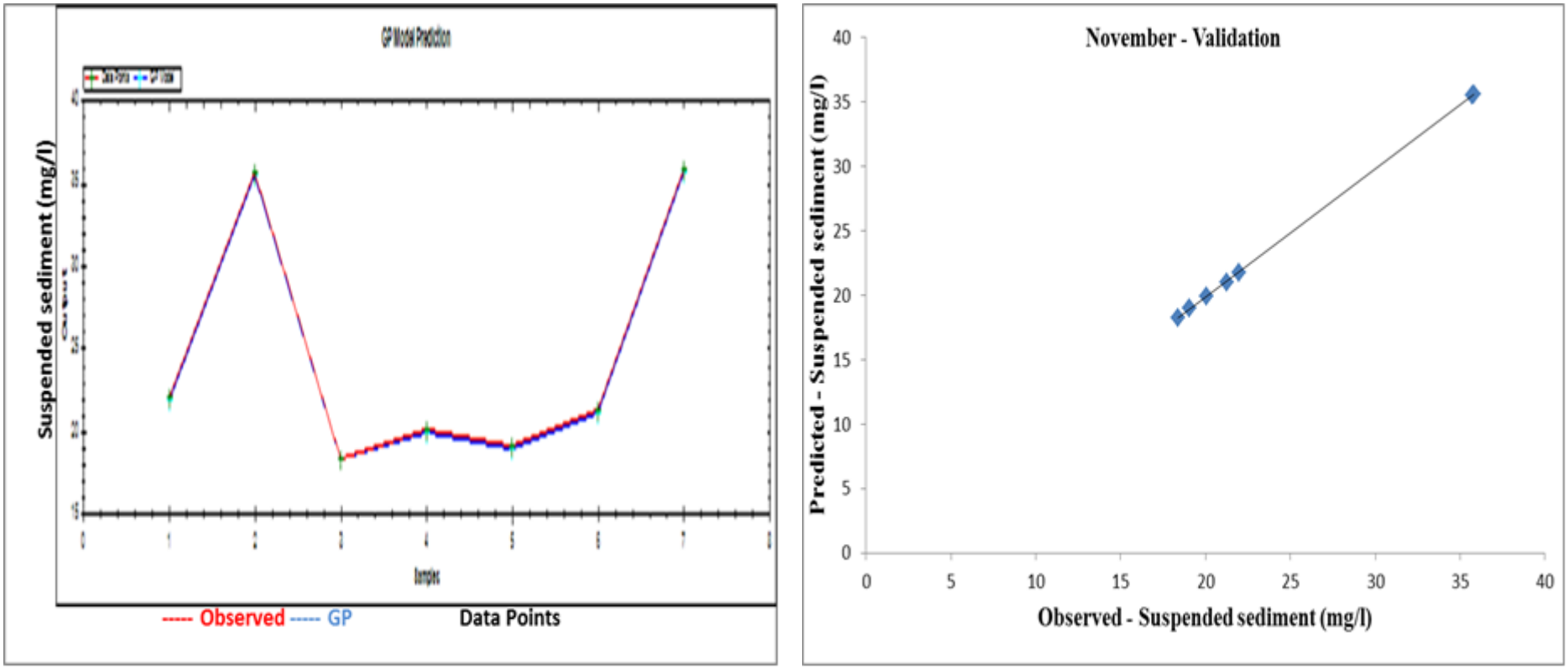


Figure 33: Plots of observed and GP-predicted suspended sediment (mg/l) for November during validation phase

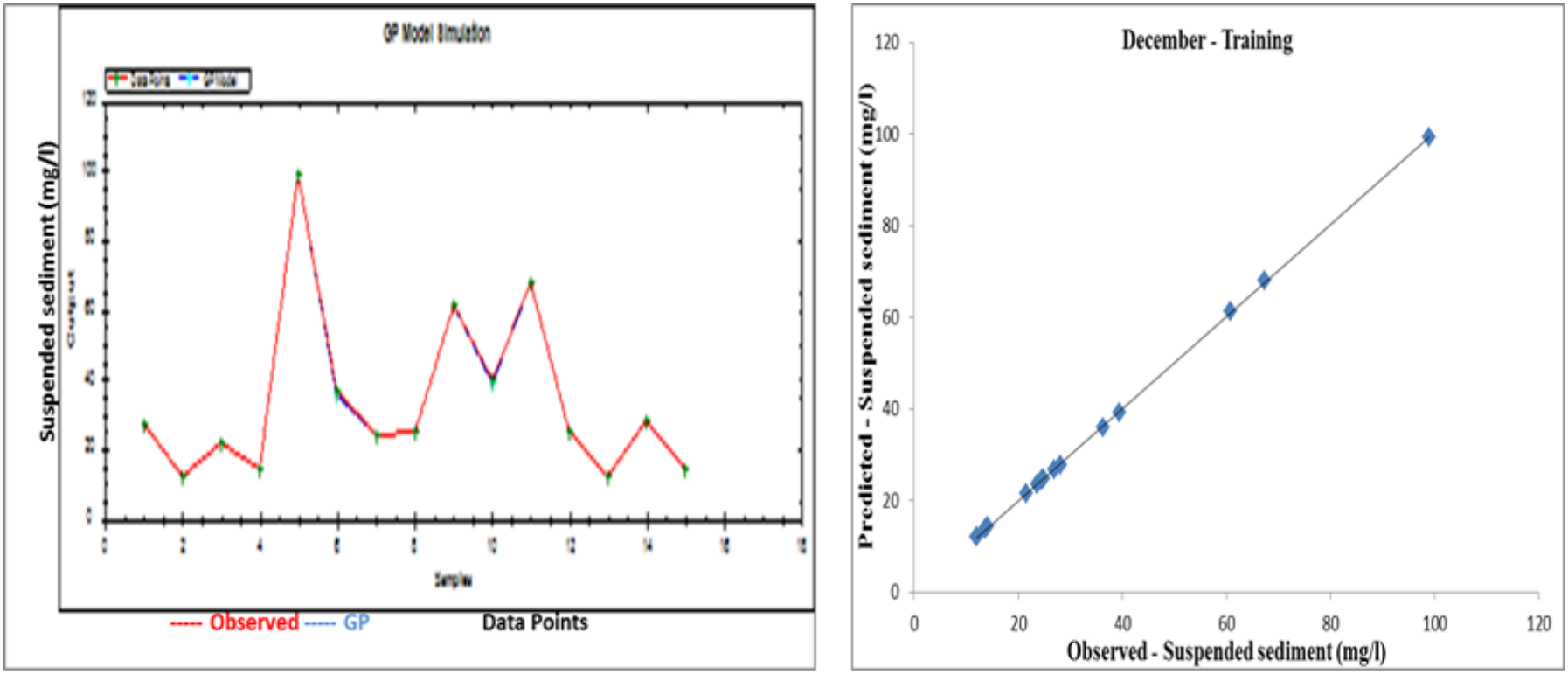


Figure 34: Plots of observed and GP-predicted suspended sediment (mg/l) for December during training phase

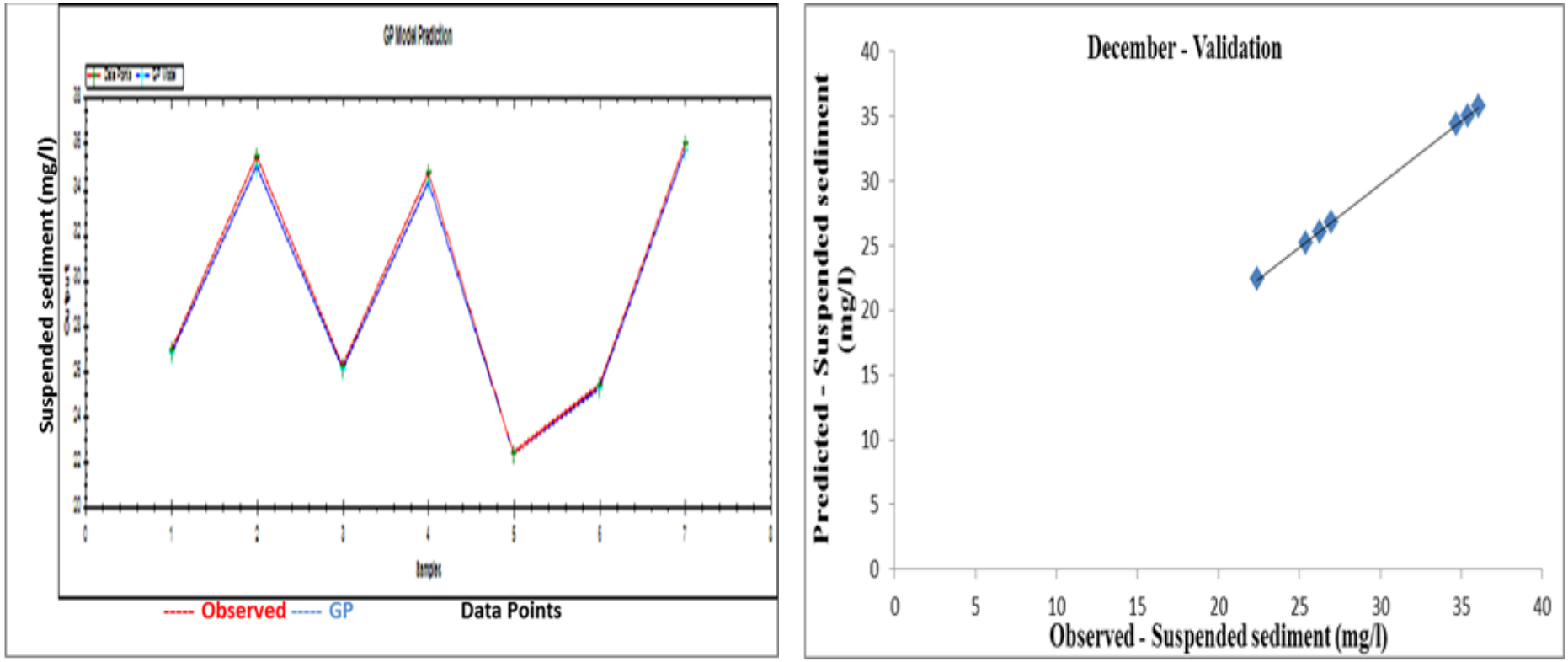


Figure 35: Plots of observed and GP-predicted suspended sediment (mg/l) for December during validation phase

It can be easily seen from Figures 12-35 that the developed monthly models were very accurate in the prediction of suspended sediment flowing into Inanda Dam with no significant difference between the measured and observed suspended sediment values. The models were also able to predict low, intermediate and high concentrations of suspended sediment. These confirm the ability of GP to predict extreme conditions and to recognise patterns (Londhe and Charhate 2010). The scatter plots between the predicted suspended sediment values and observed suspended sediment also indicated a close fit to the line of equality in all the models.

These figures also show that the average quantity of suspended sediment flowing into Inanda Dam every month is heterogeneous hence it is important to investigate this monthly variability. This monthly variability may be due to the difference in monthly streamflow values in Inanda Dam (Gay *et al.* 2014). It may also be due to the changes in land cover, rainfall intensity, land use and crop rotation in the Inanda Catchment throughout the year (Cerdan *et al.* 2010). In all the months in figures 12 to 35 the developed models present a positive relationship between suspended sediment and streamflow. It can be seen that the models in the autumn months predicted high quantity of suspended sediment. This is followed by models in the summer months, spring months and winter months. This variation may also be as a result of the remobilisation of the sediment stored in the Umgeni River due to streamflow and climatic changes (Navratil *et al.* 2010). This confirms that the quantities of suspended sediment flowing into Inanda Dam not only depend on streamflow values but also on the time of the year when there are heavy rains and availability of sediment. Though the figures show a trend in the variation of the monthly suspended sediment values flowing into Inanda Dam due to variation in its physical properties, it can still be viewed as a homogeneous when compared with larger rivers or dams (Gay *et al.* 2014).

3.6 CONCLUSION

The results from this study show that a genetic programming approach can be used to accurately predict the relationship between the streamflow and the suspended sediment load flowing into Inanda Dam. This result is in agreement with the results from the study by Sirdari *et al.* (2012) where the GP model accurately predicted the quantity of bed load transported in the Kurau River, Malaysia. The developed monthly models were accurate, simple and explicit, proving that GP can be applied in other areas of water resource modelling. The twelve developed monthly models (the best model in each month of the year)

produced a significantly low difference when the observed suspended sediment load was compared with the predicted suspended sediment load. The average R^2 values and RMS error for the 12 developed models were 0.999669 and 0.356623 respectively during the validation phase. The models were also able to replicate extreme hydrological events like predicting low and high suspended sediment load flowing into the dam. Consequently, the results are very promising and encourage the use of GP in predicting the nonlinear and dynamics of suspended sediment and streamflow patterns flowing into Inanda Dam. Furthermore, GP has been successfully used in the field of water resources management. It was used by Guven (2009) to predict daily stream flow in Schuylkill River, USA. It was also used by Wang *et al.* (2009) to predict monthly streamflow using long term observations. So GP techniques which incorporate inaccuracy and noise in the input data can satisfactory develop accurate robust models in the forms of symbolic functions to predict unknown information (Kumar *et al.* 2014).

However the use of data driven models like GP models in conjunction with knowledge based models to interpret complex hydrological processes could lead to the development of more accurate models and the better management of water resources. Hence, the use of hybrid models like the combination of artificial intelligence models such as Genetic Programming (GP) and Extended Kalman Filter (EKF) for the prediction of suspended sediment will also help to reduce the risk associated with online forecasting for water resources managers. This approach will be very effective in the forecasting of complicated temporal phenomena (Nasseri *et al.* 2011). In the nearest future it will be appropriate to develop models that is able to quantify the quantity of suspended sediment flowing into other dams in other catchments in South Africa and also to quantify suspended sediment from different tributaries into the Inanda Dam (Meshgi *et al.* 2015).

CHAPTER 4

SEDIMENT RATING CURVE FOR SUSPENDED SEDIMENT PREDICTION AND PERFORMANCE COMPARISON BETWEEN GP AND SRC

4.1 INTRODUCTION

The study of sediment, including its load estimation and prediction in reservoirs, has been a major focus for hydrologists, sedimentologists, engineers and scientists in general for many years (Horowitz 2003). Their areas of interest include the anthropogenic effects on sediment (Syvitski and Milliman 2007) and the effect of changes in sediment on aquatic life (Gao and Wang 2008). According to Leahy *et al.* (2008) the study of rivers and their particulates is very important for reliable prediction, but this task is very difficult due to the inherent nonlinearity and complexity of hydrological systems. Therefore, the study of suspended sediment flowing into reservoirs, estuaries or oceans is an important aspect of water resources management. Most of the sediment flowing into a reservoir is suspended sediment with bed load accounting for between 1% to 10% (Meade *et al.* 1990; Asselman 2000). Its measurement involves a good sample method which is usually very expensive and lengthy. According to Khanchoul *et al.* (2012) the quantity of suspended sediment flowing into a reservoir at a particular time is a product of two quantities, namely the discharge and the concentration of sediment in the water flowing into the reservoir at that particular time. In most cases the streamflow is measured continuously but its concentration measurement is not that frequent. So, if the two quantities are not measured frequently, the estimation of suspended sediment during storms may be missed or substantial errors may occur (Khanchoul *et al.* 2012). Hence, considerable efforts have been made by hydrologists and researchers to apply a rating curve to describe the empirical relationship between the suspended sediment concentration and water discharge at a particular location (Syvitski *et al.* 2000; Morehead *et al.* 2003). It is therefore necessary to develop methods that can accurately predict the suspended sediment loads from continuous water data sets and infrequently or non-sampled rivers. One of the widely used methods using empirical models when or where measurements are not available, apart from artificial intelligence models, are rating curves (Horowitz 2003; Khanchoul and Jansson 2008; Cherif *et al.* 2009). Therefore, this chapter presents the use of

sediment rating curves for suspended sediment prediction into Inanda Dam in KwaZulu Natal province of South Africa. A comparative study of the results of using GP for the same study is done.

4.2 METHODOLOGY

4.2.1 Sediment rating curve (SRC)

Sediment rating curve techniques have been used by researchers where there is no continuous set of data to estimate the quantity of suspended sediment load in a river (Horowitz 2008). According to Harrington and Harrington (2013), the quantity of suspended sediment transported in a river over a period of time can be calculated using:

$$LS = \int_{t2}^{t1} Q\text{tSSCtdt} \quad (6)$$

Where Ls is the quantity of suspended sediment in g (gram) flowing during a period between time t_1 and time t_2 ($t_2 - t_1$), Qt is the streamflow in m^3 /s at time t, SSCt is quantity of suspended sediment in mg/l at time t. The equation is commonly expressed in log space as (Kişi 2007):

$$\log SSC = \log a + b \log Q \quad (7)$$

The equation can be expressed/transformed to a normal space in the form of a power curve (Ferguson 1987) as:

$$SSC = 10^{\alpha}Q^{\beta} \quad (8)$$

This transformation of the variables from log space to normal space result in the introduction of statistical bias which can be reduced by multiplying the predicted suspended sediment load by a correction factor (CF) (Olyaie *et al.* 2015) expressed as:

$$CF = \exp(2.651s^2) \quad (9)$$

where s^2 is the mean square error of the log-transformed regression (in $\log_{10}$ units). The introduction of CF is not a perfect solution to eliminating the bias so Cohn and Gilroy (1991) suggested the use of a nonparametric smearing factor which was recommended by (Duan 1983), but Asselman (2000) suggested the fitting of a power curve to the normal data, resulting in a regression of the form:

$$SSC = aQ^b \quad (10)$$

a and b being the constants of the non-linear power regression.

Equation (8) and Equation (10) are two different equations. Equation (8) is a multiplication model because an increase in streamflow, Q, will result in double increase in suspended sediment concentration, while equation (10) is an additive model because an increase in streamflow, Q, will result in a proportional increase in suspended sediment concentration (SSC). However, according to Jansson (1985) both models can produce similar regression of suspended sediment load if streamflow is the independent variable and the primary cause variable, as in this study, and also when the data points are on or close to the regression line. More details on the suitability of these methods for the prediction of suspended sediment load in a river are presented by Harrington and Harrington (2013). In this study the suitability of equation 10 is assessed for the monthly prediction of suspended sediment load flowing into Inanda Dam.

4.3 IMPLICATIONS AND APPLICABILITY

Hydrologists have used sediment rating curves to predict and determine long-term suspended sediment load where there is an absence of actual suspended sediment load (SSL) measurements (Boukhrissa *et al.* 2013). In this study, the historical average monthly suspended sediment load (SSL) and streamflow measurements for 14 years from 1990 to 2013 were used for regression analyses and also to determine the relationship between the average monthly suspended sediment load (SSL) in mg/l and the average monthly streamflow (Q) in m^3/s using the power function rating curve (eqn 10). The results from the monthly rating curves are illustrated in Table 7 and are validated by comparing the predicted against observed values on scatter plots as illustrated in Figures 36 to 47. The values of a and b are determined from data through the linear regression between (log SSL) and (log Q) (Kişi 2007).

Table 7: The form of the equations and the best fit of the monthly rating relationships developed for Inanda Dam

Months	a	b	R^2	RMSE
January	9.9656	0.8130	0.9953	4.290848898
February	9.8847	0.8261	0.9940	11.88517045
March	9.6027	0.8399	0.9962	9.608239280
April	11.268	0.7616	0.9944	3.608530400
May	13.673	0.6276	0.9907	1.308945824
June	14.966	0.5342	0.9918	0.397089212
July	13.343	0.6762	0.9880	2.595326539
August	14.435	0.5879	0.9918	2.595326539
September	14.172	0.6033	0.9869	1.229191764
October	13.918	0.6100	0.9833	1.456225833
November	12.863	0.6752	0.9900	1.648834234
December	11.138	0.7658	0.9873	8.369357336
Average	12.43575	0.6934	0.990808	4.082757000

a, b= constant and coefficient in equation $SSC= aQ^b$

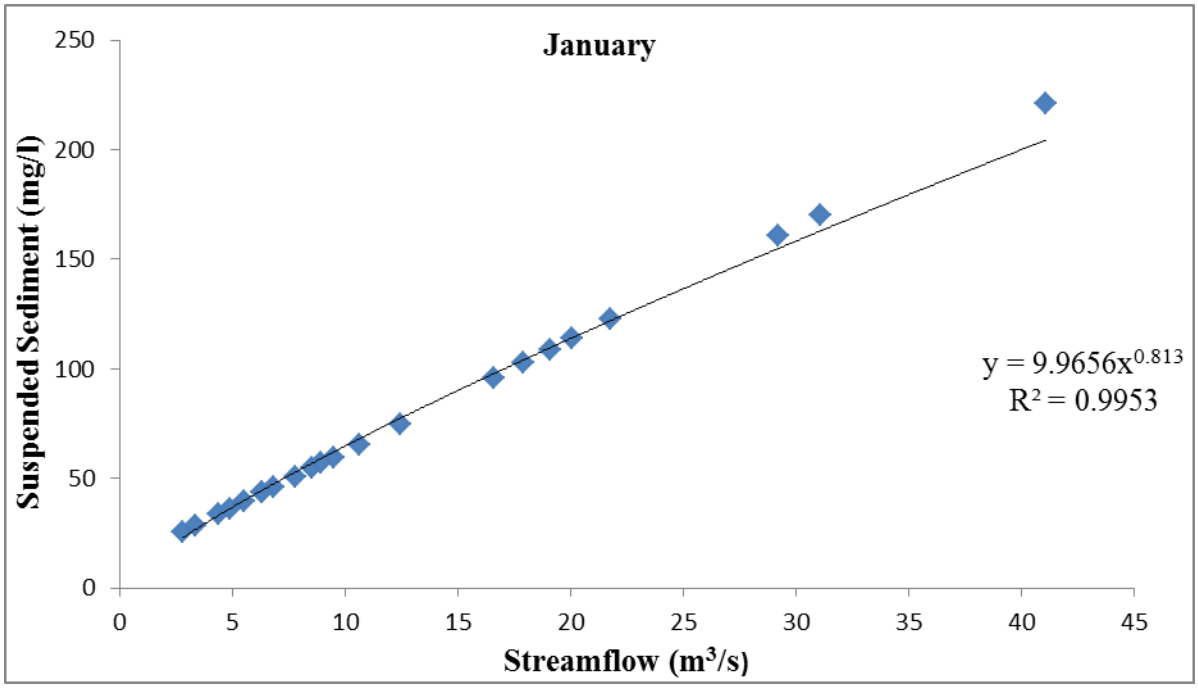


Figure 36: Sediment rating curves at Inanda Dam for January

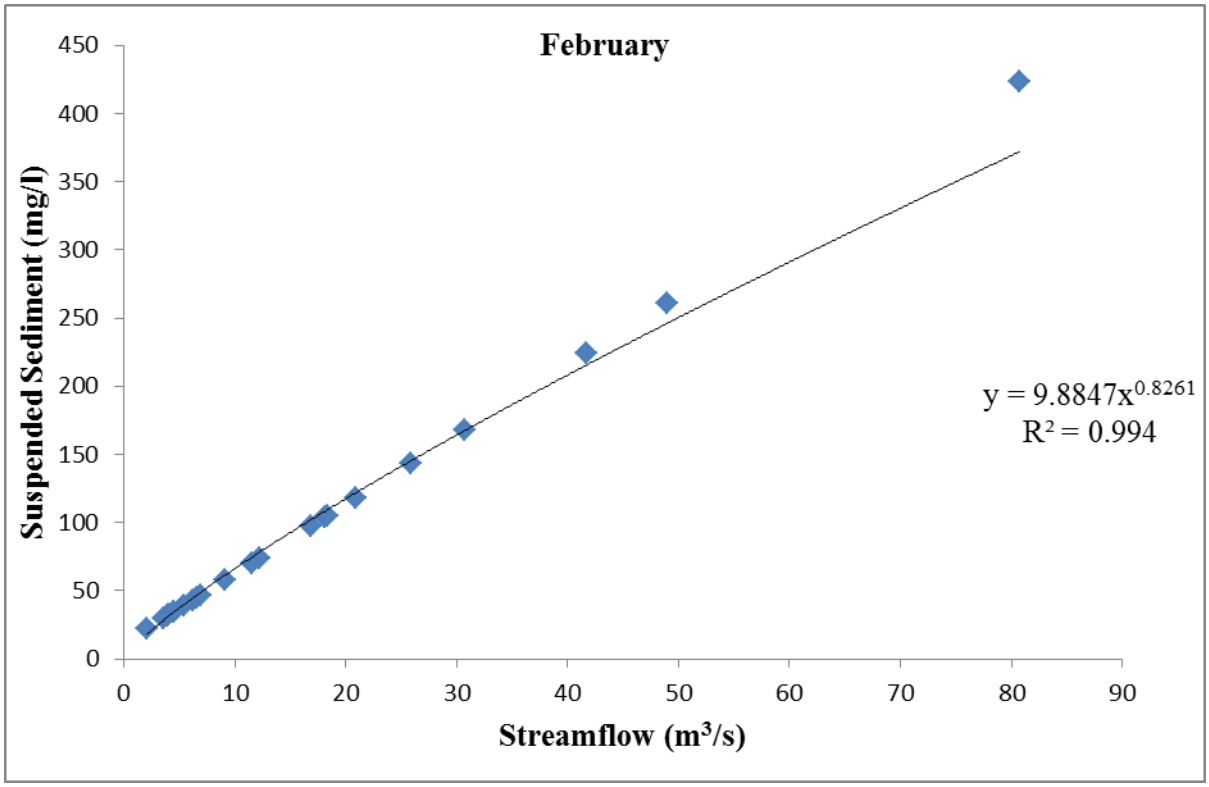


Figure 37: Sediment rating curves at Inanda Dam for February

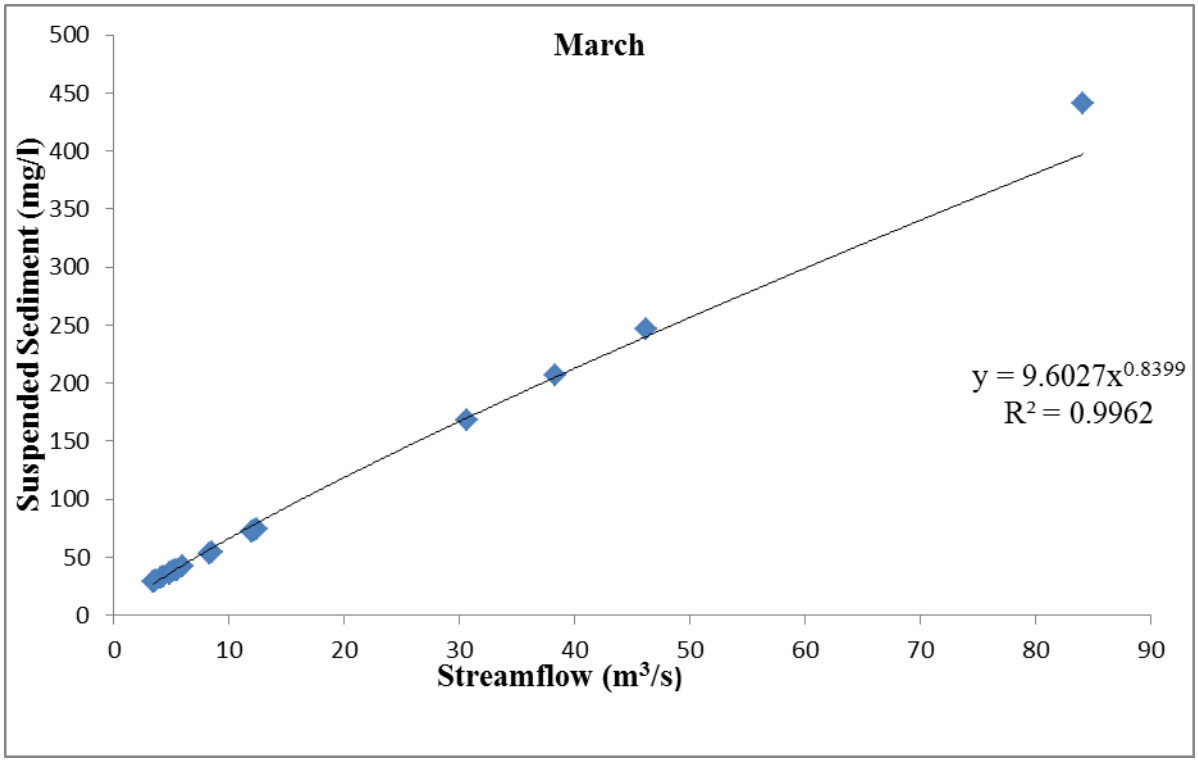


Figure 38: Sediment rating curves at Inanda Dam for March

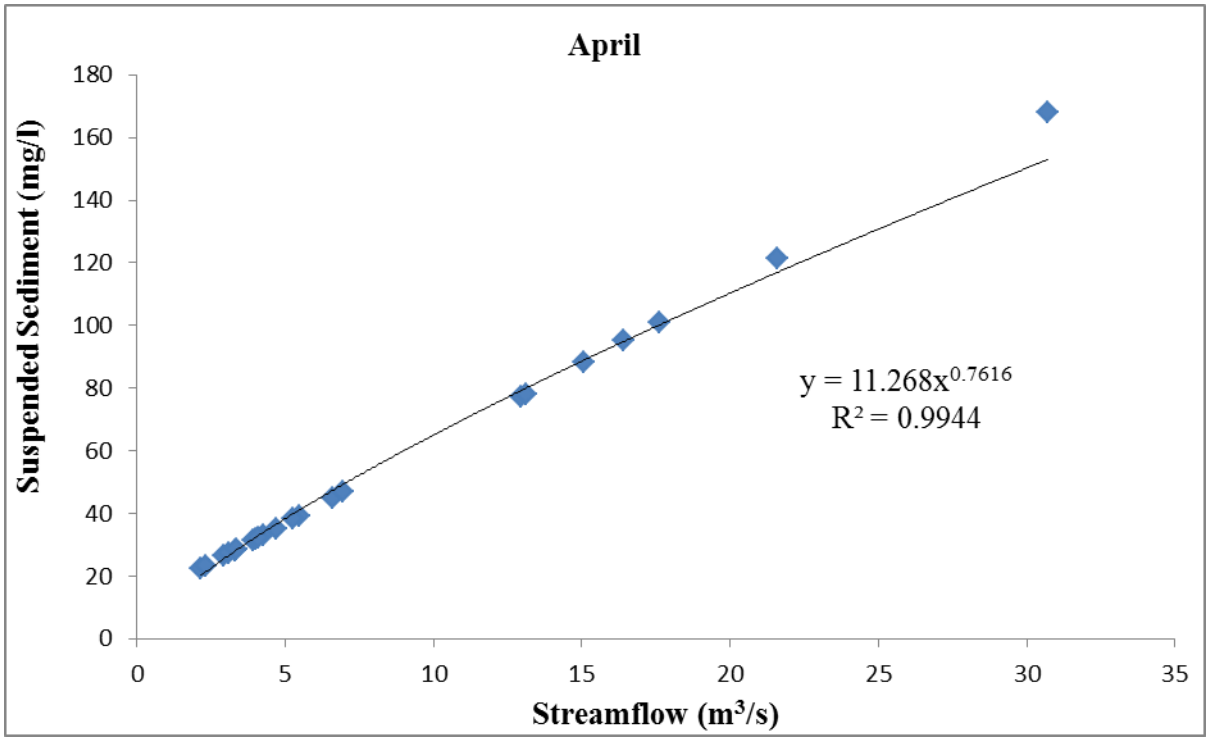


Figure 39: Sediment rating curves at Inanda Dam for April

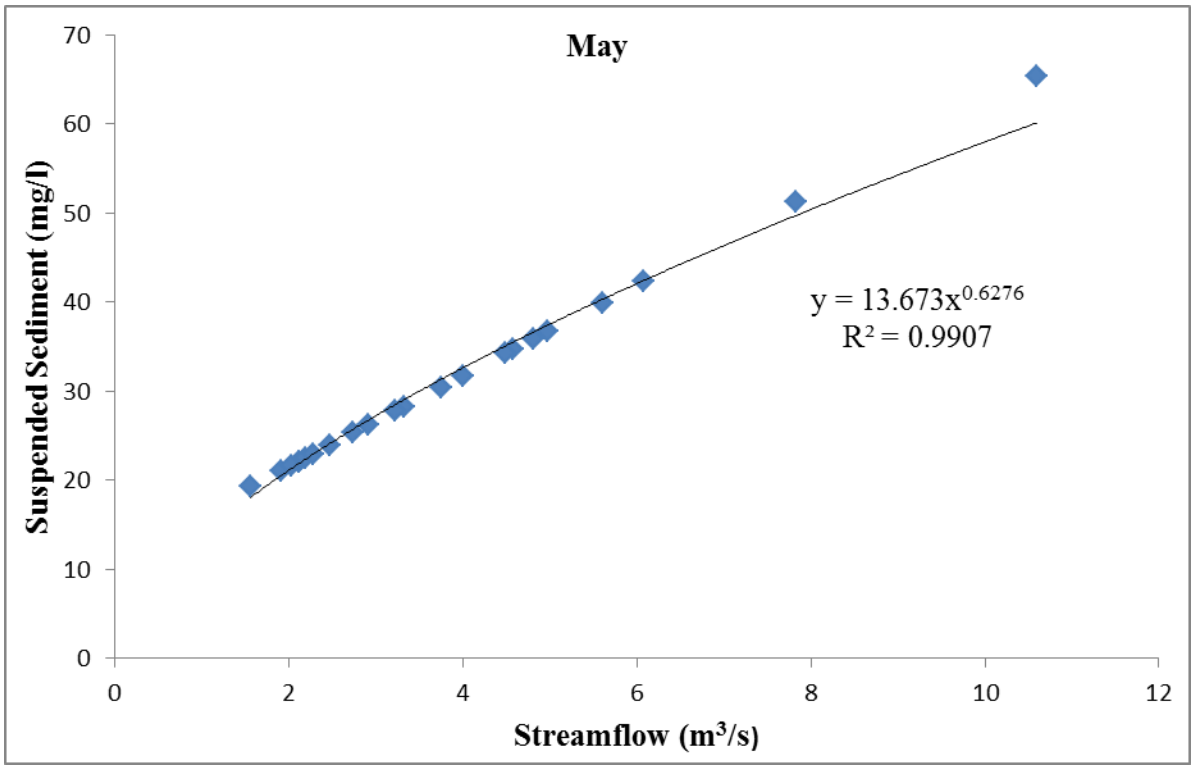


Figure 40: Sediment rating curves at Inanda Dam for May

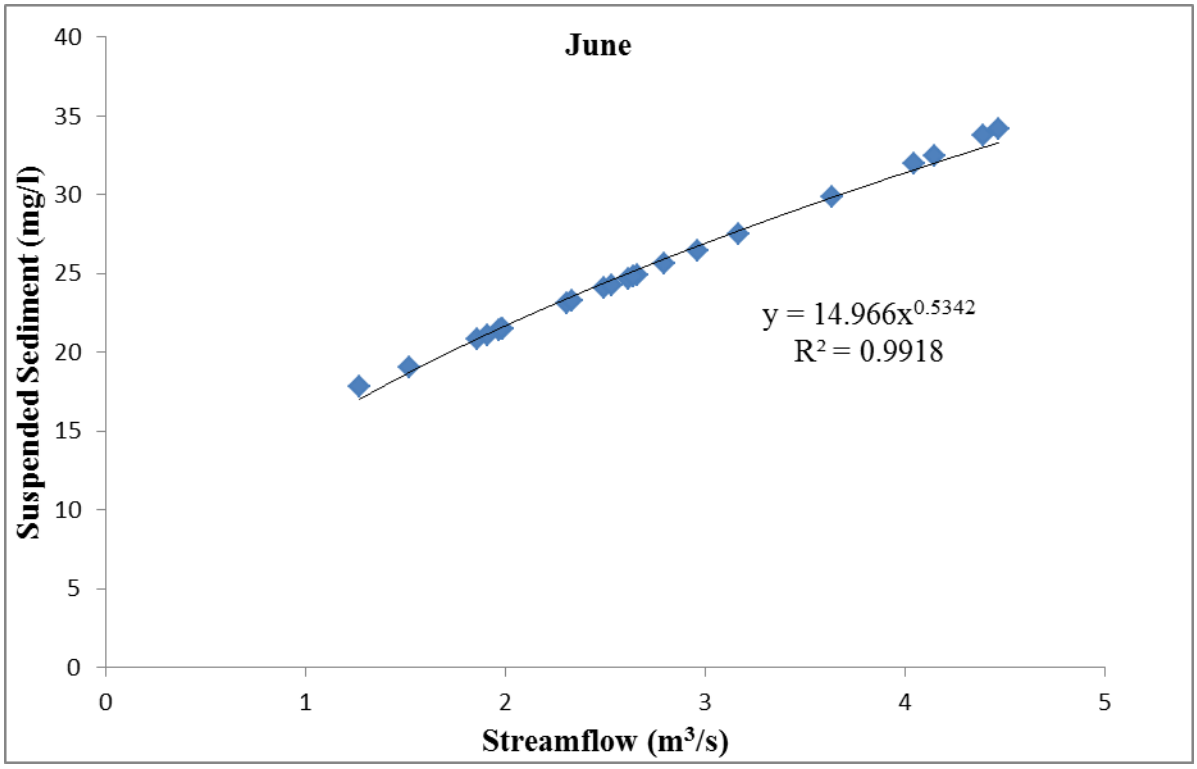


Figure 41: Sediment rating curves at Inanda Dam for June

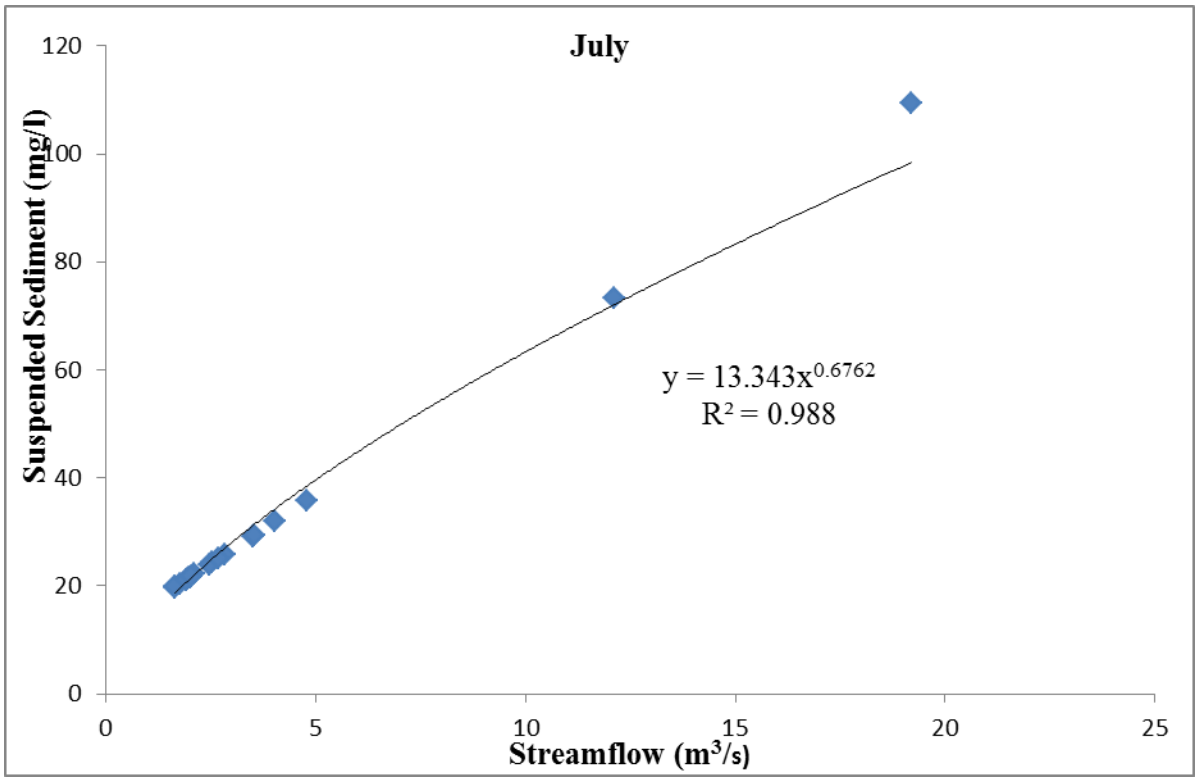


Figure 42: Sediment rating curves at Inanda Dam for July

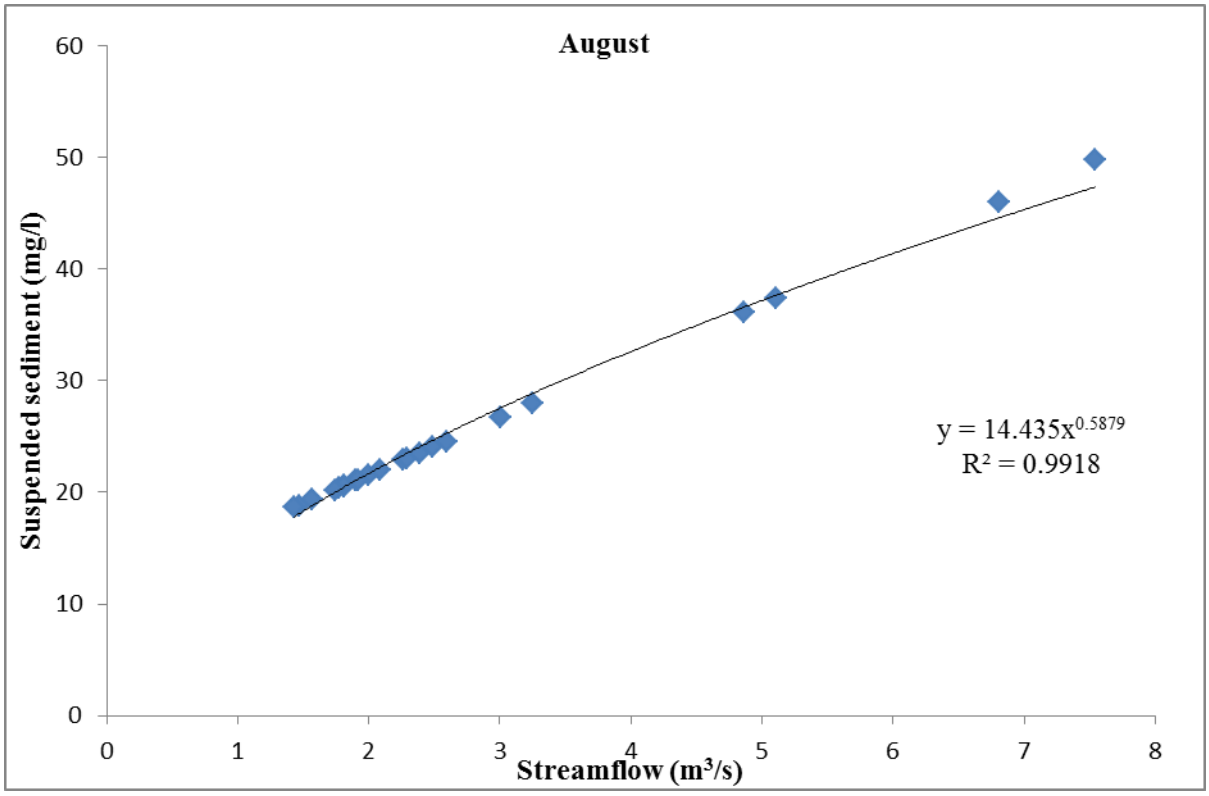


Figure 43: Sediment rating curves at Inanda Dam for August

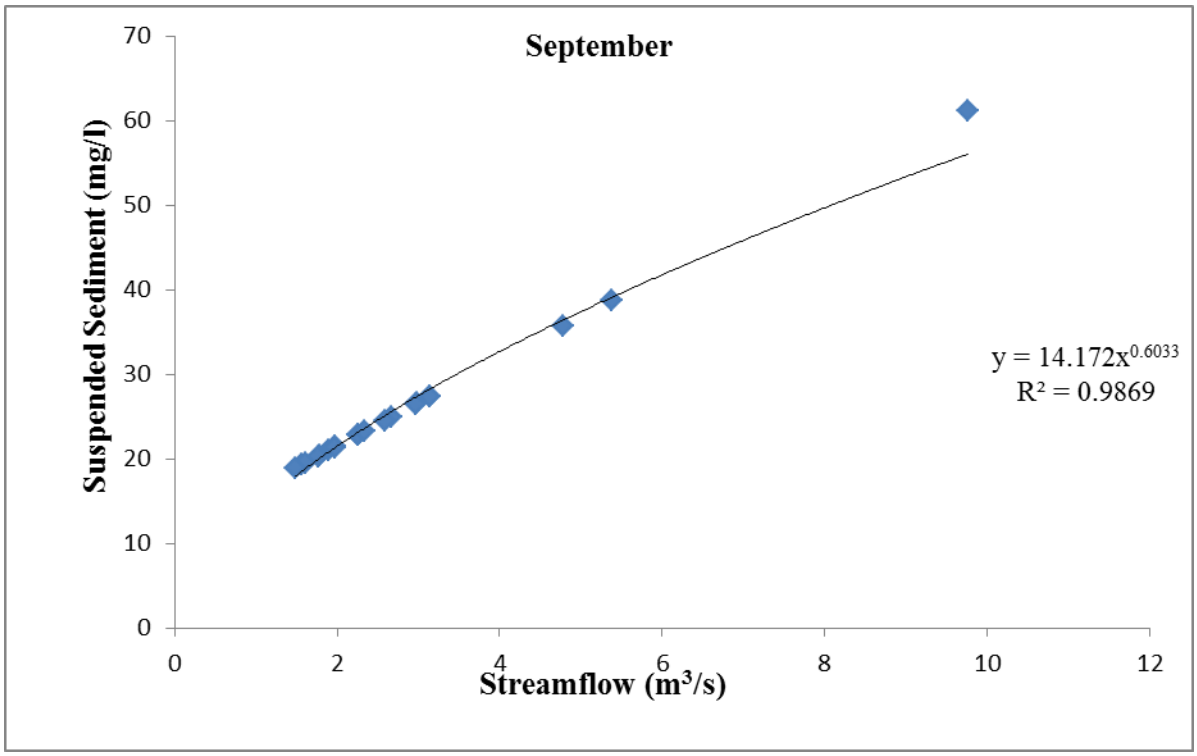


Figure 44: Sediment rating curves at Inanda Dam for September

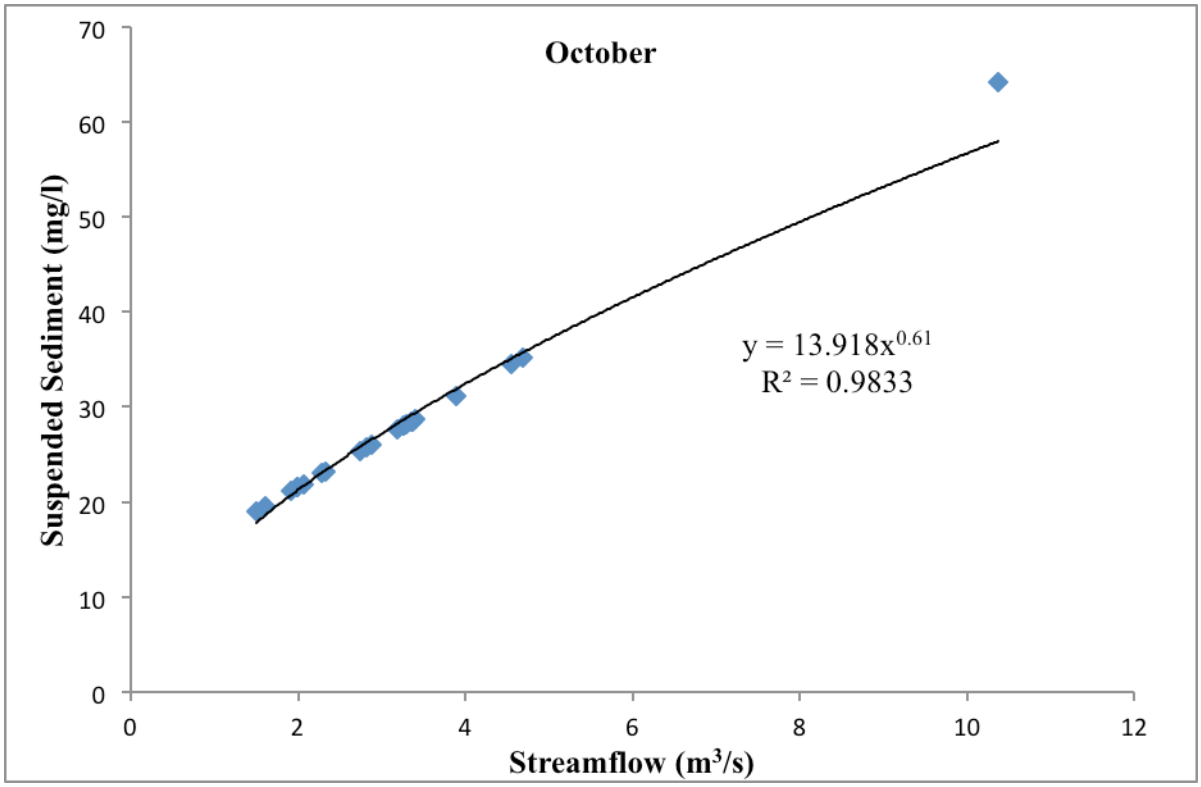


Figure 45: Sediment rating curves at Inanda Dam for October

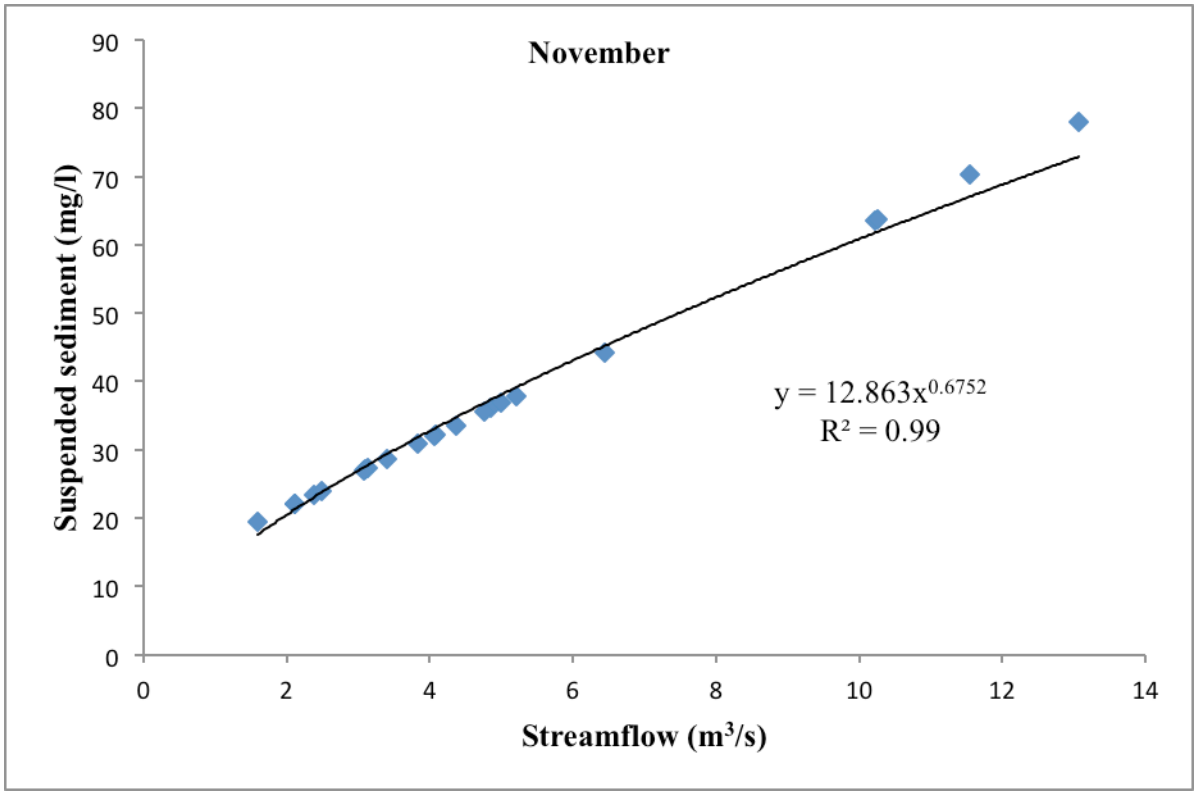


Figure 46: Sediment rating curves at Inanda Dam for November

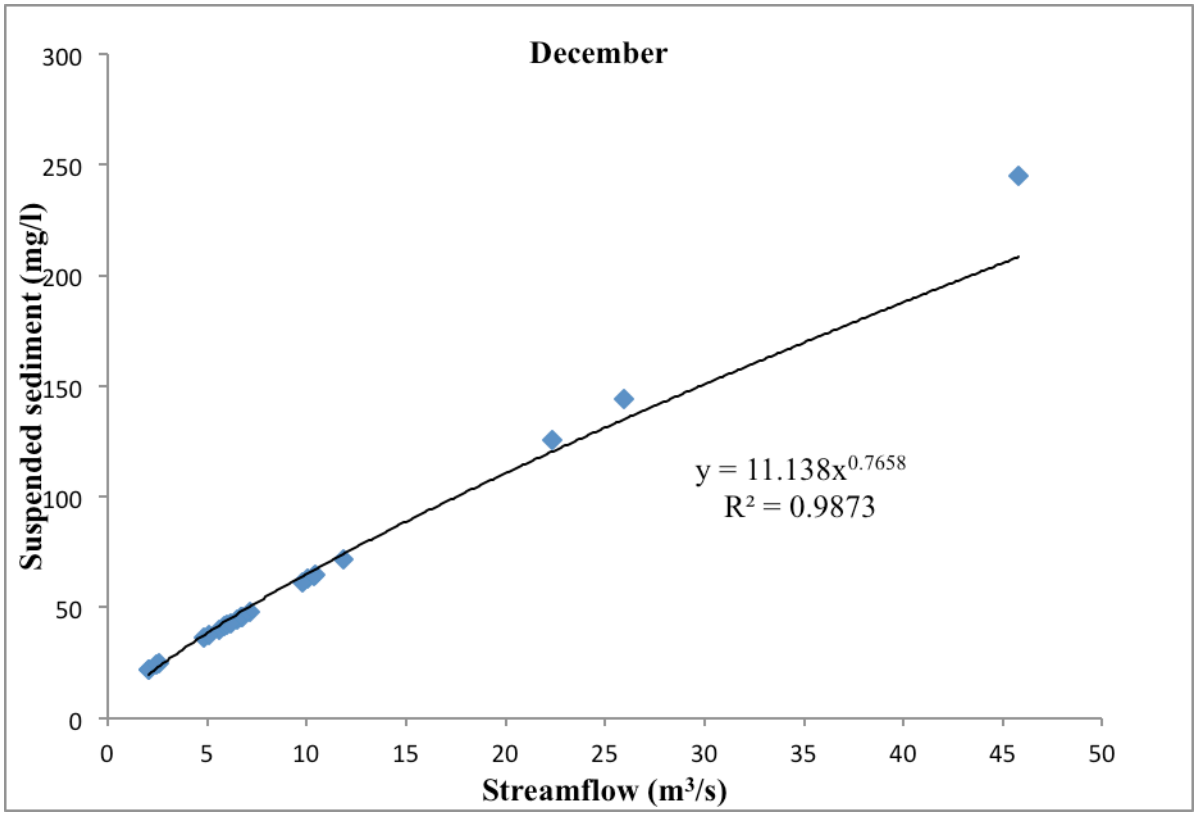


Figure 47: Sediment rating curves at Inanda Dam for December

4.3.1 Interpretation of sediment rating

The monthly sediment rating curves developed using average suspended sediment concentration (SSC) in mg/l and average streamflow (Q) in m^3/s for all available data from upstream of Inanda Dam were presented in Figures 36 to 47. In all the sediment rating curve equations developed in Figures 36 to 47, the average suspended sediment concentration (SSC) in mg/l is represented as y while the average streamflow (Q) in m^3/s is represented as x. The best fit of the theoretical functions of the sediment rating curve equations shows RMSE values from 0.397089212 to 11.88517045 and R^2 values from 0.9833 to 0.9962. From the best-fit power function lines through the data, it can be observed that the curves underestimated the suspended sediment concentrations at a high streamflow. This indicated that the rating curves did not capture the complex behaviour of suspended sediments well, so the rating curves were unable to predict the suspended sediment concentration at peak flow events. This is in agreement with the results obtained by Boukhrissa *et al.* (2013). However, at low and medium streamflow the rating curves were fairly accurate and give a better

understanding of the relationship between suspended sediment concentration and streamflow upstream of Inanda Dam.

As previously stated in Chapter 2, the values of ‘a’ and ‘b’ of the regression coefficients of the rating curves indicate the following characteristics of the Umgeni River flowing into Inanda Dam: the erosive capacity of the area, the presence of sediment in the catchment, the ability of the river to transport sediment, the erosive ability of the river and the availability of new sources of sediment in the area (Asselman 2000). The high values of ‘a’ coefficients in all the monthly sediment rating curve models indicate that the Inanda Catchment is characterized by the presence of weatherable materials and the low values of ‘b’ coefficients indicate a low erosive power of the river according to Morgan (2009). The steepness of the rating curves which are a combination of the ‘a’ coefficients and ‘b’ coefficients indicate soil erosivity and erodibility. The steepness of the sediment rating curve due to high ‘a’ values and low ‘b’ values also implies that a low streamflow will transport a low quantity of suspended sediment and at high streamflow there will be a large quantity of suspended sediment transported (Heng and Suetsugi 2014). It also implies that the erosive power of the dam is high or there is a new source of sediment available when the water level of the river is high. The steepness of the sediment rating curves in Figures 36 to 47 could also be as a result of weirs in the dam. At low streamflow little suspended sediment will be transported by the river because most of it would have settled behind the weirs. But at high streamflow the settled sediment will be flushed out thereby greatly increasing the quantity of suspended sediment flowing into the dam (Asselman 2000).

4.4 RESULTS AND DISCUSSION

From the results of this study it was observed that sediment rating curves can be used to predict historical missing data of the quantity of suspended sediment flowing into Inanda Dam using existing streamflow datasets. Likewise, it can also be used to predict the quantity of suspended sediment that will flow into the dam. However, the inaccuracies from the use of these rating curves to predict suspended sediment load could be increased when the developed sediment rating curve models were applied to suspended sediment and streamflow conditions which are significantly different from the conditions from which they were developed. Therefore, in this study, 12 sediment rating curves were developed, one for each month of year, so that the developed models can be an accurate representation of a particular month. From the results in Figures 36 to 47, it can be seen that accurate sediment rating curve

models were developed with low RMSE values of between 0.397089212 and 11.88517045 and high R^2 values of between 0.9833 and 0.9962. This shows that the rating curves were able to capture the relationship between the average monthly suspended sediment load and streamflow in Inanda Dam. The rating curve models predicted accurate suspended sediment at low, medium and high stream flows. This implies that the majority of the suspended sediment flowing into the dam was at high streamflow events (Fan *et al.* 2012). It can be concluded therefore, that the large quantity of suspended sediment flowing into Inanda Dam could be as a result of water impaired by impurities (sediment) from upstream activities like industry, agriculture, and urbanization (Kumar and Series 2012). The mining of sand close to the inflow of the dam on the banks of the Mshazi and Umgeni rivers may also increase the quantity of suspended sediment flowing into the dam.

As illustrated in Figures 36 to 47 and Table 7, all the developed monthly sediment rating curves were evaluated using the same performance evaluation criteria namely RMSE and R^2 for equal bases of comparison. The performance of these models through their RMSE and R^2 values were compared with those of the corresponding monthly developed GP models for each month of the year. The results of the performance evaluation of the models are presented in Table 8.

Table 8: RMSE and R^2 values for both GP and SRC

MONTHS	**GP**		**SRC**	
	R^2	**RMSE**	**R^2**	**RMSE**
JAN	0.999970	0.214396	0.9953	4.290848898
FEB	0.996507	2.382141	0.9940	11.88517045
MAR	0.999962	0.548428	0.9962	9.608239280
APR	0.999741	0.201124	0.9944	3.608530400
MAY	0.999999	0.032855	0.9907	1.308945824
JUN	0.999997	0.064152	0.9918	0.397089212
JUL	0.999978	0.066710	0.9880	2.595326539
AUG	0.999990	0.039383	0.9918	2.595326539
SEP	0.999992	0.148439	0.9869	1.229191764
OCT	0.999965	0.150845	0.9833	1.456225833
NOV	0.999967	0.173291	0.9900	1.648834234
DEC	0.999965	0.257714	0.9873	8.369357336
AVG	0.999669	0.356623	0.990808	4.082757000

The results show that the GP models have higher R^2 and smaller RMSE values. The results agree with the results of Aytek and Kisi (2008) where they developed an explicit relationship between daily suspended sediment and streamflow based on genetic programming. The GP model they developed was compared with multi-linear regression and sediment rating curves models and their result show that the GP model performed best and it was easy to use. This is an indication that the GP models perform better than the SRC models. Also, when the observed and estimated suspended sediment load from both the GP models and its corresponding SRC models were compared, as illustrated in both the hydrographs in Figures 48 to 59 and the scatterplots in Figures 60 to 71, the results show that the predicted suspended sediment loads from the GP models were very close to the measured suspended sediment loads in all the months.

The results also show that all the SRC models could not predict the measured suspended sediment loads, especially in peak periods at high stream flows. The prediction of the SRC in the month of September was very low in accuracy with R^2 value of 0.1083. According to Walling (1977) it is difficult to determine the actual cause of this scatter due to the interrelated relationship between hydrometerlogical phenomena. Generally, it could be due to inaccuracy in the laboratory and field measurement and the dynamic of sediment yield and erosion. It may also be as a result of the biased predictions at the time of data detransformation or the absence of distinct pattern in the trend of the relationship between suspended sediment load and streamflow (Sadeghi *et al.* 2008; Smith 2008).

The scatter plots in Figures 60 to 71 also show that the prediction of the GP models are closer to the exact fit line (measured/observed suspended sediment loads) than the prediction of the SRC models with R^2 values ranging from 0.9965 to 1 in each month of the year. Also the values of the 'a' and 'b' coefficients in the best straight line equations in all the scatterplots of the GP models are closer to 1 and 0 respectively when compared to those of the SRC models. These results indicate clearly that both the GP models and the SRC models can be used for the prediction of suspended sediment loads using previously measured streamflow values.

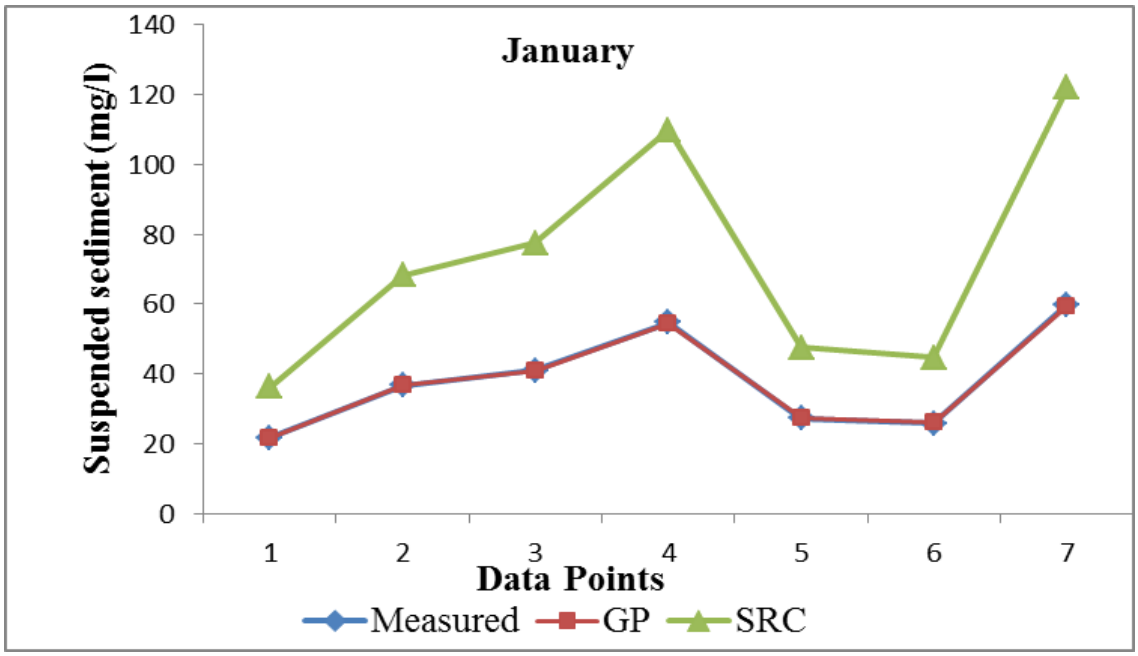


Figure 48: Comparison of measured, SRC models and GP models results in the validation phase for January

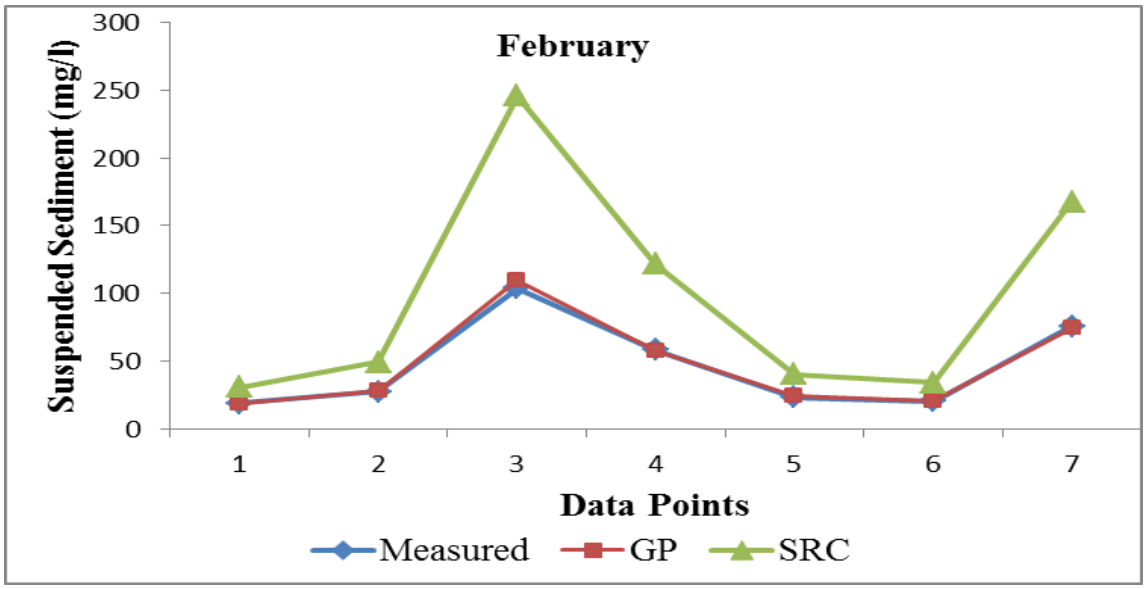


Figure 49: Comparison of measured, SRC models and GP models results in the validation phase for February

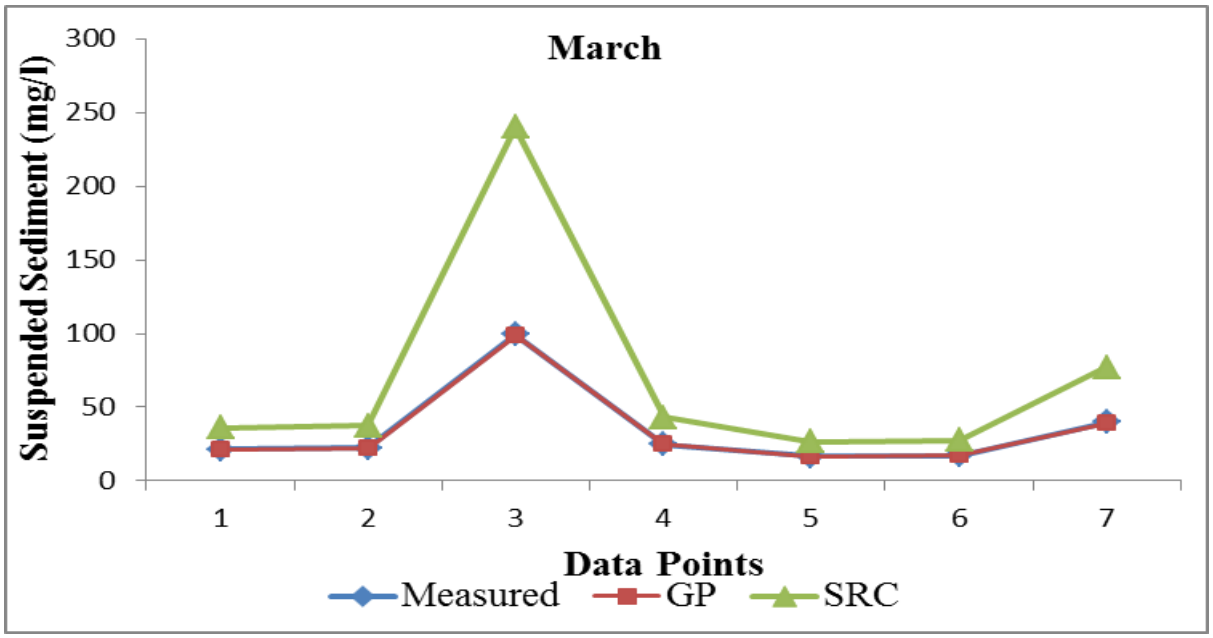


Figure 50: Comparison of measured, SRC models and GP models results in the validation phase for March

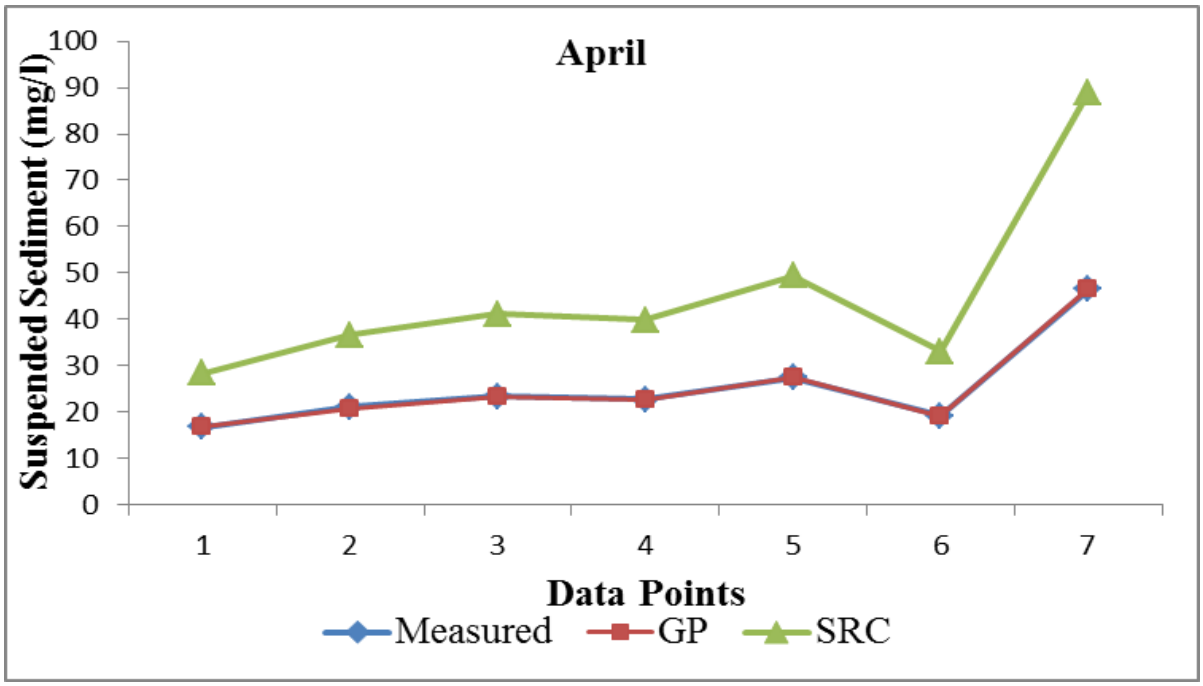


Figure 51: Comparison of measured, SRC models and GP models results in the validation phase for April

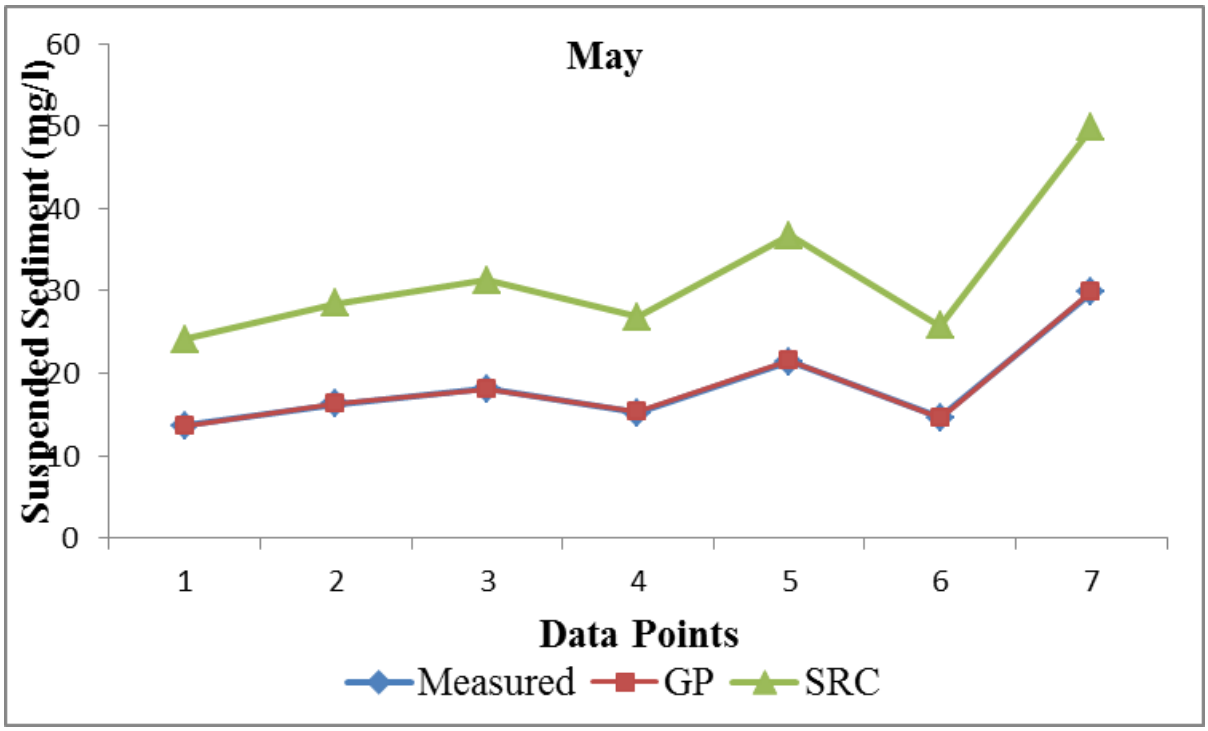


Figure 52: Comparison of measured, SRC models and GP models results in the validation phase for May

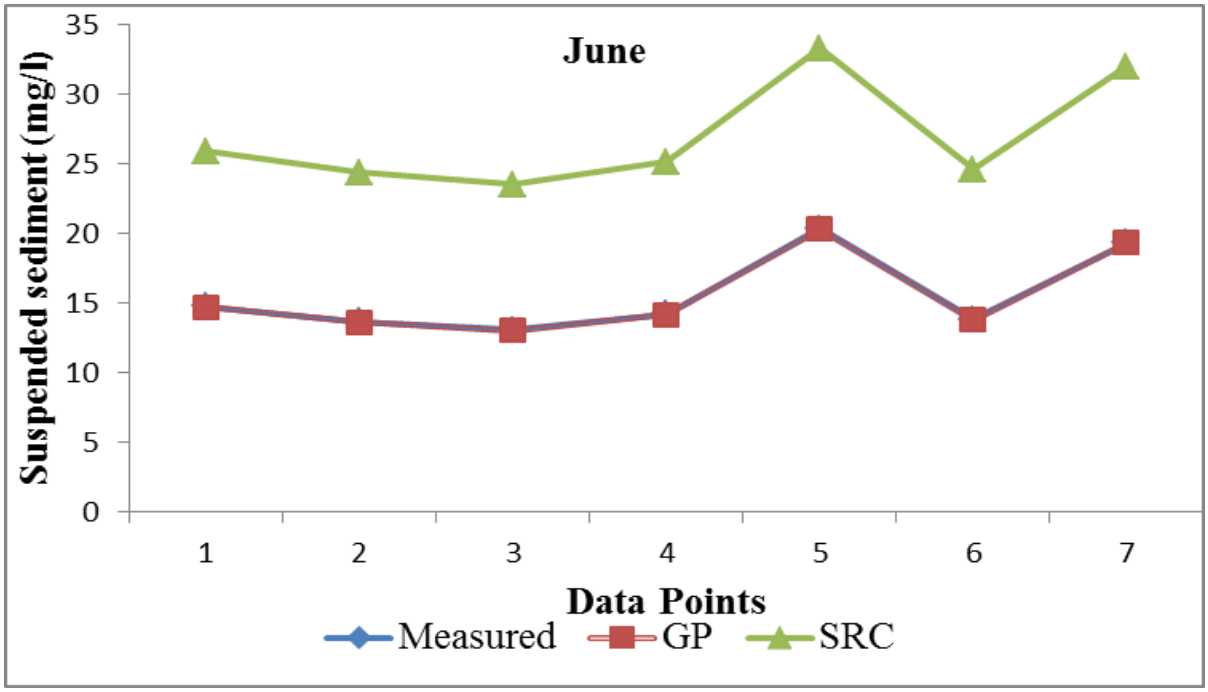


Figure 53: Comparison of measured, SRC models and GP models results in the validation phase for June

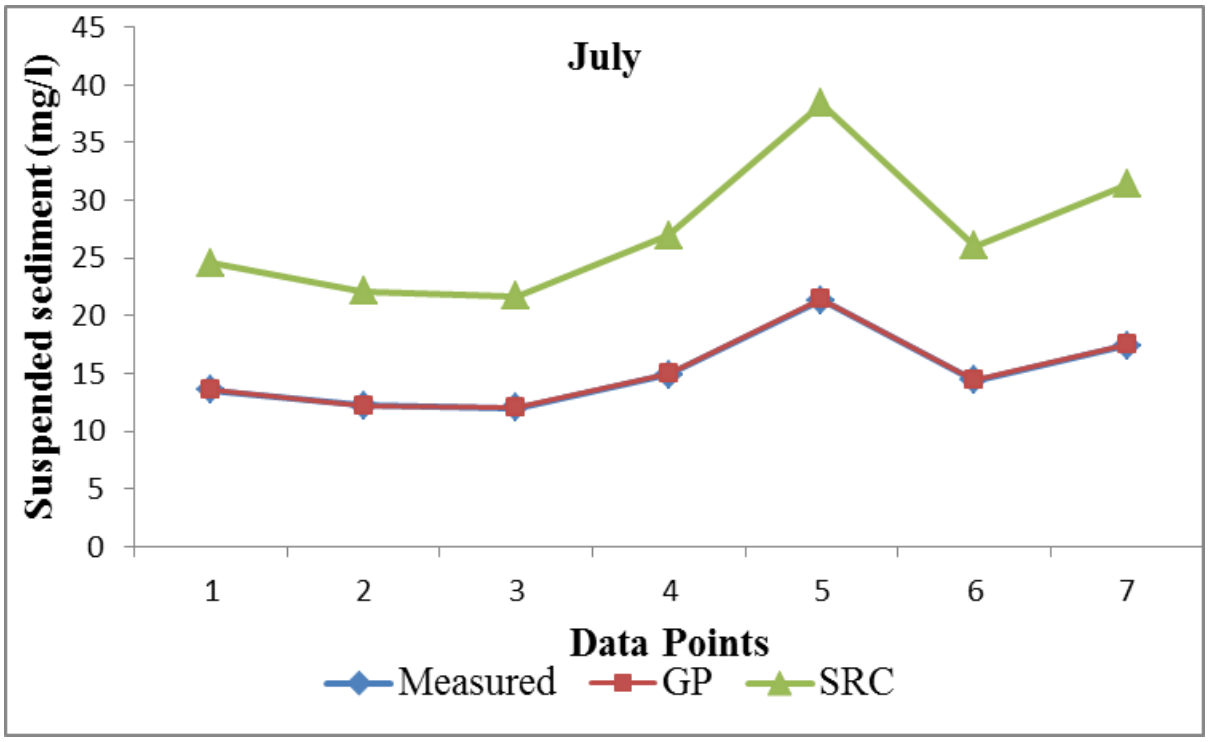


Figure 54: Comparison of measured, SRC models and GP models results in the validation phase for July

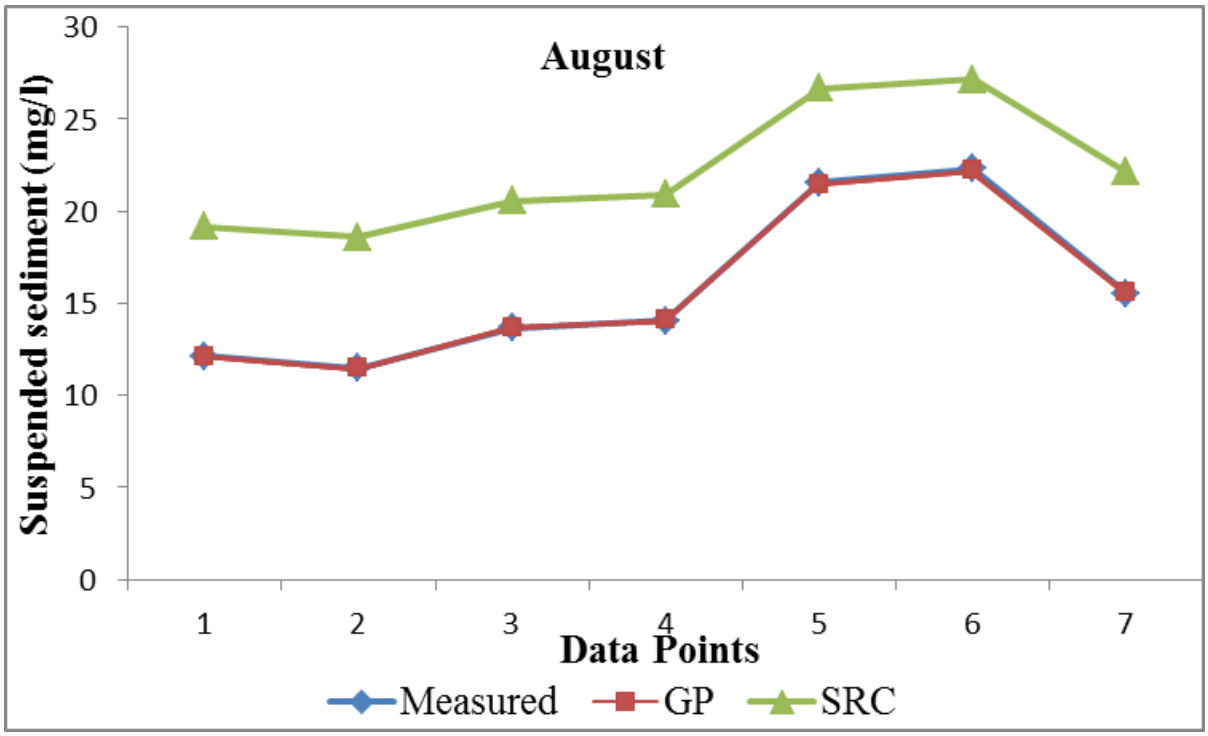


Figure 55 Comparison of measured, SRC models and GP models results in the validation phase for August

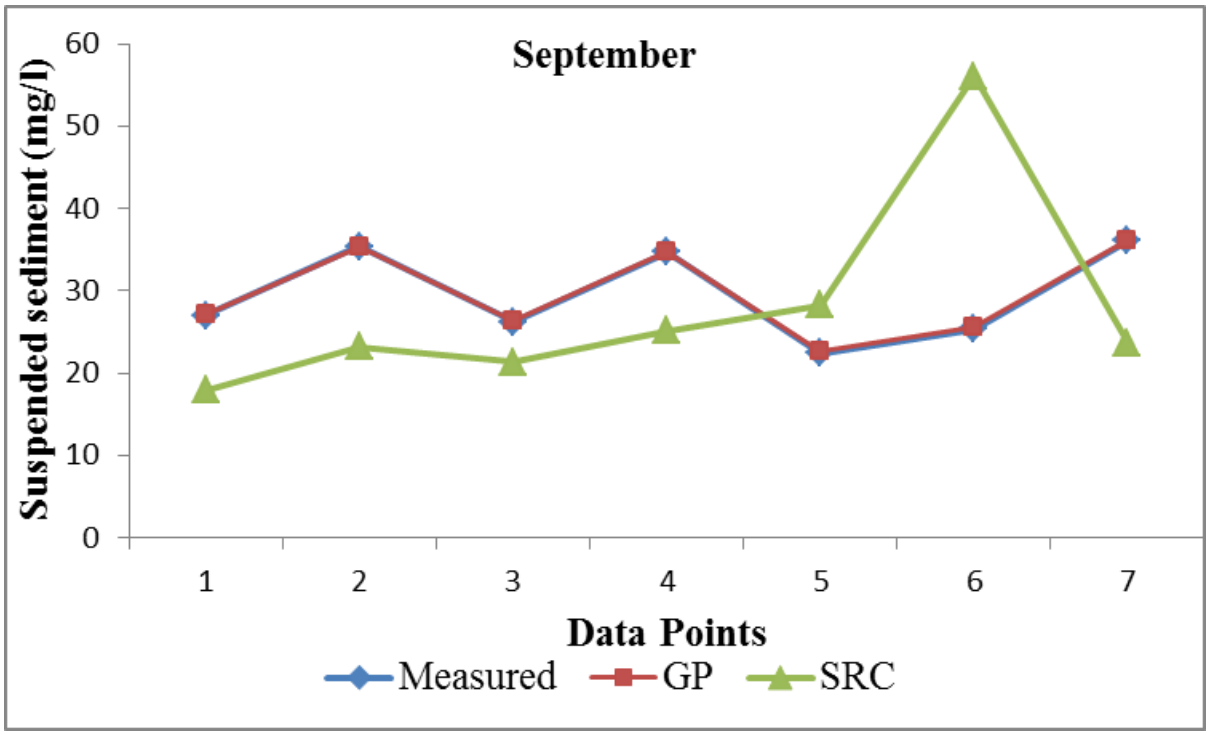


Figure 56: Comparison of measured, SRC models and GP models results in the validation phase for September

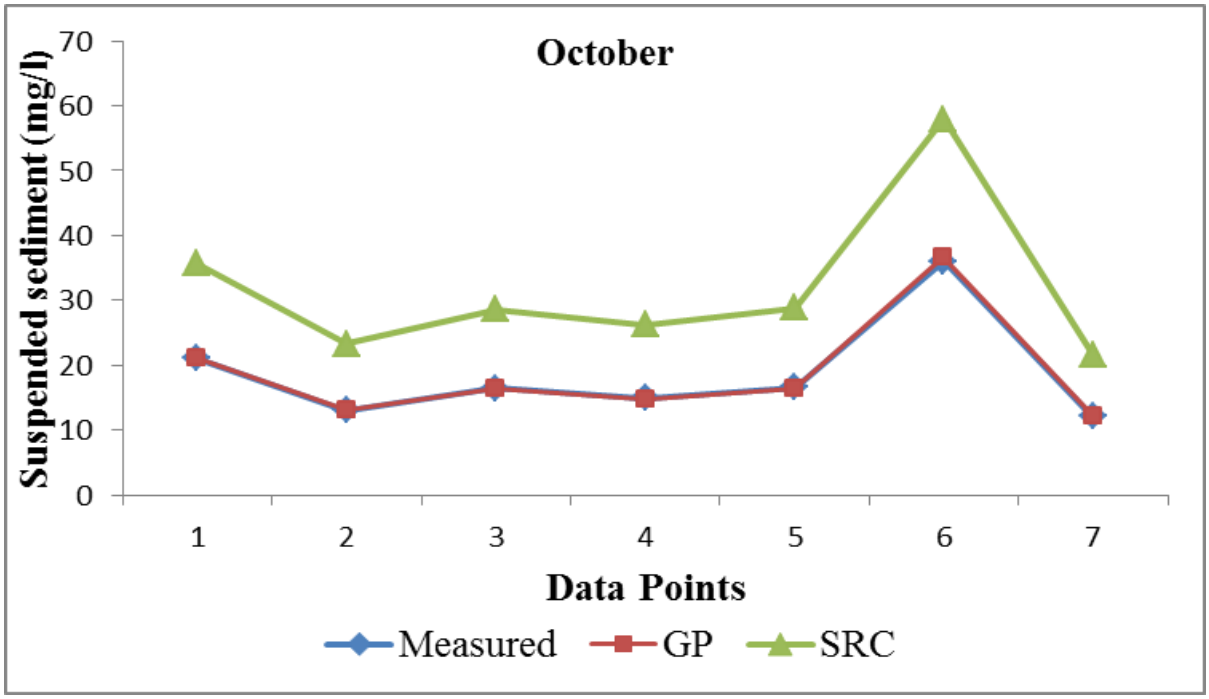


Figure 57: Comparison of measured, SRC models and GP models results in the validation phase for October

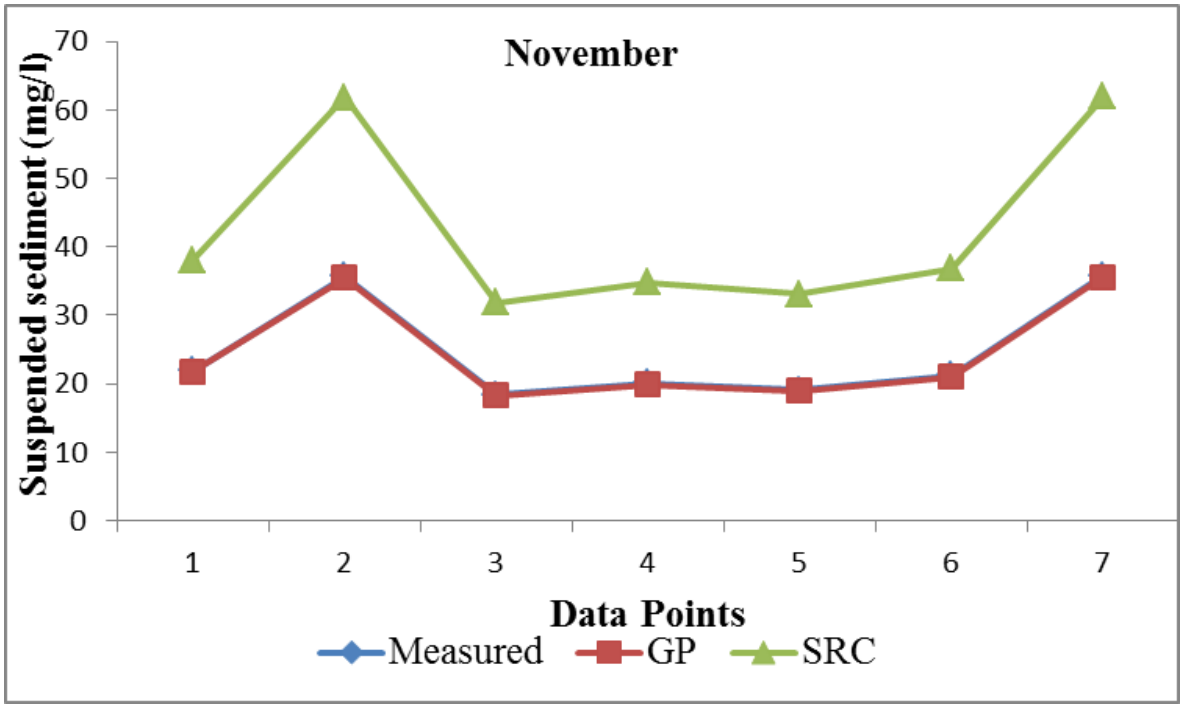


Figure 58 Comparison of measured, SRC models and GP models results in the validation phase for November

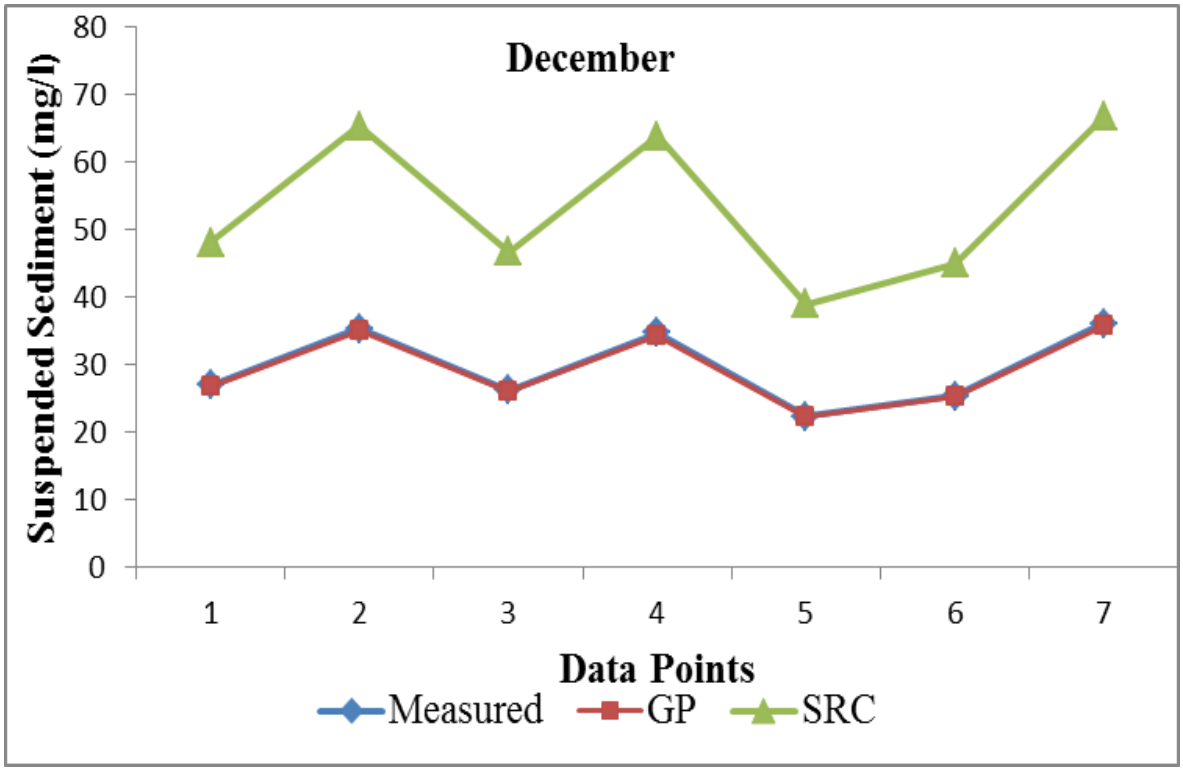


Figure 59: Comparison of measured, SRC models and GP models results in the validation phase for December

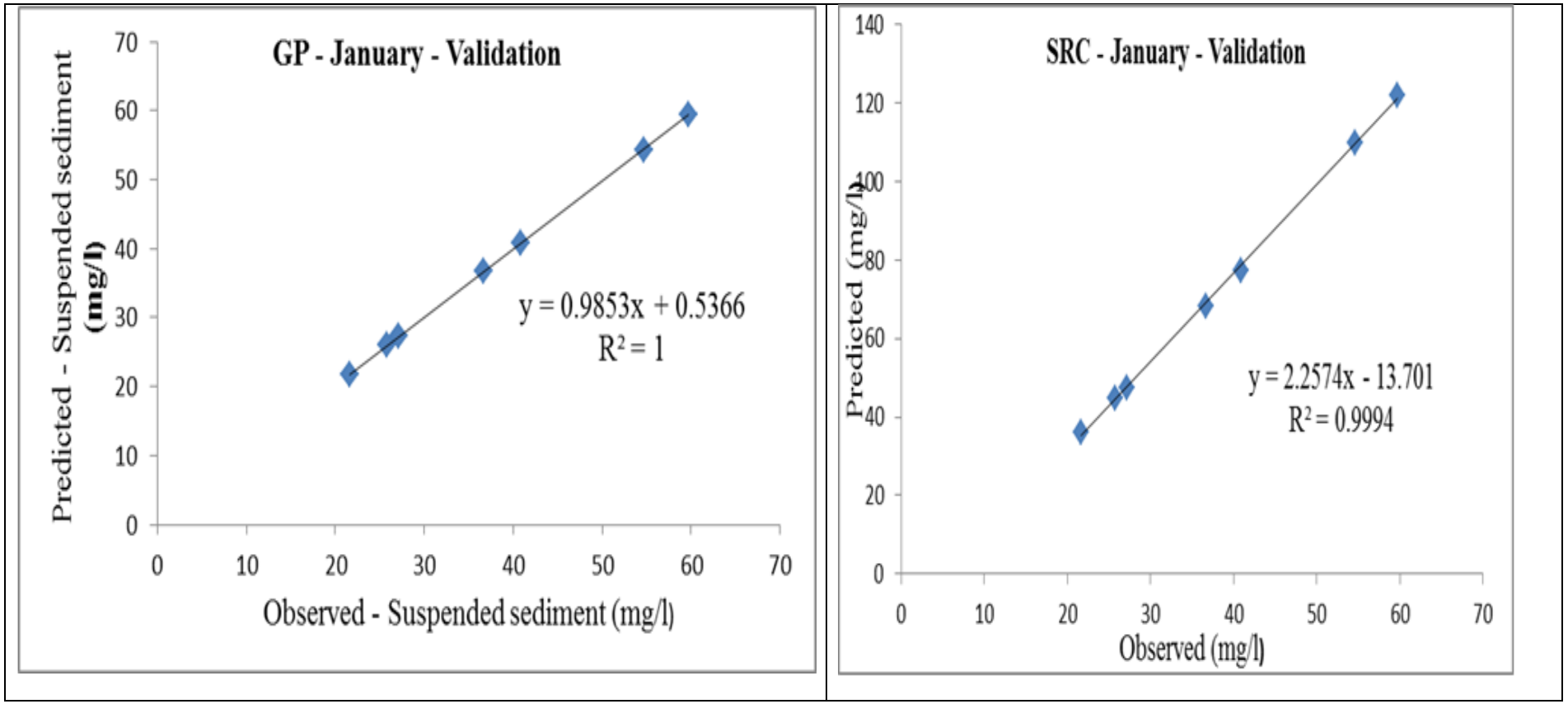


Figure 60: Observed and predicted suspended sediment load by the GP and SRC models in the validation phase for January

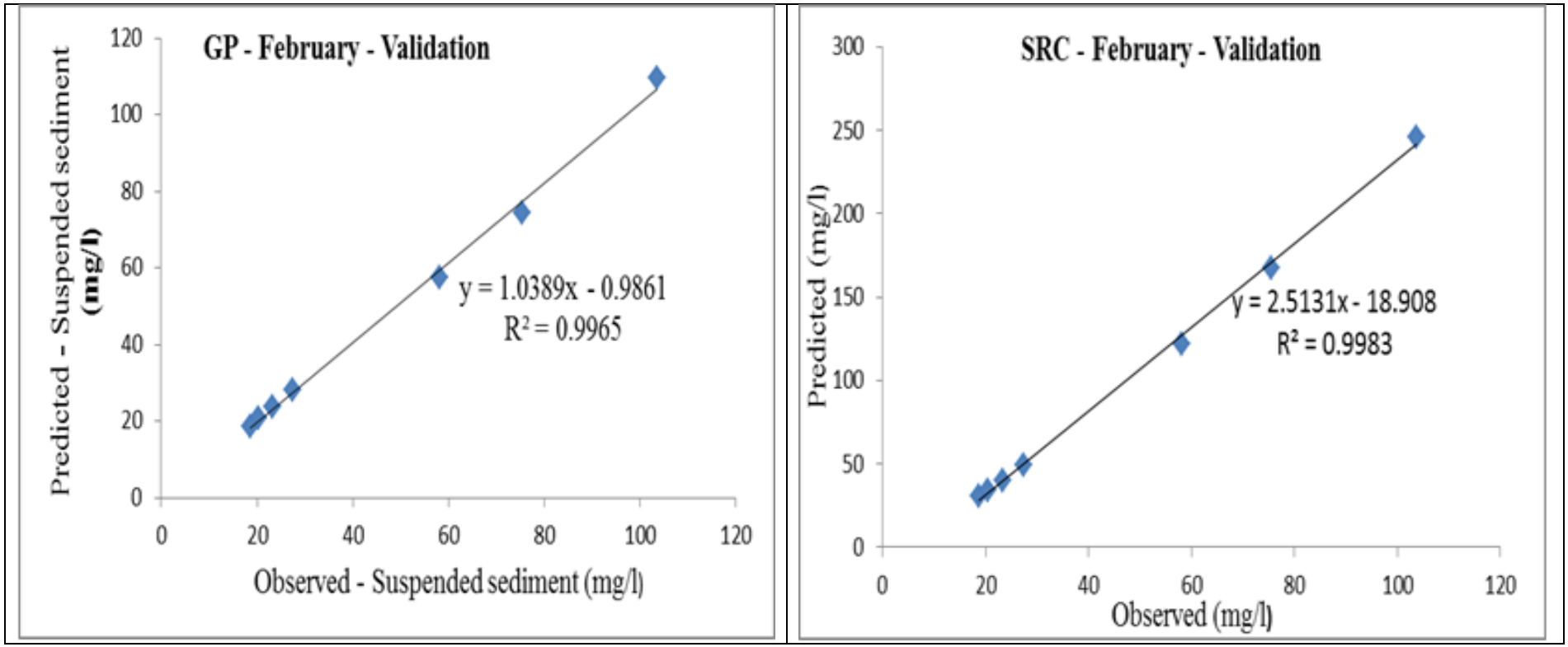


Figure 61: Observed and predicted suspended sediment load by the GP and SRC models in the validation phase for February

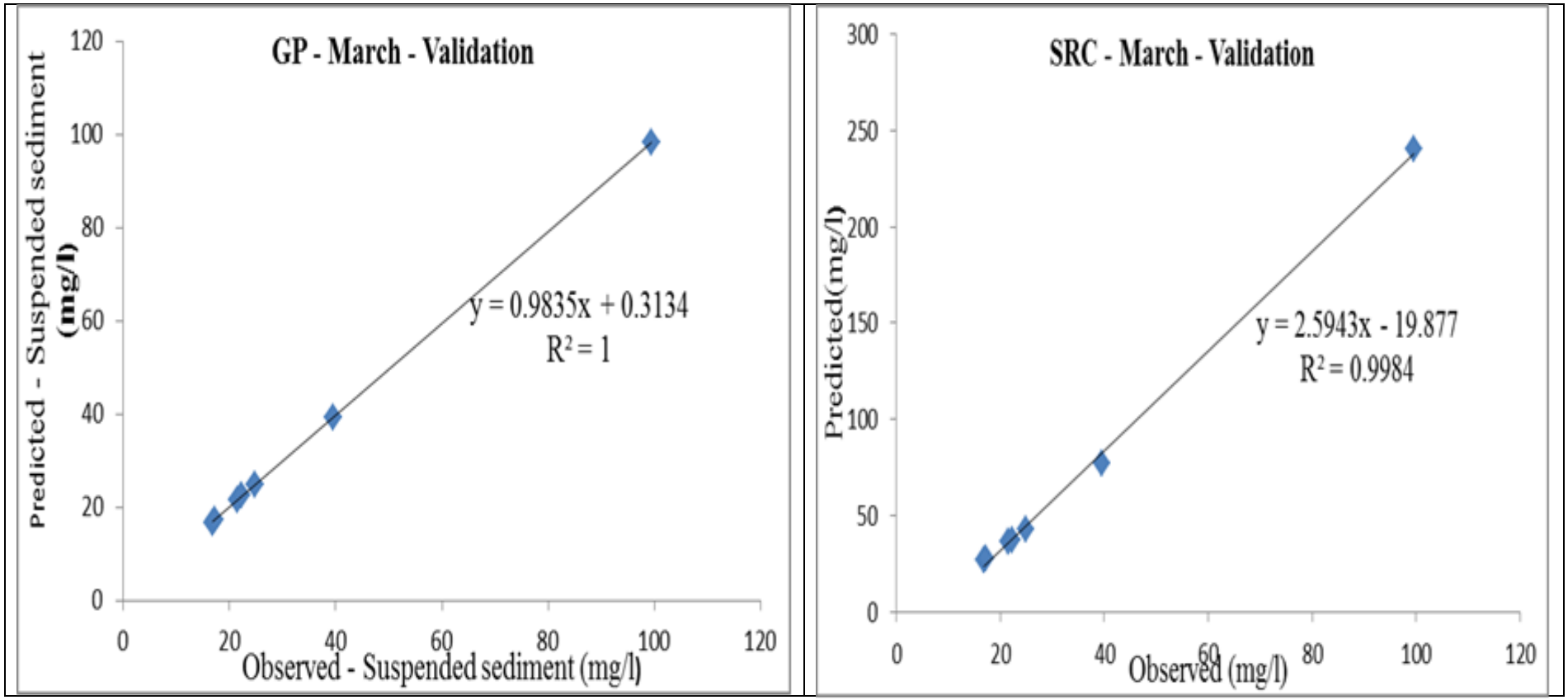


Figure 62: Observed and predicted suspended sediment load by the GP and SRC models in the validation phase for March

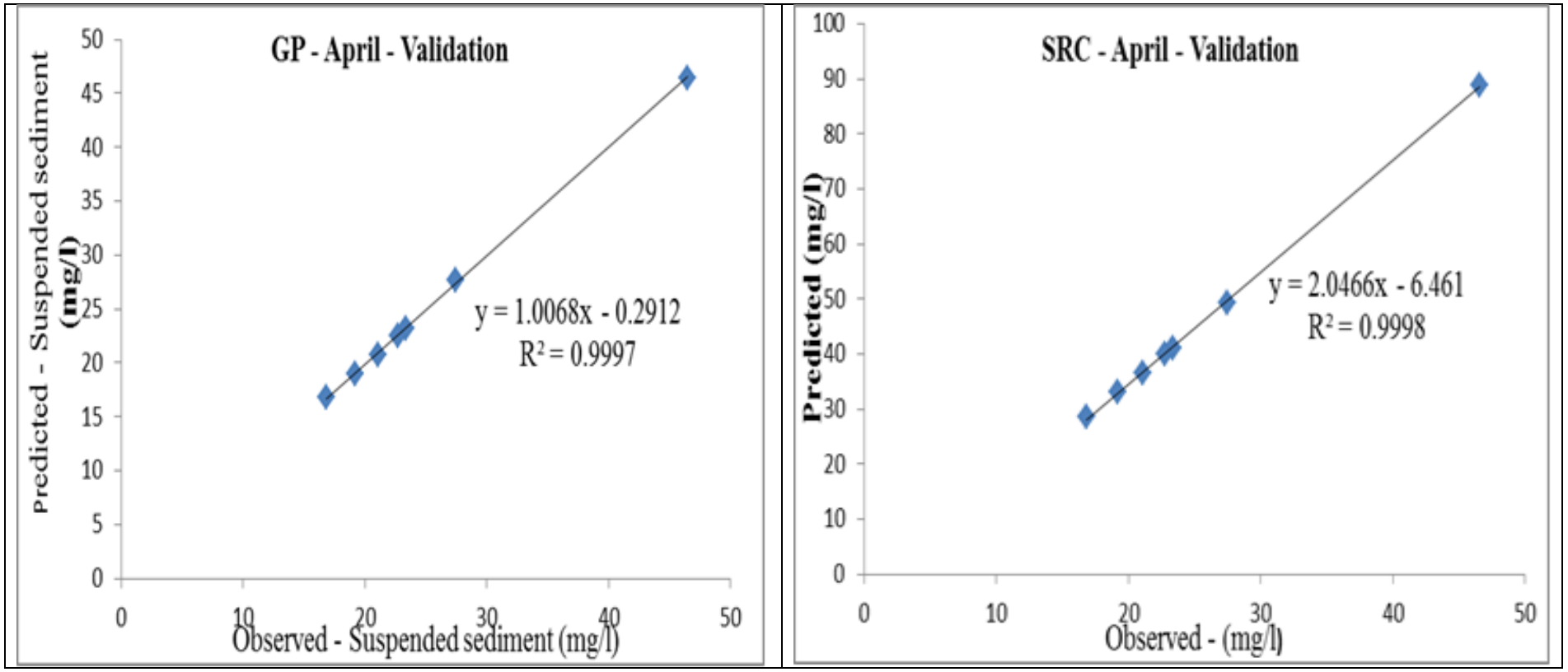


Figure 63: Observed and predicted suspended sediment load by the GP and SRC models in the validation phase for April

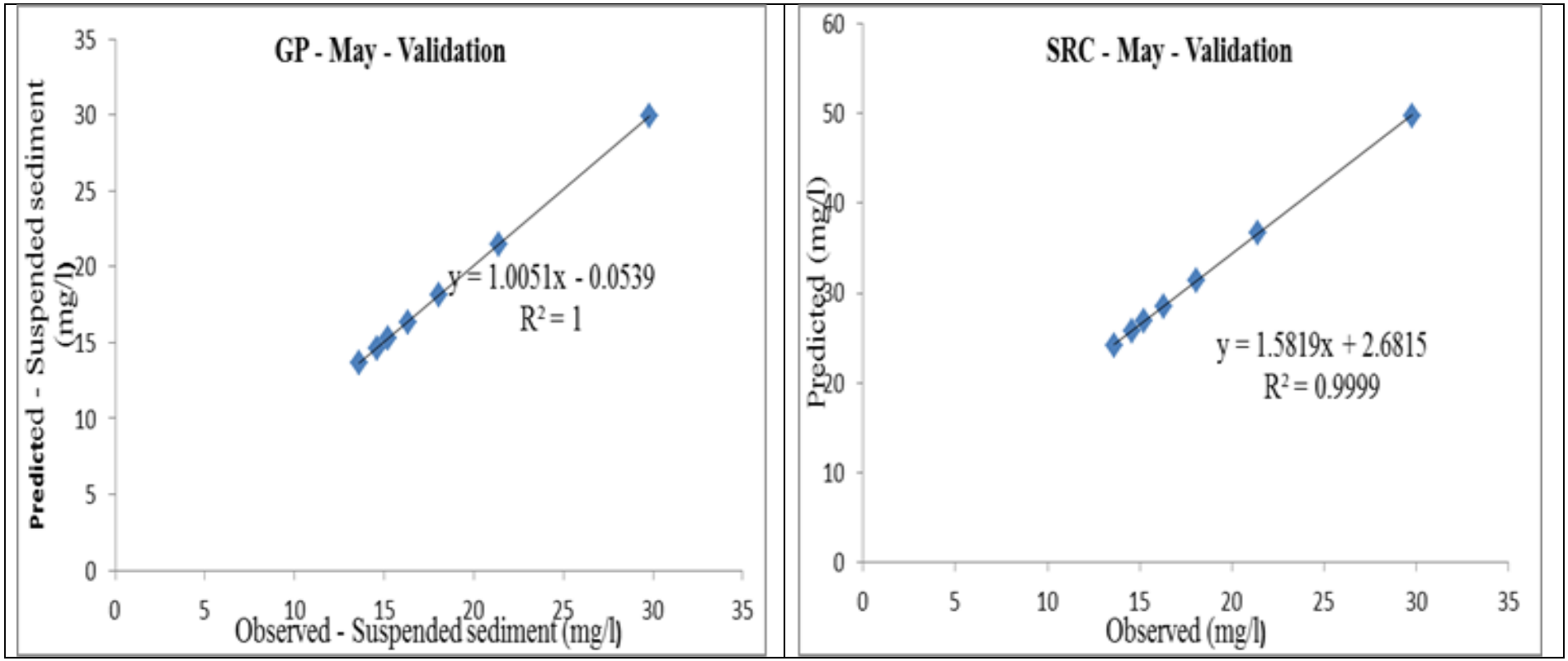


Figure 64: Observed and predicted suspended sediment load by the GP and SRC models in the validation phase for May

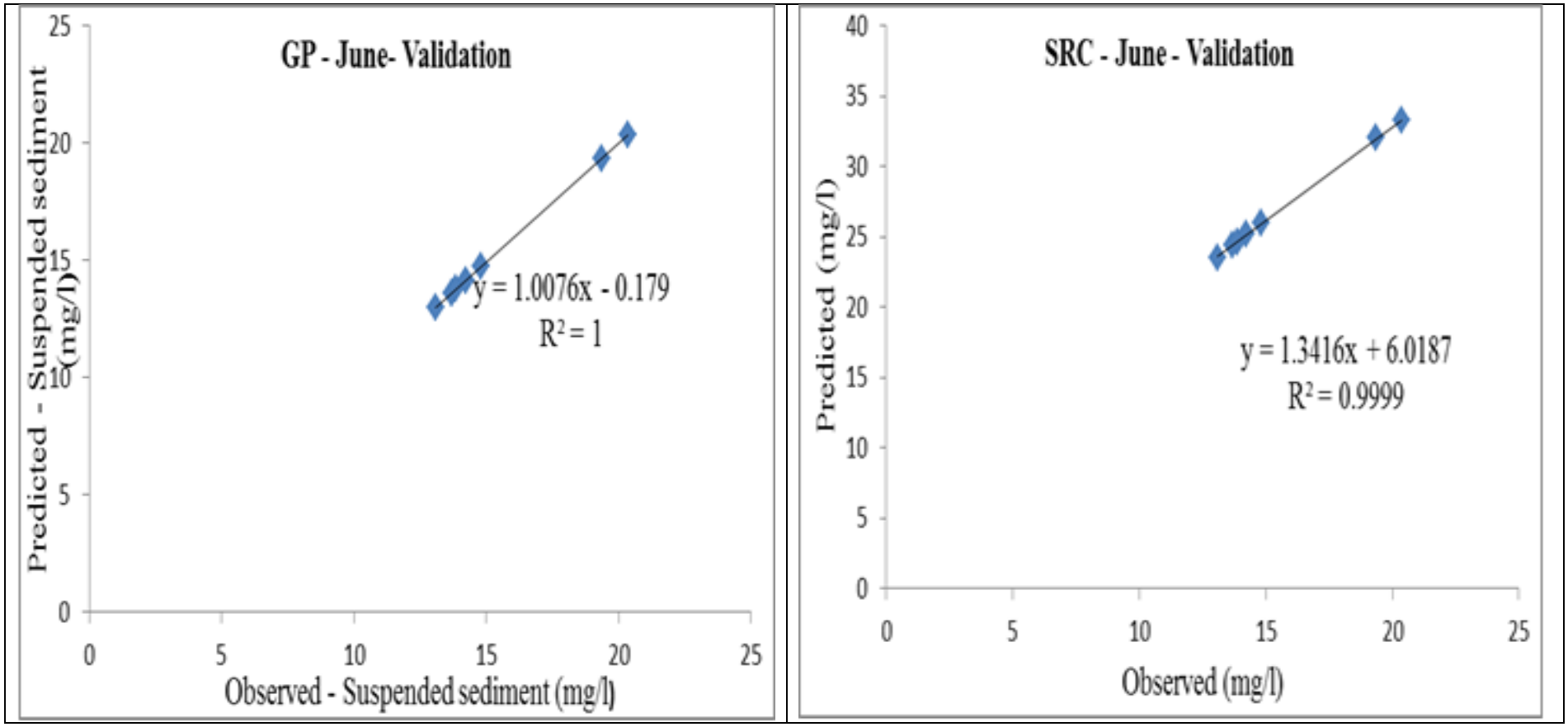


Figure 65: Observed and predicted suspended sediment load by the GP and SRC models in the validation phase for June

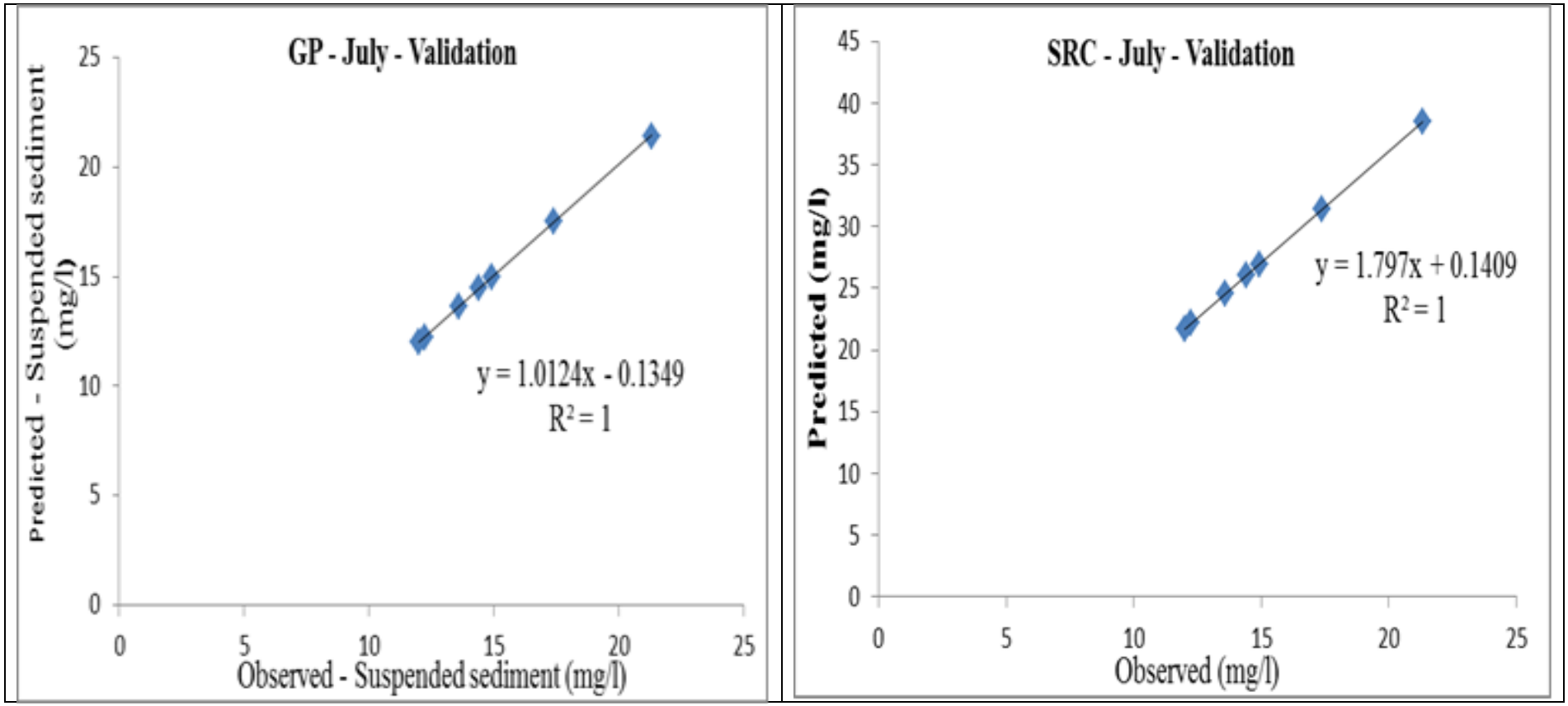


Figure 66: Observed and predicted suspended sediment load by the GP and SRC models in the validation phase for July

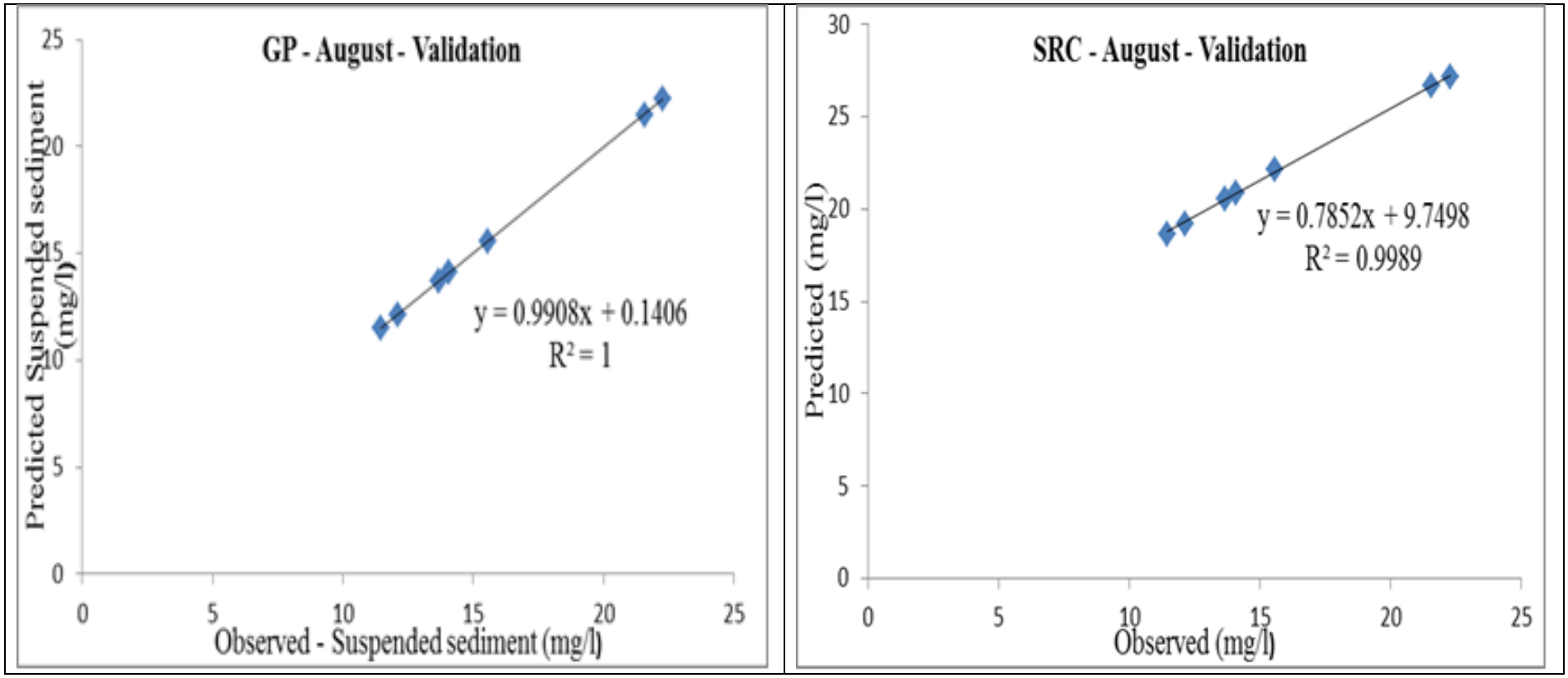


Figure 67: Observed and predicted suspended sediment load by the GP and SRC models in the validation phase for August

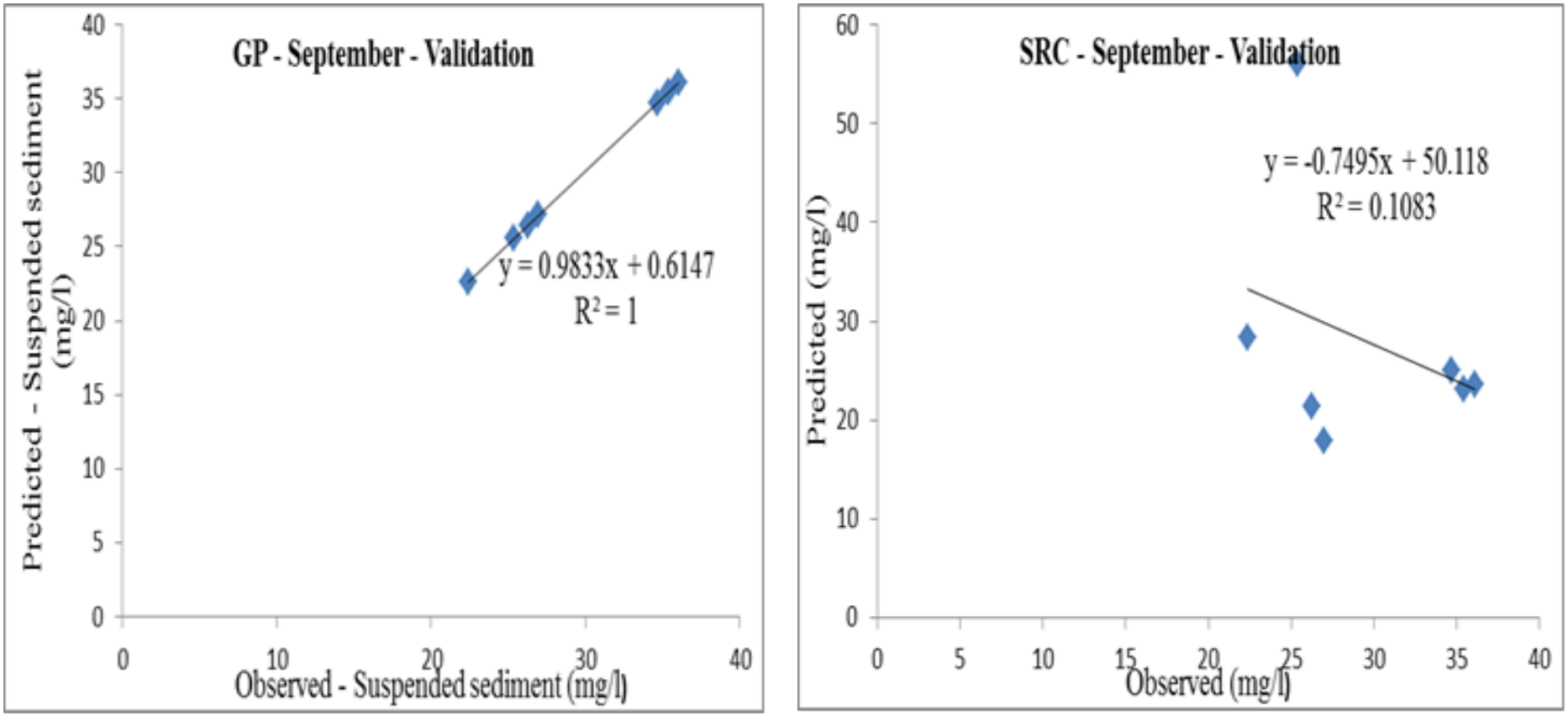


Figure 68: Observed and predicted suspended sediment load by the GP and SRC models in the validation phase for September

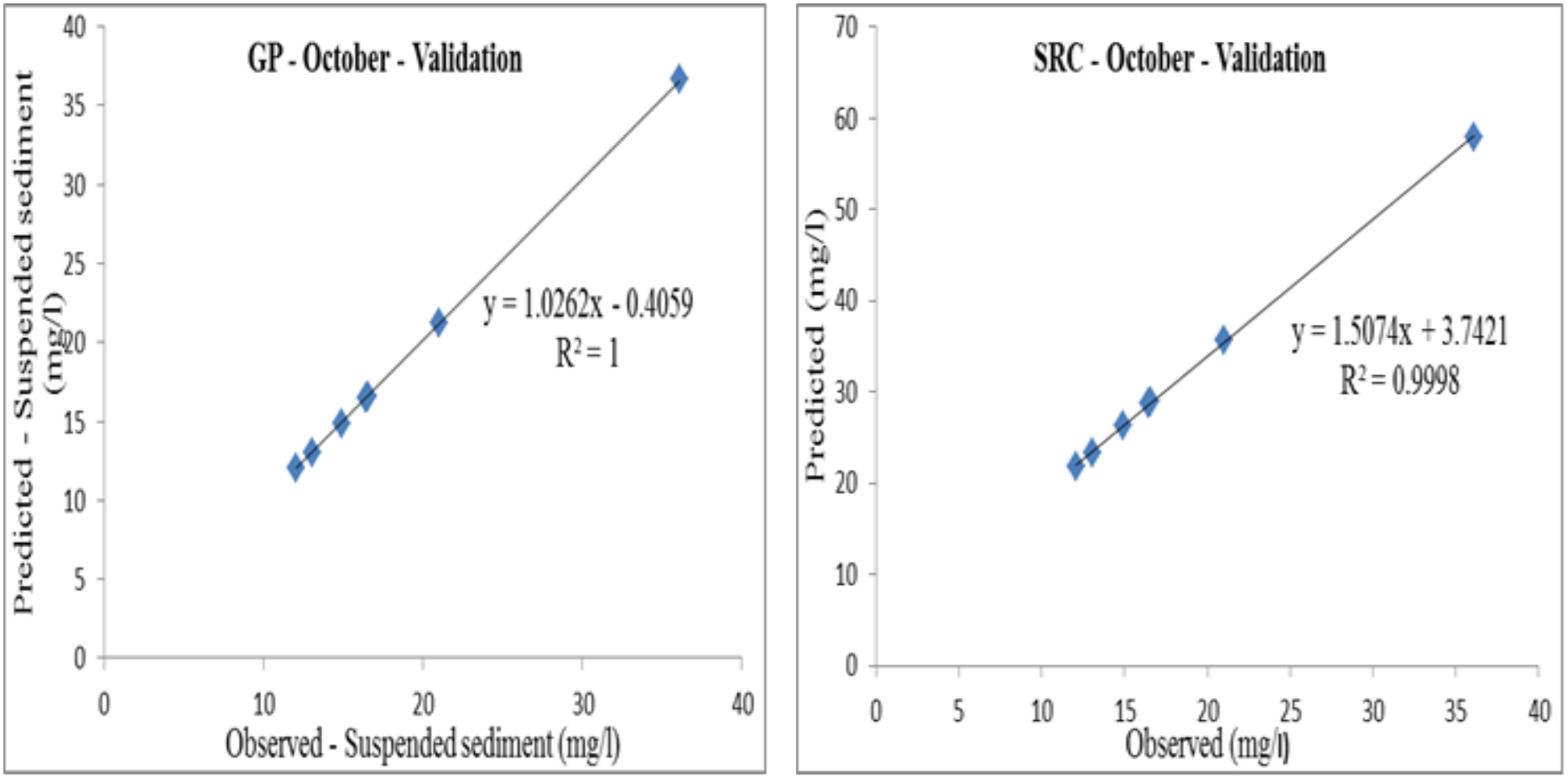
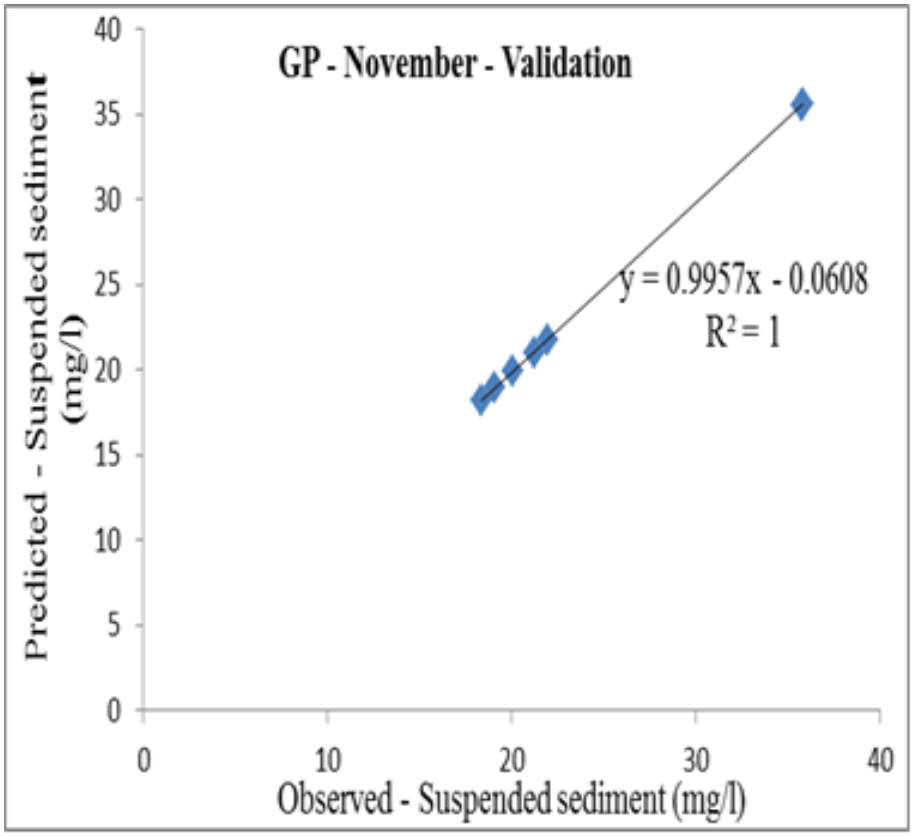


Figure 69: Observed and predicted suspended sediment load by the GP and SRC models in the validation phase for October



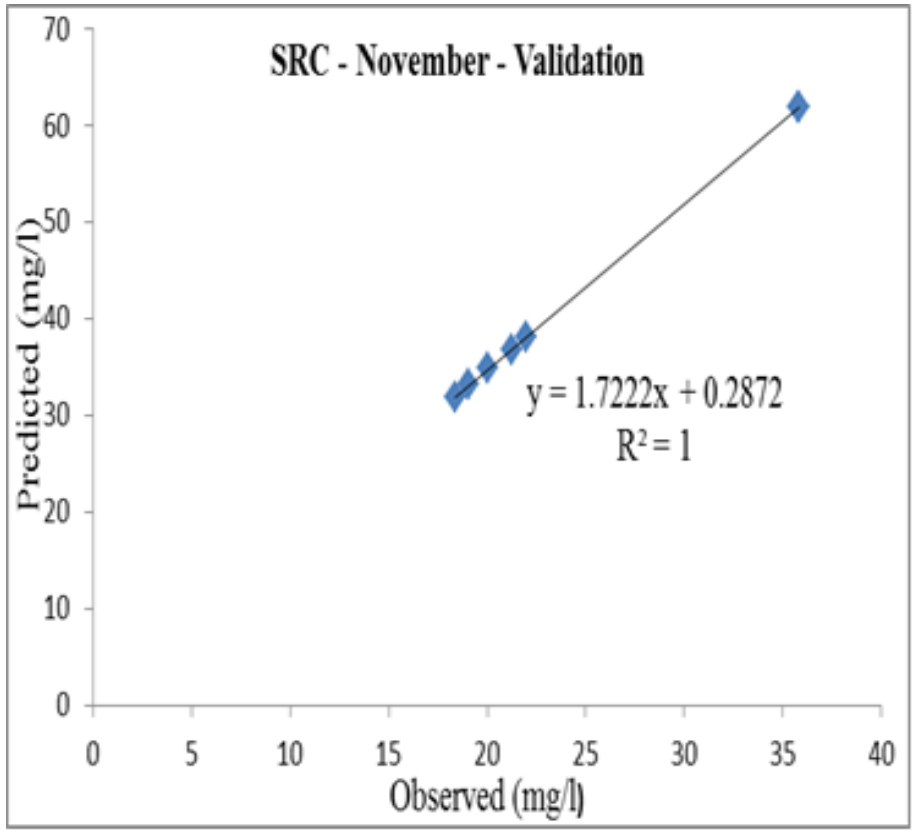


Figure 70: Observed and predicted suspended sediment load by the GP and SRC models in the validation phase for November

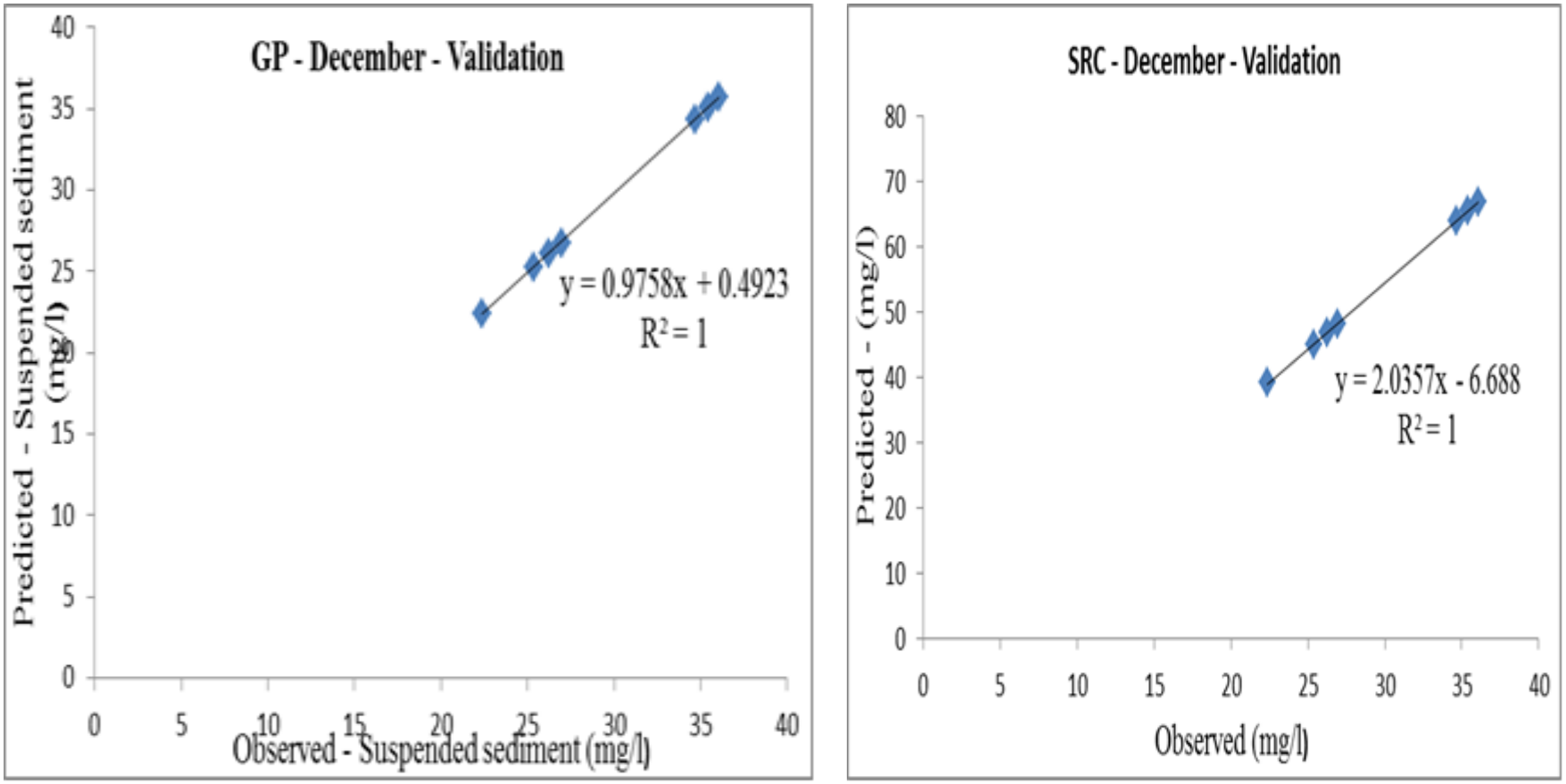


Figure 71: Observed and predicted suspended sediment load by the GP and SRC models in the validation phase for December

4.5 CONCLUSION

In this chapter, the performances of the best developed GP models and corresponding SRC models for each month of the year were investigated comparatively for average monthly suspended sediment load prediction using a limited number of datasets. The results show that the predictions from the GP models are better than the predictions from the SRC models, especially in predicting large quantities of suspended sediment load during high streamflow such as during flood events. This proves that the use of GP as a better alternative to sediment rating curve in the prediction of suspended sediment load in a small to medium basin like Inanda Dam, based on the results from this study. However, SRC technique can also be used if the observed suspended sediment load data is critically evaluated before use and any inaccuracies are taken into consideration. It also needs to be established that its predictions will serve the purpose for which it is required. The use of SRC may be preferred because it is simpler to apply than GP and it can be used to estimate long term suspended sediment data records for rivers with limited sediment database or for rivers not monitored for sediment load (Harrington and Harrington 2013). However the GP technique is also very useful when there is no clear understanding on the interrelationships between variables; when large amount of data are available in computer readable form; when conventional mathematical techniques cannot be used to solve a problem; approximate solution is acceptable; and when it is very important to measure any slight increase in performance regularly (Aytek and Kisi 2008).

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

5.1 GENERAL CONCLUSIONS

The main aim of this study was the use of an artificial intelligence model, GP, to estimate the quantity of suspended sediment yield flowing into Inanda Dam. This involves the capturing of the complex relationship beween suspended sediment load and streamflow using historical dataset. To achieve this aim, genetic programming (GP), a suitable technique capable of representing the relationship on monthly basis under limited availability of datasets was applied. In chapter 3, several monthly GP models were devloped to predict the quantity of suspended sediment load flowing into Inanda Dam and the performance of each of these models was evaluated using Root Mean Square Error (RMSE) and Coeficient of Determination (R^2) and the best performed model in each month of the year was chosen to represent the model for that month. The results from this chapter show the ability of GP technique to capture the relationship between monthly suspended sediment load and monthly streamflow in form of a model that is simple and explicit for use by anyone. In chapter 4, sediment rating curves for each month of the year was also developed to capture the relationship beween suspended sediment load and streamflow. The results from this chapter indicate that each developed monthly sediment rating curve is an accurate representation of its particular month due to the low RMSE and high R^2 values. The performance of each sediment rating curves was also investigated comparatively with that of the best monthly GP models and the results show that the GP models are more accurate for estimating the quantity of suspended sediment flowing into Inanda Dam.

To achieve the above main aim, the following specific objectives were outlined in chapter 1 of this study as follows:

a. To develop monthly models using genetic programming techniques and sediment rating curve to estimate the quantity of suspended sediment flowing into Inanda Dam.
b. To evaluate the accuracy of the developed monthly models by applying standard evaluation criteria to individual models.

c. To compare the accuracy of the developed models in order to determine the most accurate model for each month for predicting the quantity of suspended sediment flowing into Inanda Dam.
d. To estimate the monthly quantity of suspended sediment flowing into Inanda Dam using sediment rating curve.
e. To compare the performance of the most accurate GP models in each month with its corresponding sediment rating curve.

The specific objective (a) above was accomplish in chapter 3, as GP models were developed to accurately predict the quantity of suspended sediment flowing into Inanda Dam. These models were developed using different combinations of upstream flow and suspended sediment dataset in the dam. The upstream flow was used as the primary variable because it had the highest correlation values with suspended sediment load when compared with other hydrological variables. Also the historical suspended sediment load and upstream flow dataset in Inanda Dam were used to develop monthly sediment rating curve models. The performance of each developed monthly models were evaluated using Root Mean Square Error (RMSE) and coefficient of determination (R^2) as outlined in specific objective (b). From the results, all the developed models were very accurate in the prediction of suspended sediment load flowing into Inanda Dam with little significant difference between the measured and observed suspended sediment values and low RMSE values and high R^2 values. The models were also able to predict low, intermediate and high concentrations of suspended sediment. These confirm the ability of GP to predict extreme conditions and to recognise patterns (Aytek and Kisi 2008; Sirdari *et al.* 2012).

The performances of the individual GP models were evaluated based on their RMSE and R^2 values. The models with the lowest RMSE value and highest R^2 value in each month of the year were selected to represent the best model for those months for the prediction of suspended sediment flowing into Inanda Dam and this addressed specific objective (c). The historical dataset were used to develop sediment rating curves for each month of the year thereby addressing specific objective (d).

To address the specific objective (e), the most accurate monthly GP models in each month were compared with their corresponding developed SRC models in terms of their RMSE and R^2 values. The results, as highlighted in chapter 4, show that all the GP models are more accurate in the prediction of the quantity of suspended sediment flowing into Inanda Dam.

The results also show that both GP and SRC models were able to capture the complex relationship between suspended sediment load and streamflow but the GP models were more accurate with no significant differences between observed suspended sediment loads and predicted suspended sediment loads. This confirms the superiority of the GP models over conventional SRC models.

In general the results from this study are very promising and support the use of GP in predicting the nonlinear and complex relationship between suspended sediment load and streamflow.

5.2 RECOMMENDATIONS FOR FUTURE RESEARCH

The GP and SRC modelling techniques used in this study for the estimation of suspending sediment load flowing into Inanda Dam can also be applied in real-time forecasting on short-term basis like daily or hourly suspended sediment load predictions. This will help planners and managers of water resource systems to understand the system better in terms of its problems and to find alternative ways to address them. Generally the present and future users of GP need to be trained to exploit the growing GP capabilities and its successful applications in real-life problems. They should also be encouraged on its inclusion in tertiary institution curricula.

In the last two decades, EAs like GP, have been extremely researched in water resources problems especially in the areas of developing and improving optimisation algorithms and determining whether natural phenomena or heuristics inspired algorithms can be applied successfully in water resources. Now there is the need to lay more emphasis on the challenges that are hindering the full application of the algorithm to real life problems. The main challenges that need more attention in the next decade to move the application of GP techniques in the field of water resources includes (1) algorithmic issues (2) problem formulation and decision-making and, (3) benchmarking. For more details on these challenges, reader are referred to Maier *et al.* (2014).

There are other optimization techniques that could also be used to formulate the explicit relationship between suspended sediment load and streamflow. These include Particle Swarm, Differential Evolution, Artificial Fish Swarm, Imperialist Competitive Algorithm, Bee Algorithm, Ant Colony, Cuckoo Optimization Algorithm and Tabu Search. The accuracies of these techniques to capture the relationship between suspended sediment load

and streamflow could also be compared with each other. The combination of these techniques and the use of various pre-processing methodologies could enhance the efficiency, accuracy and effectiveness of the models for long term predictions.

Another direction in which present researchers are moving is multi-model simulations to determine how to accurately acquire real-time data at required times and their integration with Geographical Information System (GIS). Efforts are ongoing to incorporate the user-generated content in the futuristic models to improve the accuracy and efficiency of water resource modelling systems.

REFERENCES

Abbott, M. B., Bathurst, J. C., Cunge, J. A., O'Connell, P. E. and Rasmussen, J. 1986. An introduction to the European Hydrological System— Systeme Hydrologique Europeen, SHE. 1. History and philosophy of a physically-based, distributed modelling system. *Journal of*

Hydrology, 87: 45–59.

Abrahart, R. J., Anctil, F., Coulibaly, P., Dawson, C. W., Mount, N. J., See, L. M., Shamseldin, A. Y., Solomatine, D. P., Toth, E. and Wilby, R. L. 2012. Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting. *Progress in Physical Geography*, 36 (4): 480-513.

Adami, C. 1998. *Introduction to artificial life*. Springer Science & Business Media.

Adriaenssens, V., Baets, B. D., Goethals, P. L. and Pauw, N. D. 2004. Fuzzy rule-based models for decision support in ecosystem management. *Science of the Total Environment*, 319 (1): 1-12.

Ahmadaali, K., Liaghat, A., Haddad, O. B. and Heydari, N. 2013. Estimation of Virtual Water Using Support Vector Machine, K-nearest neighbour, and Radial Basis Function Neural Network Models. *International Journal of Agronomy and Plant Production*, 4 (11): 2926-2936.

Akbari-Alashti, H., Haddad, O. B., Fallah-Mehdipour, E. and Mariño, M. A. 2014. Multi-reservoir real-time operation rules: a new genetic programming approach.

Al-Alawi, S. M., Abdul-Wahab, S. A. and Bakheit, C. S. 2008. Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone. *Environmental Modelling & Software*, 23 (4): 396-403.

Alcázar, A. I. E. and Sharman, K. C. 1996. Some applications of genetic programming in digital signal processing. In: Proceedings of *Late Breaking Papers at the Genetic Programming 1996 Conference Stanford University*. Citeseer, 24-31.

Altman, N. S. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46 (3): 175-185.

Anctil, F., Perrin, C. and Andreassian, V. 2003. *ANN output updating of lumped conceptual rainfall/runoff forecasting models1*: Wiley Online Library.

Aqil, M., Kita, I., Yano, A. and Nishiyama, S. 2007. A comparative study of artificial neural networks and neuro-fuzzy in continuous modeling of the daily and hourly behaviour of runoff. *Journal of Hydrology*, 337 (1): 22-34.

Ashab, H. A.-D., Kozlowski, P., Goldenberg, S. L. and Moradi, M. 2014. Solutions for Missing Parameters in Computer-Aided Diagnosis with Multiparametric Imaging Data. In: *Machine Learning in Medical Imaging*. Springer, 289-296.

Asselman, N. 2000. Fitting and interpretation of sediment rating curves. *Journal of Hydrology*, 234 (3): 228-248.

Aytac, G. and Kisi, Ö. 2010. Estimation of Suspended Sediment Yield in Natural Rivers Using Machine-Coded Linear Genetic Programming. *Water Resources Management*, (25): 691–704.

Aytek, A. and Kisi, O. 2008. A genetic programming approach to suspended sediment modelling. *Journal of Hydrology*, 351: 288– 298.

Aytek, A. and Kişi, Ö. 2008. A genetic programming approach to suspended sediment modelling. *Journal of Hydrology*, 351 (3–4): 288-298.

Azamathulla, H. M. and Ghani, A. A. 2011. Genetic programming for predicting longitudinal dispersion coefficients in streams. *Water resources management*, 25 (6): 1537-1544.

Babovic, V. 2000. Data mining and knowledge discovery in sediment transport. *Computer-Aided Civil and Infrastructure Engineering*, 15 (5): 383-389.

Babovic, V. and Keijzer, M. 2000. Genetic programming as a model induction engine. *Journal of Hydroinformatics*, 2: 35-60.

Babovic, V., Keijzer, M., Aguilera, D. and Harrington, J. 2001. Automatic discovery of settling velocity equations. *D2K Technical Rep*, (D2K-0201): 1.

Back, A. D. and Trappenberg, T. P. 1999. Input variable selection using independent component analysis. In: Proceedings of *Proceedings of International Joint Conference on Neural Networks*. Citeseer, 989-992.

Bai, Y., Wang, P., Li, C., Xie, J. and Wang, Y. 2014. A multi-scale relevance vector regression approach for daily urban water demand forecasting. *Journal of Hydrology*, 517: 236-245.

Banerjee, P., Singh, V., Chatttopadhyay, K., Chandra, P. and Singh, B. 2011. Artificial neural network model as a potential alternative for groundwater salinity forecasting. *Journal of Hydrology*, 398 (3): 212-220.

Banzhaf, W., Lakhtakia, A. and Martin-Palma, R. J. 2013. Evolutionary Computation and Genetic Programming. *Engineered Biomimicry*: 429-447.

Banzhaf, W., P, Keller, R. E. and Francone, F. D. 1998. *Genetic programming: an introduction*. San Francisco (CA): Morgan Kaufmann.

Beck, M. B. 1987. Water quality modeling: a review of the analysis of uncertainty. *Water Resources Research*, 23 (8): 1393-1442.

Beck, M. B., Jakeman, A. J. and McAleer, M. J. 1995. Construction and evaluation of models of environmental systems. *In: Beck, M.B., McAleer, M.J. (Eds.), Modelling Change in Environmental Systems.John Wiley and Sons*: pp. 3–35.

Behzad, M., Asghari, K., Eazi, M. and Palhang, M. 2009. Generalization performance of support vector machines and neural networks in runoff modeling. *Expert Systems with applications*, 36 (4): 7624-7629.

Bender, M. J., Sawatsky, L. F., Long, D. and Anderson, P. 2005. *A strategy for determining acceptable sediment yield for reclaimed mine lands*. Italy: UNESCO.

Bennett, J. P. 1974. Concepts of mathematical modeling of sediment yield. *Water Resources Research*, 10 (3): 485-492.

Benyahya, L., Caissie, D., St-Hilaire, A., Ouarda, T. B. and Bobée, B. 2007. A review of statistical water temperature models. *Canadian Water Resources Journal*, 32 (3): 179-192.

Bergström, S. and Singh, V. 1995. The HBV model. *Computer models of watershed hydrology.*: 443-476.

Beven, K., Calver, A. and Morris, E. 1987. *The Institute of Hydrology distributed model*. Wallingford, UK: Institute of Hydrology. IH Report No 98.

Bhattacharya, B., Price, R. and Solomatine, D. 2005. Data-driven modelling in the context of sediment transport. *Physics and Chemistry of the Earth, Parts A/B/C*, 30 (4): 297-302.

Bhattacharya, B. and Solomatine, D. P. 2006. Machine learning in sedimentation modelling. *Neural Networks*, 19 (2): 208-214.

Bieroza, M., Baker, A. and Bridgeman, J. 2011. Classification and calibration of organic matter fluorescence data with multiway analysis methods and artificial neural networks: an operational tool for improved drinking water treatment. *Environmetrics*, 22 (3): 256-270.

Bjørnholt Karlsson, I., Obel Sonnenborg, T., Refsgaard, J. C. and Høgh Jensen, K. 2014. Significance of hydrological model choice and land use changes when doing climate change impact assessment. In: Proceedings of *EGU General Assembly Conference Abstracts*. EGU2014-2602.

Borah, D. K. 2011. Hydrologic procedures of storm event watershed models: a comprehensive review and comparison. *Hydrological Processes*, 25 (22): 3472-3489.

Boukhrissa, Z., Khanchoul, K., Le Bissonnais, Y. and Tourki, M. 2013. Prediction of sediment load by sediment rating curve and neural network (ANN) in El Kebir catchment, Algeria. *Journal of earth system science*, 122 (5): 1303-1312.

Bowden, G. J., Dandy, G. C. and Maier, H. R. 2005. Input determination for neural network models in water resources applications. Part 1—background and methodology. *Journal of Hydrology*, 301 (1): 75-92.

Box, G. E., Jenkins, G. M. and Reinsel, G. C. 2013. *Time series analysis: forecasting and control*. John Wiley & Sons.

Brierley, G. J. and Fryirs, K. A. 2008. *River futures: an integrative scientific approach to river repair*. Cambridge Univ Press.

Burke, E. K. and Kendall, G. 2005. *Search methodologies: introductory tutorials in optimization and decision support techniques*. Springer.

Castelfranchi, C. 2013. Alan Turing's "Computing Machinery and Intelligence". *Topoi*, 32 (2): 293-299.

Cerdan, O., Govers, G., Le Bissonnais, Y., Van Oost, K., Poesen, J., Saby, N., Gobin, A., Vacca, A., Quinton, J., Auerswald, K., Klik, A., Kwaad, F. J. P. M., Raclot, D., Ionita, I., Rejman, J., Rousseva, S., Muxart, T., Roxo, M. J. and Dostal, T. 2010. Rates and spatial variations of soil erosion in Europe: A study based on erosion plot data. *Geomorphology*, 122 (1–2): 167-177.

Chaker, N. and Hampel, R. 1996. Genetic programming designs hierarchic fuzzy logic controllers. In: Proceedings of *10. Workshop Fuzzy Control des GMA-FA 5.22*. 205.

Chaves, H. M. and Alipaz, S. 2007. An integrated indicator based on basin hydrology, environment, life, and policy: the watershed sustainability index. *Water Resources Management*, 21 (5): 883-895.

Chen, H.-W. and Chang, N.-B. 2010. Using fuzzy operators to address the complexity in decision making of water resources redistribution in two neighboring river basins. *Advances in water resources*, 33 (6): 652-666.

Chen, H., Xu, C.-Y. and Guo, S. 2012. Comparison and evaluation of multiple GCMs, statistical downscaling and hydrological models in the study of climate change impacts on runoff. *Journal of Hydrology*, 434: 36-45.

Chen, S.-T., Yu, P.-S. and Tang, Y.-H. 2010. Statistical downscaling of daily precipitation using support vector machines and multivariate analysis. *Journal of Hydrology*, 385 (1): 13-22.

Chen, V. Y., Lien, H.-P., Liu, C.-H., Liou, J. J., Tzeng, G.-H. and Yang, L.-S. 2011. Fuzzy MCDM approach for selecting the best environment-watershed plan. *Applied Soft Computing*, 11 (1): 265-275.

Cherif, E. A., Errih, M. and Cherif, H. M. 2009. Modélisation statistique du transport solide du bassin versant de l'Oued Mekerra (Algérie) en zone semi-aride méditerranéenne. *Hydrological sciences journal*, 54 (2): 338-348.

Chiang, Y. M. and Chang, F. J. 2009. Integrating hydrometeorological information for rainfall-runoff modelling by artificial neural networks. *Hydrological Processes*, 23 (11): 1650-1659.

Cisty, M. 2010. Hybrid genetic algorithm and linear programming method for least-cost design of water distribution systems. *Water resources management*, 24 (1): 1-24.

Cohn, T. and Gilroy, E. 1991. Estimating loads from periodic records. *U.S. Geological Survey Branch of Systems Analysis Technical Memo*, 91 (01): 81.

Cramer, N. L. 1985. A Representation for the Adaptive Generation of Simple Sequential Programs. In: Proceedings of *Proceedings of the 1st International Conference on Genetic Algorithms*. L. Erlbaum Associates Inc., 183-187.

Daida, J. M., Hommes, J. D., Bersano-Begey, T. F., Ross, S. J. and Vesecky, J. F. 1996. 21 Algorithm Discovery Using the Genetic Programming Paradigm: Extracting Low-Contrast Curvilinear Features from SAR Images of Arctic Ice.

Dams, W. C. o. 2000. *Dams and Development: A New Framework for Decision-making: the Report of the World Commission on Dams*. 1 ed. Earthscan.

Danandeh Mehr, A., Kahya, E. and Olyaie, E. 2013. Streamflow prediction using linear genetic programming in comparison with a neuro-wavelet technique. *Journal of Hydrology*, 505: 240-249.

Datta, B., Prakash, O. and Sreekanth, J. 2014. Application of Genetic Programming Models Incorporated in Optimization Models for Contaminated Groundwater Systems Management. In: *EVOLVE-A Bridge between Probability, Set Oriented Numerics, and Evolutionary Computation V*. Springer, 183-199.

De Vente, J. and Poesen, J. 2005. Predicting soil erosion and sediment yield at the basin scale: scale issues and semi-quantitative models. *Earth-Science Reviews*, 71 (1): 95-125.

Dekker, L. J., Boogerd, W., Stockhammer, G., Dalebout, J. C., Siccama, I., Zheng, P., Bonfrer, J. M., Verschuuren, J. J., Jenster, G. and Verbeek, M. M. 2005. MALDI-TOF mass spectrometry analysis of

cerebrospinal fluid tryptic peptide profiles to diagnose leptomeningeal metastases in patients with breast cancer. *Molecular & Cellular Proteomics*, 4 (9): 1341-1349.

Deo, M. C. 2009. Recent Data Driven Methods and Applications in Coastal and Hydrologic Data Analysis. *ISH Journal of Hydraulic Engineering*, 15 (sup1): 310-327.

Dessai, S. and Hulme, M. 2007. Assessing the robustness of adaptation decisions to climate change uncertainties: A case study on water resources management in the East of England. *Global Environmental Change*, 17 (1): 59-72.

Dhamge, N. R., Atmapoojya, S. and Kadu, M. S. 2012. Genetic Algorithm Driven ANN Model for Runoff Estimation. *Procedia Technology*, 6: 501-508.

Dorado, J., RabuñAL, J. R., Pazos, A., Rivero, D., Santos, A. and Puertas, J. 2003. Prediction and modeling of the rainfall-runoff transformation of a typical urban basin using ANN and GP. *Applied Artificial Intelligence*, 17 (4): 329-343.

Duan, N. 1983. Smearing estimate: a nonparametric retransformation method. *Journal of the American Statistical Association*, 78 (383): 605-610.

Duan, W., Takara, K., He, B., Luo, P., Nover, D. and Yamashiki, Y. 2013. Spatial and temporal trends in estimates of nutrient and suspended sediment loads in the Ishikari River, Japan, 1985 to 2010. *Science of The Total Environment*, 461: 499-508.

Dunlop, J., Kefford, B., McNeil, V., McGregor, G., Choy, S. and Nugegoda, D. 2008. A review of guideline development for suspended solids and salinity in tropical rivers of Queensland, Australia. *Australasian Journal of Ecotoxicology*, 14 (2/3): 129.

DWAF. 2008. *Resource Management Plan for Inanda Dam Final Draft*. Durban: eThekwini Municipality and the Department of Water Affairs and Forestry.

East, A. E., Pess, G. R., Bountry, J. A., Magirl, C. S., Ritchie, A. C., Logan, J. B., Randle, T. J., Mastin, M. C., Minear, J. T. and Duda, J. J. 2015. Large-scale dam removal on the Elwha River, Washington, USA: River channel and floodplain geomorphic change. *Geomorphology*, 228: 765-786.

Elshorbagy, A., Corzo, G., Srinivasulu, S. and Solomatine, D. 2010. Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology-Part 1: Concepts and methodology. *Hydrology and Earth System Sciences*, 14 (10): 1931-1941.

Fallah-Mehdipour, E., Bozorg Haddad, O. and Mariño, M. 2013a. Prediction and simulation of monthly groundwater levels by genetic programming. *Journal of Hydro-environment Research*, 7 (4): 253-260.

Fallah-Mehdipour, E., Haddad, O. B. and Mariño, M. 2012. Real-time operation of reservoir system by genetic programming. *Water resources management*, 26 (14): 4091-4103.

Fallah-Mehdipour, E., Haddad, O. B. and Mariño, M. 2013b. Developing reservoir operational decision rule by genetic programming. *Journal of Hydroinformatics*, 15 (1): 103-119.

Fallah-Mehdipour, E., Haddad, O. B. and Mariño, M. A. 2014. Genetic programming in groundwater modeling. *Journal of Hydrologic Engineering*, 19 (12)

Fan, X., Shi, C., Zhou, Y. and Shao, W. 2012. Sediment rating curves in the Ningxia-Inner Mongolia reaches of the upper Yellow River and their implications. *Quaternary International*, 282: 152-162.

Ferguson, R. 1987. Accuracy and precision of methods for estimating river loads. *Earth surface processes and landforms*, 12 (1): 95-104.

Ferreira, C. 2001. Gene expression programming: a new adaptive algorithm for solving problems. *Complex System*, 13 (2): 87-129.

Fogel, D. B. 1992. Evolving artificial intelligence. Doctoral Dissertation, University of California, San Diego, USA.

Fogel, D. B. 1998. Evolutionary Computation. The Fossil Record. Selected Readings on the History of Evolutionary Algorithms. *New York: The Institute of Electrical and Electronic Engineers*,

Fogel, L. J., Owens, A. J. and Walsh, M. J. 1966. *Artificial intelligence through simulated evolution*. Chichester, UK:

Forsyth, R. 1981. A DARWINIAN APPROACH TO PATTERN RECOGNITION. *Kybernetes*, 10 (3): 159-166.

Fu, G. and Kapelan, Z. 2011. Fuzzy probabilistic design of water distribution networks. *Water Resources Research*, 47 (5)

Fujiko, C. and Dickinson, J. 1987. Using the genetic algorithm to generate LISP source code to solve the prisoner's dilemma. In: Proceedings of *Proceedings of the Second International Conference on Genetic Algorithms on Genetic algorithms and their application*. L. Erlbaum Associates Inc., 236-240.

Gao, S. and Wang, Y. P. 2008. Changes in material fluxes from the Changjiang River and their implications on the adjoining continental shelf ecosystem. *Continental Shelf Research*, 28 (12): 1490-1500.

Garg, A., Garg, A., Tai, K., Barontini, S. and Stokes, A. 2014. A computational intelligence-based genetic programming approach for the simulation of soil water retention curves. *Transport in porous media*, 103 (3): 497-513.

Garg, V. 2011. Modeling catchment sediment yield: a genetic programming approach. *Natural Hazards*, 70 (1): 39-50.

Garg, V. and Jothiprakash, V. 2013. Evaluation of reservoir sedimentation using data driven techniques. *Applied Soft Computing*, 13 (8): 3567–3581.

Gay, A., Cerdan, O., Delmas, M. and Desmet, M. 2014. Variability of suspended sediment yields within the Loire river basin (France). *Journal of Hydrology*, 519: 1225-1237.

Ghani, A. A. 1993. Sediment transport in sewers. Doctoral Thesis, University of Newcastle, Upon Tyne, United Kingdom.

Ghorbani, M. A., Kisi, O. and Aalinezhad, M. 2010. A probe into the chaotic nature of daily streamflow time series by correlation dimension and largest Lyapunov methods. *Applied Mathematical Modelling*, 34 (12): 4050-4057.

Giustolisi, O. 2004. Using genetic programming to determine Chezy resistance coefficient in corrugated channels. *Journal of Hydroinformatics*, 6: 157-173.

Giustolisi, O. and Laucelli, D. 2005. Improving generalization of artificial neural networks in rainfall–runoff modelling/Amélioration de la généralisation de réseaux de neurones artificiels pour la modélisation pluie-débit. *Hydrological Sciences Journal*, 50 (3): 457.

Govindaraju, R. S. and Rao, A. R. 2010. *Artificial neural networks in hydrology*. 1 ed. New York, NY 10087-0777, USA: Springer Publishing Company, Incorporated.

Gray, H. F., Maxwell, R. J., Martínez-Pérez, I., Arús, C. and Cerdán, S. 1998. Genetic programming for classification and feature selection: analysis of 1H nuclear magnetic resonance spectra from human brain tumour biopsies. *NMR in Biomedicine*, 11 (4-5): 217-224.

Griffiths, P. G., Hereford, R. and Webb, R. H. 2006. *Sediment yield and runoff frequency of small drainage basins in the Mojave Desert, California and Nevada* Tuscon, Arizona: U.S. Geological Survey.

Grimes, C. 1995. Application of genetic techniques to the planning of railway track maintenance work. In: Proceedings of *Genetic Algorithms in Engineering Systems: Innovations and Applications, 1995. GALESIA. First International Conference on (Conf. Publ. No. 414)*. IET, 467-472.

Guo, J., Zhou, J., Qin, H., Zou, Q. and Li, Q. 2011. Monthly streamflow forecasting based on improved support vector machine model. *Expert Systems with Applications*, 38 (10): 13073-13081.

Guven, A. 2009. Linear genetic programming for time-series modelling of daily flow rate. *Journal of earth system science*, 118 (2): 137-146.

Guven, A. and Kişi, Ö. 2011a. Daily pan evaporation modeling using linear genetic programming technique. *Irrigation science*, 29 (2): 135-145.

Guven, A. and Kişi, Ö. 2011b. Estimation of suspended sediment yield in natural rivers using machine-coded linear genetic programming. *Water resources management*, 25 (2): 691-704.

Handley, S. and Forrest, S. 1993. Automatic Learning of a Detector for alpha-helices in protein. In: Proceedings of *Proceedings of the 5th International Conference on*. Morgan Kaufmann, 271-278.

Harrington, S. T. and Harrington, J. R. 2013. An assessment of the suspended sediment rating curve approach for load estimation on the Rivers Bandon and Owenabue, Ireland. *Geomorphology*, 185: 27-38.

Harris, E., Babovic, V. and Falconer, R. 2003. Velocity predictions in compound channels with vegetated floodplains using genetic programming. *International Journal of River Basin Management*, 1 (2): 117-123.

Harvey, N. R., Theiler, J., Brumby, S. P., Perkins, S., Szymanski, J. J., Bloch, J. J., Porter, R. B., Galassi, M. and Young, A. C. 2002. Comparison of GENIE and conventional supervised classifiers for multispectral image feature extraction. *Geoscience and Remote Sensing, IEEE Transactions on*, 40 (2): 393-404.

He, B., Oki, T., Sun, F., Komori, D., Kanae, S., Wang, Y., Kim, H. and Yamazaki, D. 2011. Estimating monthly total nitrogen concentration in streams by using artificial neural network. *Journal of Environmental Management*, 92 (1): 172-177.

Hearst, M. A., Dumais, S., Osman, E., Platt, J. and Scholkopf, B. 1998. Support vector machines. *Intelligent Systems and their Applications, IEEE*, 13 (4): 18-28.

Heng, S. and Suetsugi, T. 2013. Using artificial neural network to estimate sediment load in Ungauged catchments of the Tonle Sap River Basin, Cambodia. *Journal of Water Resource and Protection*, 5 (2): 111-123.

Heng, S. and Suetsugi, T. 2014. Comparison of regionalization approaches in parameterizing sediment rating curve in ungauged catchments for subsequent instantaneous sediment yield prediction. *Journal of Hydrology*, 512: 240-253.

Hessenmoller, D. and Elsenhans, A. 2002. Comparison between parametric and non-parametric methods to predict the annual increment of beech. *ALLGEMEINE FORST UND JAGDZEITUNG*, 173 (11-12): 216-223.

Hicklin, J. F. 1986. Application of the genetic algorithm to automatic program generation. University of Idaho, USA.

Holland, J. H. 1962. Outline for a logical theory of adaptive systems. *Journal of the ACM (JACM)*, 9 (3): 297-314.

Holtschlag, D. J. 2001. Optimal estimation of suspended-sediment concentrations in streams. *Hydrological processes*, 15 (7): 1133-1155.

Horowitz, A. J. 2003. An evaluation of sediment rating curves for estimating suspended sediment concentrations for subsequent flux calculations. *Hydrological processes*, 17 (17): 3387-3409.

Horowitz, A. J. 2008. Determining annual suspended sediment and sediment-associated trace element and nutrient fluxes. *Science of the Total Environment*, 400 (1): 315-343.

Howard, D., Roberts, S. and Brankin, R. 1999. Target detection in SAR imagery by genetic programming. *Advances in Engineering Software*, 30 (5): 303-311.

Hrnjica, B. I. 2015. GPdotNET V4.0- artificial intelligence tool [Computer program]

(computer software and manual). Available: http://gpdotnet.codeplex.com (Accessed 04/03/2015).

Hsu, K. l., Gupta, H. V. and Sorooshian, S. 1995. Artificial Neural Network Modeling of the Rainfall-Runoff Process. *Water resources research*, 31 (10): 2517-2530.

Hu, B., Wang, H., Yang, Z. and Sun, X. 2011. Temporal and spatial variations of sediment rating curves in the Changjiang (Yangtze River) basin and their implications. *Quaternary International*, 230 (1): 34-43.

Huang, J., Lin, X., Wang, J. and Wang, H. 2015. The precipitation driven correlation based mapping method (PCM) for identifying the critical source areas of non-point source pollution. *Journal of Hydrology*, 525: 100-110.

Hudson, N., Furness, H. and Boucher, K. 1990. *Water quality of Inanda dam-some indicators for the management of a raw water reservoir*. Water Quality Department, Umgeni Water, Pietermarizburg.

Hudson, P. F. 2003. Event sequence and sediment exhaustion in the lower Panuco Basin, Mexico. *Catena*, 52 (1): 57-76.

Jacquin, A. P. and Shamseldin, A. Y. 2006. Development of rainfall–runoff models using Takagi–Sugeno fuzzy inference systems. *Journal of Hydrology*, 329 (1): 154-173.

Jahanshahi, M., Rahmani, S. and Ghaderi, S. 2015. An efficient cluster head selection algorithm for wireless sensor networks using fuzzy inference systems. *International Journal of Smart Electrical Engineering*, 2 (2)

Jakeman, A. and Hornberger, G. 1993. How much complexity is warranted in a rainfall-runoff model? *Water Resources Research*, 29 (8): 2637-2649.

Jakeman, A. J., Green, T. R., Beavis, S. G., Zhang, L., Dietrich, C. R. and Crapper, P. F. 1999. Modelling upland and in-stream erosion, sediment and phosphorus transport in a large catchment. *Hydrological Processes* 13 (5): 745–752.

Jansson, M. 1985. A comparison of detransformed logarithmic regressions and power function regressions. *Geografiska Annaler. Series A. Physical Geography*: 61-70.

Jayawardena, A., Muttil, N. and Fernando, T. 2005. Rainfall-runoff modelling using genetic programming. In: Proceedings of *Proceedings of the MODSIM 2005 international congress on modelling and simulation: advances and applications for management and decision making. Melbourne, Australia*. 1841-1847.

Johari, A., Habibagahi, G. and Ghahramani, A. 2006. Prediction of soil–water characteristic curve using genetic programming. *Journal of Geotechnical and Geoenvironmental Engineering*, 132 (5): 661-665.

Jordaan, E. M. 2002. Development of robust inferential sensors: industrial applications of support vector machines for regression. Doctoral thesis, Technical University Eindhoven. The Netherlands.

Kachroo, R. 1992. River flow forecasting. Part 1. A discussion of the principles. *Journal of Hydrology*, 133 (1): 1-15.

Kakaei Lafdani, E., Moghaddam Nia, A. and Ahmadi, A. 2013. Daily suspended sediment load prediction using artificial neural networks and support vector machines. *Journal of Hydrology*, 478: 50-62.

Kalma, J. D. and Sivapalan, M. 1995. *Scale issues in hydrological modelling*. Chichester, UK: John Wiley and Sons.

Karim, M. F. and Kennedy, J. F. 1990. Menu of coupled velocity and sediment-discharge relations for rivers. *Journal of Hydraulic Engineering*, 116 (8): 978-996.

Karlsson, M. and Yakowitz, S. 1987. Nearest-neighbor methods for nonparametric rainfall-runoff forecasting. *Water Resources Research*, 23 (7): 1300-1308.

Khalil, B., Ouarda, T. and St-Hilaire, A. 2011. Estimation of water quality characteristics at ungauged sites using artificial neural networks and canonical correlation analysis. *Journal of Hydrology*, 405 (3): 277-287.

Khanchoul, K., El Abidine Boukhrissa, Z., Acidi, A. and Altschul, R. 2012. Estimation of suspended sediment transport in the Kebir drainage basin, Algeria. *Quaternary International*, 262: 25-31.

Khanchoul, K. and Jansson, M. B. 2008. Sediment rating curves developed on stage and seasonal means in discharge classes for the Mellah wadi, Algeria. *Geografiska Annaler: Series A, Physical Geography*, 90 (3): 227-236.

Khayet, M., Cojocaru, C. and Essalhi, M. 2011. Artificial neural network modeling and response surface methodology of desalination by reverse osmosis. *Journal of Membrane Science*, 368 (1): 202-214.

Kim, S., Seo, D.-J., Riazi, H. and Shin, C. 2014. Improving water quality forecasting via data assimilation–Application of maximum likelihood ensemble filter to HSPF. *Journal of Hydrology*, 519: 2797-2809.

Kinnear, K. E., Spector, L. C. and Angeline, P. J. 1999. *Advances in genetic programming*. Cambridge MA: MIT press.

Kişi, Ö. 2007. Development of stream flow-suspended sediment rating curve using a range dependent neural network. *International Journal of Science & Technology*, 2 (1): 49-61.

Kişi, Ö. 2009. Neural networks and wavelet conjunction model for intermittent streamflow forecasting. *Journal of Hydrologic Engineering*, 14 (8): 773-782.

Kisi, O. and Cimen, M. 2011. A wavelet-support vector machine conjunction model for monthly streamflow forecasting. *Journal of Hydrology*, 399 (1): 132-140.

Kisi, O., Dailr, A. H., Cimen, M. and Shiri, J. 2012. Suspended sediment modeling using genetic programming and soft computing techniques. *Journal of Hydrology*, 450-451: 48-58.

Kisi, O. and Guven, A. 2010. A machine code-based genetic programming for suspended sediment concentration estimation. *Advances in Engineering Software*, 41 (7–8): 939-945.

Kisi, O. and Shiri, J. 2010. A comparison of genetic programming and ANFIS in forecasting daily, monthly and daily streamflows. In: Proceedings of *Proceedings of the international symposium on innovations in intelligent systems and applications*. 118–122.

Kisi, O. and Shiri, J. 2011. Precipitation forecasting using wavelet-genetic programming and wavelet-neuro-fuzzy conjunction models. *Water resources management*, 25 (13): 3135-3152.

Kisi, O. and Shiri, J. 2012. River suspended sediment estimation by climatic variables implication: Comparative study among soft computing techniques. *Computers & Geosciences*, 43: 73-82.

Kizhisseri, A. S., Simmonds, D., Rafiq, Y. and Borthwick, M. 2005. An evolutionary computation approach to sediment transport modeling. In: Proceedings of *Fifth International Conference on Coastal Dynamics*. 4-8.

Kleissen, F., Beck, M. and Wheater, H. 1990. The identifiability of conceptual hydrochemical models. *Water Resources Research*, 26 (12): 2979-2992.

Kondolf, G. M. 1997. PROFILE: hungry water: effects of dams and gravel mining on river channels. *Environmental management*, 21 (4): 533-551.

Kouli, M., Soupios, P. and Vallianatos, F. 2009. Soil erosion prediction using the revised universal soil loss equation (RUSLE) in a GIS framework, Chania, Northwestern Crete, Greece. *Environmental Geology*, 57 (3): 483-497.

Koza, J. R. 1992. *Genetic programming: on the programming of computers by means of natural selection*. Cambridge MA: MIT press.

Koza, J. R. and Andre, D. 1996. Classifying protein segments as transmembrane domains using architecture-altering operations in genetic programming. *Advances in genetic programming*, 2: 155-176.

Kumar, B., Jha, A., Deshpande, V. and Sreenivasulu, G. 2014. Regression model for sediment transport problems using multi-gene symbolic genetic programming. *Computers and Electronics in Agriculture*, 103 (0): 82-90.

Kumar, R. R. R. and Series, S. C. 2012. *WATER, SOIL, AND SEDIMENT CHARACTERISATION: SHARAVATHI RIVER BASIN, WESTERN GHATS* (ENVIS Technical Report). Indian Institute of Science, Bangalore, INDIA: Environmental Information System [ENVIS]

Centre for Ecological Sciences

Lachtermacher, G. and Fuller, J. D. 1994. Backpropagation in Hydrological Time Series Forecasting. In: Hipel, K., McLeod, A. I., Panu, U. S. and Singh, V. eds. *Stochastic and Statistical Methods in Hydrology and Environmental Engineering*. Springer Netherlands, 229-242. Available: http://dx.doi.org/10.1007/978-94-017-3083-9_18 (Accessed 23/02/2015).

Laio, F., Porporato, A., Revelli, R. and Ridolfi, L. 2003. A comparison of nonlinear flood forecasting methods. *Water Resources Research*, 39 (5)

Leahy, P., Kiely, G. and Corcoran, G. 2008. Structural optimisation and input selection of an artificial neural network for river level prediction. *Journal of Hydrology*, 355 (1): 192-201.

Lee, T. and Ouarda, T. B. M. J. 2011. Identification of model order and number of neighbors for k-nearest neighbor resampling. *Journal of Hydrology*, 404 (3–4): 136-145.

Legates, D. R. and McCabe, G. J. 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water resources research*, 35 (1): 233-241.

Letcher, R. A., Jakeman, A. J., Merritt, W. S., McKee, L. J., Eyre, B. D. and Baginska, B. 1999. *Review of techniques to estimate catchment exports* (Technical Report). New South Wales, Australia: Environment Protection Authority.

Li, T.-S., Tong, S.-C. and Feng, G. G. 2010. A novel robust adaptive-fuzzy-tracking control for a class of nonlinearmulti-input/multi-output systems. *Fuzzy Systems, IEEE Transactions on*, 18 (1): 150-160.

Li, X., Gao, G., Hu, T., Ma, H. and Li, T. 2014. Multiple time scales analysis of runoff series based on the Chaos Theory. *Desalination and Water Treatment*, 52 (13-15): 2741-2749.

Lin, J.-Y., Cheng, C.-T. and Chau, K.-W. 2006. Using support vector machines for long-term discharge prediction. *Hydrological Sciences Journal*, 51 (4): 599-612.

Lin, J. and Lewis, F. L. 2003. Two-time scale fuzzy logic controller of flexible link robot arm. *Fuzzy sets and systems*, 139 (1): 125-149.

Liong, S.-Y. and Sivapragasam, C. 2002. Flood Stage Forecasting With Support Vector Machines. *Journal of the American Water Resources Association*, 38 (1): 173-186.

Liong, S. Y., Gautam, T. R., Khu, S. T., Babovic, V., Keijzer, M. and Muttil, N. 2002. GENETIC PROGRAMMING: A NEW PARADIGM IN RAINFALL RUNOFF MODELING1. *JAWRA Journal of the American Water Resources Association*, 38 (3): 705-718.

Lloret, J. 2013. Underwater sensor nodes and networks. *Sensors*, 13 (9): 11782-11796.

Loch, R. and Silburn, D. 1996. Constraints to sustainability—soil erosion. *Sustainable Crop Production in the Sub-tropics: an Australian Perspective. QDPI*,

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

Lopes, V. L. and Ffolliott, P. F. 1993. Sediment rating curves for a clearcut ponderosa pine watershed in Northern Arizona. *JAWRA Journal of the American Water Resources Association*, 29 (3): 369-382.

Loucks, D. P., Van Beek, E., Stedinger, J. R., Dijkman, J. P. and Villars, M. T. 2005. *Water resources systems planning and management: an introduction to methods, models and applications*. Paris: UNESCO.

Lu, H.-W., Huang, G. H. and He, L. 2010. Development of an interval-valued fuzzy linear-programming method based on infinite α-cuts for water resources management. *Environmental Modelling & Software*, 25 (3): 354-361.

Luke, S. and Panait, L. 2001. A survey and comparison of tree generation algorithms. In: Proceedings of *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2001)*. 81-88.

Maier, H., Kapelan, Z., Kasprzyk, J., Kollat, J., Matott, L., Cunha, M., Dandy, G., Gibbs, M., Keedwell, E. and Marchi, A. 2014. Evolutionary algorithms and other metaheuristics in water resources: current status, research challenges and future directions. *Environmental Modelling & Software*, 62: 271-299.

Maier, H. R. and Dandy, G. C. 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environmental modelling & software*, 15 (1): 101-124.

Maier, H. R., Jain, A., Dandy, G. C. and Sudheer, K. P. 2010. Methods used for the development of neural networks for the prediction of water resource variables in river systems: current status and future directions. *Environmental Modelling & Software*, 25 (8): 891-909.

Makkeasorn, A., Chang, N.-B. and Li, J. 2009. Seasonal change detection of riparian zones with remote sensing images and genetic programming in a semi-arid watershed. *Journal of Environmental Management*, 90 (2): 1069-1080.

Mamdani, E. H. 1974. Application of fuzzy algorithms for control of simple dynamic plant. In: Proceedings of *Proceedings of the Institution of Electrical Engineers*. IET, 1585-1588.

May, R. J., Maier, H. R., Dandy, G. C. and Fernando, T. 2008. Non-linear variable selection for artificial neural networks using partial mutual information. *Environmental Modelling & Software*, 23 (10): 1312-1326.

McDermott, J., White, D. R., Luke, S., Manzoni, L., Castelli, M., Vanneschi, L., Jaskowski, W., Krawiec, K., Harper, R. and De Jong, K. 2012. Genetic programming needs better benchmarks. In: Proceedings

of *Proceedings of the fourteenth international conference on Genetic and evolutionary computation conference*. ACM, 791-798.

Meade, R. H., Yuzyk, T. R. and Day, T. J. 1990. Movement and storage of sediment in rivers of the United States and Canada. *In: Wolman, M.G., Riggs, H.C. (Eds.), The Geology of North America-Surface Water Hydrology vol. 1. Geological Society of America, Boulder, Colorado.*: 255-280.

Melesse, A. M., Ahmad, S., McClain, M. E., Wang, X. and Lim, Y. H. 2011. Suspended sediment load prediction of river systems: An artificial neural network approach. *Agricultural Water Management*, 98 (5): 855-866.

Mellit, A. and Kalogirou, S. A. 2008. Artificial intelligence techniques for photovoltaic applications: A review. *Progress in Energy and Combustion Science*, 34 (5): 574-632.

Merritt, W. S., Letcher, R. A. and Jakeman, A. J. 2003. A review of erosion and sediment transport models. *Environmental Modelling & Software*, 18 (8): 761-799.

Meshgi, A., Schmitter, P., Chui, T. F. M. and Babovic, V. 2015. Development of a modular streamflow model to quantify runoff contributions from different land uses in tropical urban environments using Genetic Programming. *Journal of Hydrology*, 525 (0): 711-723.

Miller, J. F. 1999. An empirical study of the efficiency of learning boolean functions using a cartesian genetic programming approach. In: Proceedings of *Proceedings of the Genetic and Evolutionary Computation Conference*. 1135-1142.

Mitra, A. P., Almal, A. A., George, B., Fry, D. W., Lenehan, P. F., Pagliarulo, V., Cote, R. J., Datar, R. H. and Worzel, W. P. 2006. The use of genetic programming in the analysis of quantitative gene expression profiles for identification of nodal status in bladder cancer. *BMC cancer*, 6 (1): 159.

Modaresi, F. and Araghinejad, S. 2014. A Comparative Assessment of Support Vector Machines, Probabilistic Neural Networks, and K-Nearest Neighbor Algorithms for Water Quality Classification. *Water Resources Management*, 28 (12): 4095-4111.

Mohsenifar, N., Mohsenifar, N. and Mohsenifar, K. 2011. Using Artificial Neural Network (ANN) for Estimating Rainfall Relationship with River Pollution. *Advances in Environmental Biology*, 5 (6): 1202-1208.

Morehead, M. D., Syvitski, J. P., Hutton, E. W. and Peckham, S. D. 2003. Modeling the temporal variability in the flux of sediment from ungauged river basins. *Global and Planetary Change*, 39 (1): 95-110.

Morgan, R. P. C. 2009. *Soil erosion and conservation*. Chichester, UK: John Wiley & Sons.

Morris, G. L., Annandale, G. and Hotchkiss, R. 2008. Reservoir sedimentation. *Sedimentation engineering process measurements, modeling and practices. Amer Soc Civil Engineer (ASCE), Reston, USA*: 1-17.

Mugumo, M. 2012. A simple operating model of the Van der Kloof Reservoir using ANN streamflow forecasts. Masters Dissertation University of Witwatersrand, Johannesburg, South Africa.

Mukundan, R., Pradhanang, S. M., Schneiderman, E. M., Pierson, D. C., Anandhi, A., Zion, M. S., Matonse, A. H., Lounsbury, D. G. and Steenhuis, T. S. 2013. Suspended sediment source areas and future climate impact on soil erosion and sediment yield in a New York City water supply watershed, USA. *Geomorphology*, 183: 110-119.

Mulvaney, T. 1851. On the use of self-registering rain and flood gauges in making observations of the relations of rainfall and flood discharges in a given catchment. *Proceedings of the institution of Civil Engineers of Ireland*, 4 (2): 18-33.

Muttil, N. and Chau, K.-W. 2006. Neural network and genetic programming for modelling coastal algal blooms. *International Journal of Environment and Pollution*, 28 (3): 223-238.

Muttil, N. and Lee, J. H. 2005. Genetic programming for analysis and real-time prediction of coastal algal blooms. *Ecological modelling*, 189 (3): 363-376.

Naik, T. R. and Dabhi, V. K. 2013. Improving generalization ability of genetic programming: comparative study. *Journal of Bioinformatics and Intelligent Control*, 2 (4): 243-252.

Nash, J. and Barsi, B. I. 1983. A hybrid model for flow forecasting on large catchments. *Journal of Hydrology*, 65 (1): 125-137.

Nasseri, M., Moeini, A. and Tabesh, M. 2011. Forecasting monthly urban water demand using Extended Kalman Filter and Genetic Programming. *Expert Systems with Applications*, 38 (6): 7387-7395.

Natale, L. and Todini, E. 1976. A stable estimator for linear models: 1. Theoretical development and Monte Carlo Experiments. *Water Resources Research*, 12 (4): 667-671.

Navratil, O., Legout, C., Gateuille, D., Esteves, M. and Liebault, F. 2010. Assessment of intermediate fine sediment storage in a braided river reach (southern French Prealps). *Hydrological processes*, 24 (10): 1318-1332.

Nayak, P., Venkatesh, B., Krishna, B. and Jain, S. K. 2013. Rainfall-runoff modeling using conceptual, data driven, and wavelet based computing approach. *Journal of Hydrology*, 493: 57-67.

Nikoo, M. R. and Mahjouri, N. 2013. Water quality zoning using probabilistic support vector machines and self-organizing maps. *Water resources management*, 27 (7): 2577-2594.

Noori, R., Karbassi, A., Moghaddamnia, A., Han, D., Zokaei-Ashtiani, M., Farokhnia, A. and Gousheh, M. G. 2011. Assessment of input variables determination on the SVM model performance using PCA, Gamma test, and forward selection techniques for monthly stream flow prediction. *Journal of Hydrology*, 401 (3): 177-189.

Nourani, V., Ejlali, R. G. and Alami, M. T. 2011. Spatiotemporal groundwater level forecasting in coastal aquifers by hybrid artificial neural network-geostatistics model: a case study. *Environmental Engineering Science*, 28 (3): 217-228.

Nowak, K., Prairie, J., Rajagopalan, B. and Lall, U. 2010. A nonparametric stochastic approach for multisite disaggregation of annual to daily streamflow. *Water Resources Research*, 46 (8)

Nunkesser, R., Bernholt, T., Schwender, H., Ickstadt, K. and Wegener, I. 2007. Detecting high-order interactions of single nucleotide polymorphisms using genetic programming. *Bioinformatics*, 23 (24): 3280-3288.

O'Neill, M., Vanneschi, L., Gustafson, S. and Banzhaf, W. 2010. Open issues in genetic programming. *Genetic Programming and Evolvable Machines*, 11 (3-4): 339-363.

Oakley, H. 1994. Two scientific applications of genetic programming: Stack filters and non-linear equation fitting to chaotic data. *Advances in genetic programming*: 369-389.

Odan, F. K. and Reis, L. F. R. 2012. Hybrid water demand forecasting model associating artificial neural network with fourier series. *Journal of Water Resources Planning and Management*, 138 (3): 245-256.

Olyaie, E., Banejad, H., Chau, K.-W. and Melesse, A. M. 2015. A comparison of various artificial intelligence approaches performance for estimating suspended sediment load of river systems: a case study in United States. *Environmental monitoring and assessment*, 187 (4): 1-22.

Ömer Faruk, D. 2010. A hybrid neural network and ARIMA model for water quality time series prediction. *Engineering Applications of Artificial Intelligence*, 23 (4): 586-594.

Opricovic, S. 2011. Fuzzy VIKOR with an application to water resources planning. *Expert Systems with Applications*, 38 (10): 12983-12990.

Orouji, H., Haddad, O. B., Fallah-Mehdipour, E. and Marino, M. A. 2014. Flood routing in branched river by genetic programming. *Water management*, 167 (2): 115-123.

Oyebode, O., Adeyemo, J. and Otieno, F. 2014. Monthly stream flow prediction with limited hydro-climatic variables in the upper Mkomazi River, South Africa using genetic programming. *Fresenius Environmental Bulletin*, 23 (3): 708-719.

Oyebode, O. K. 2014. Modelling streamflow response to hydro-climatic variables in the Upper Mkomazi River, South Africa. Master's dissertation, Durban University of Technology, Durban, South Africa.

Ozgur, K. and Jalal, S. 2012. River suspended sediment estimation by climatic variables implication:Comparative study among soft computing techniques. *Computers & Geosciences*, 43: 73-82.

Park, E. and Latrubesse, E. M. 2014. Modeling suspended sediment distribution patterns of the Amazon River using MODIS data. *Remote Sensing of Environment*, 147: 232-242.

Partal, T. and Cigizoglu, H. K. 2008. Estimation and forecasting of daily suspended sediment data using wavelet–neural networks. *Journal of Hydrology*, 358 (3): 317-331.

Perrin, C., Michel, C. and Andréassian, V. 2001. Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. *Journal of Hydrology*, 242 (3): 275-301.

Petke, J., Harman, M., Langdon, W. B. and Weimer, W. 2014. Using genetic improvement and code transplants to specialise a C++ program to a problem class. In: *Genetic Programming*. Springer, 137-149.

Phillips, J., Webb, B., Walling, D. and Leeks, G. 1999. Estimating the suspended sediment loads of rivers in the LOIS study area using infrequent samples. *Hydrological processes*, 13 (7): 1035-1050.

Piotrowski, A. P. and Napiorkowski, J. J. 2011. Optimizing neural networks for river flow forecasting – Evolutionary Computation methods versus the Levenberg–Marquardt approach. *Journal of Hydrology*, 407 (1-4): 12-27.

Poli, R. 2001. Exact Schema Theory for Genetic Programming and Variable-length Genetic Algorithms with One-Point crossover *Genetic Programming and Evolvable Machines*, 2 (23): 163.

Poli, R. and Koza, J. 2014. Genetic Programming. In: Burke, E. K. and Kendall, G. eds. *Search Methodologies*. Springer US, 143-185. Available: http://dx.doi.org/10.1007/978-1-4614-6940-7_6 (Accessed 02/03/2015).

Poli, R., Langdon, W. B., McPhee, N. F. and Koza, J. R. 2008. *A field guide to genetic programming*. Morrisville, NC: Lulu Press.

Poli, R., Vanneschi, L., Langdon, W. B. and McPhee, N. F. 2010. Theoretical results in genetic programming:the next ten year? *Genetic Programming and Evolvable Machines*, (11): 285–320.

Porter, M., Willis, M. and Hiden, H. G. 1996. Computer- Aided Polymer Design Using Genetic Programming. Master's dissertation, University of Newcastle, United Kingdom.

Pradhan, N. R., Downer, C. W. and Johnson, B. E. 2014. A physics based hydrologic modeling approach to simulate non-point source pollution for the purposes of calculating TMDLs and designing abatement measures. In: *Practical Aspects of Computational Chemistry III*. Springer, 249-282.

Quilbé, R., Rousseau, A. N., Duchemin, M., Poulin, A., Gangbazo, G. and Villeneuve, J.-P. 2006. Selecting a calculation method to estimate sediment and nutrient loads in streams: application to the Beaurivage River (Québec, Canada). *Journal of Hydrology*, 326 (1): 295-310.

Rabunal, J., Puertas, J., Suarez, J. and Rivero, D. 2007. Determination of the unit hydrograph of a typical urban basin using genetic programming and artificial neural networks. *Hydrological processes*, 21 (4): 476-485.

Raghunath, H. M. 2006. *Hydrology: principles, analysis and design*. New Delhi: New Age International.

Rajaee, T., Nourani, V., Zounemat-Kermani, M. and Kisi, O. 2010. River suspended sediment load prediction: Application of ANN and wavelet conjunction model. *Journal of Hydrologic Engineering*, 16 (8): 613-627.

Rajagopalan, B. and Lall, U. 1999. A k-nearest-neighbor simulator for daily precipitation and other weather variables. *Water Resources Research*, 35 (10): 3089-3101.

Ran, L., Lu, X., Xin, Z. and Yang, X. 2013. Cumulative sediment trapping by reservoirs in large river basins: a case study of the Yellow River basin. *Global and Planetary Change*, 100: 308-319.

Rani, D. and Moreira, M. M. 2010. Simulation–optimization modeling: a survey and potential application in reservoir systems operation. *Water Resources Management*, 24 (6): 1107-1138.

Rauss, P. J., Daida, J. M. and Chaudhary, S. 2000. Classification of spectral imagery using genetic programming. *Ann Arbor*, 1001: 48109.

Rechenberg, I. 1994. Evolution strategy. In: Zurada, J. M., Marks II, R. J. and C. J. Robinson, C. J. eds. *Computational Intelligence: Imitating Life*, 1: 147-159.

Reed, P. M., Hadka, D., Herman, J. D., Kasprzyk, J. R. and Kollat, J. B. 2013. Evolutionary multiobjective optimization in water resources: The past, present, and future. *Advances in water resources*, 51: 438-456.

Ross, B. J., Gualtieri, A. G., Fueten, F. and Budkewitsch, P. 2005. Hyperspectral image analysis using genetic programming. *Applied Soft Computing*, 5 (2): 147-156.

Ryan, C., Collins, J. and Neill, M. O. 1998. Grammatical evolution: Evolving programs for an arbitrary language. In: *Genetic Programming*. Springer, 83-96.

Sadegh, M. and Kerachian, R. 2011. Water resources allocation using solution concepts of fuzzy cooperative games: fuzzy least core and fuzzy weak least core. *Water resources management*, 25 (10): 2543-2573.

Sadeghi, S., Mizuyama, T., Miyata, S., Gomi, T., Kosugi, K., Fukushima, T., Mizugaki, S. and Onda, Y. 2008. Development, evaluation and interpretation of sediment rating curves for a Japanese small mountainous reforested watershed. *Geoderma*, 144 (1): 198-211.

Sajjan, A. K., Gyasi-Agyei, Y. and Sharma, R. H. 2014. Modeling Grass-Cover Effects on Soil Erosion on Railway Embankment Steep Slopes. *Journal of Hydrologic Engineering*,

Salas, J. D. 1980. *Applied modeling of hydrologic time series*. Littleton CO: Water Resources Publication.

Sauer, T., Yorke, J. A. and Casdagli, M. 1991. Embedology. *Journal of statistical Physics*, 65 (3-4): 579-616.

Savic, D. A., Walters, G. A. and Davidson, J. W. 1999. A genetic programming approach to rainfall-runoff modelling. *Water Resources Management*, 13 (3): 219-231.

Schindler, R. J., Parsons, D. R., Ye, L., Hope, J. A., Baas, J. H., Peakall, J., Manning, A. J., Aspden, R. J., Malarkey, J. and Simmons, S. 2015. Sticky stuff: Redefining bedform prediction in modern and ancient environments. *Geology*, 43 (5): 399-402.

Selle, B. and Muttil, N. 2011. Testing the structure of a hydrological model using Genetic Programming. *Journal of Hydrology*, 397 (1): 1-9.

Sharma, S. K. and Chandra, P. 2010. Constructive neural networks: a review. *International Journal of Engineering Science and Technology*, 2 (12): 7847-7855.

Sheikhalipour, Z. and Hassanpour, F. 2013. Estimation of Suspended Sediment Load Using Genetic Expression Programming. *Journal homepage: http://www. ojceu. ir/main*, 292: 299.

Shepherd, T. G. 2014. Atmospheric circulation as a source of uncertainty in climate change projections. *Nature Geoscience*, 7: 703-708.

Shiri, J. and Kişi, Ö. 2011. Comparison of genetic programming with neuro-fuzzy systems for predicting short-term water table depth fluctuations. *Computers & Geosciences*, 37 (10): 1692-1701.

Shiri, J., Kişi, Ö., Landeras, G., López, J. J., Nazemi, A. H. and Stuyt, L. C. 2012. Daily reference evapotranspiration modeling by using genetic programming approach in the Basque Country (Northern Spain). *Journal of Hydrology*, 414: 302-316.

Singh, K. P., Basant, N. and Gupta, S. 2011. Support vector machines in water quality management. *Analytica chimica acta*, 703 (2): 152-162.

Singh, V. P. 1995. *Computer models of watershed hydrology*. Littleton, CO: Water Resources Publications.

Singh, V. P. and Woolhiser, D. A. 2002. Mathematical modeling of watershed hydrology. *Journal of hydrologic engineering*, 7 (4): 270-292.

Sirdari, Z. Z., Ghani, A. A. and Hasan, Z. A. 2012. Prediction of bed load transport in Kurau River based on genetic programming. In: Proceedings of *3rd International Conference on Managing Rivers in the 21st century: Sustainable Solutions for Global Crisis of Flooding, Pollution and Water Scarcity*. Penang, Malaysia. , 6-9 December 2011. 385-390.

Sivakumar, B. and Berndtsson, R. 2010. Nonlinear Dynamic and Chaos in Hydrology. In: Sivakumar, B. and Berndtsson, R. eds. Advances in data-based approaches for hydrologic modeling and forecasting.

. *World Scientific*: 411-461.

Sivakumar, B., Berndtsson, R., Olsson, J. and Jinno, K. 2001. Evidence of chaos in the rainfall-runoff process. *Hydrological Sciences Journal*, 46 (1): 131-145.

Sivapragasam, C., Arun, V. and Muttil, N. 2011. Re-design of rain gauge network using genetic programming based ordinary Kriging. In: Proceedings of *Proceedings of the 34th World Congress of the International Association for Hydro-Environment Research and Engineering: 33rd Hydrology and Water Resources Symposium and 10th Conference on Hydraulics in Water Engineering*. Engineers Australia, 428-433.

Sivapragasam, C., Maheswaran, R. and Venkatesh, V. 2008. Genetic programming approach for flood routing in natural channels. *Hydrological processes*, 22 (5): 623-628.

Smith, J. and Eli, R. N. 1995. Neural-network models of rainfall-runoff process. *Journal of water resources planning and management*, 121 (6): 499-508.

Smith, R. J. 2008. Logarithmic transformation bias in allometry. *American Journal of Physical Anthropology*, 90 (2): 215-228.

Smith, S. F. 1980. A learning system based on genetic adaptive algorithms. Doctoral thesis, University of Pittsburgh, USA.

Solomatine, D. and Ostfeld, A. 2008. Data-driven modelling: some past experiences and new approaches. *Journal of hydroinformatics*, 10 (1): 3-22.

Solomatine, D., Rojas, C., Velickov, S. and Wust, H. 2000. Chaos theory in predicting surge water levels in the North Sea. In: Proceedings of *Proc. 4th Int. Conference on Hydroinformatics, Cedar-Rapids*. Cedar-Rapids, USA, July 2000. Citeseer,

Song, J., Xiang, B., Wang, X., Wu, L. and Chang, C. 2014. Application of dynamic data driven application system in environmental science. *Environmental Reviews*, 22 (999): 287-297.

Sorooshian, S. 1991. Parameter estimation, model identification, and model validation: conceptual-type models. In: *Recent advances in the modeling of hydrologic systems*. Springer, 443-467.

Spear, R. C. 1997. Large simulation models: calibration, uniqueness and goodness of fit. *Environmental Modelling & Software*, 12 (2): 219-228.

Sreekanth, J. and Datta, B. 2011. Comparative evaluation of genetic programming and neural network as potential surrogate models for coastal aquifer management. *Water resources management*, 25 (13): 3201-3218.

St-Hilaire, A., Ouarda, T. B., Bargaoui, Z., Daigle, A. and Bilodeau, L. 2012. Daily river water temperature forecast model with ak-nearest neighbour approach. *Hydrological Processes*, 26 (9): 1302-1310.

Steinwart, I. and Christmann, A. 2008. *Support vector machines*. New York, USA: Springer.

Sudheer, K., Gosain, A. and Ramasastri, K. 2002. A data-driven algorithm for constructing artificial neural network rainfall-runoff models. *Hydrological Processes*, 16 (6): 1325-1330.

Sugeno, M. and Kang, G. 1988. Structure identification of fuzzy model. *Fuzzy sets and systems*, 28 (1): 15-33.

Syvitski, J. P. and Milliman, J. D. 2007. Geology, geography, and humans battle for dominance over the delivery of fluvial sediment to the coastal ocean. *The Journal of Geology*, 115 (1): 1-19.

Syvitski, J. P., Morehead, M. D., Bahr, D. B. and Mulder, T. 2000. Estimating fluvial sediment transport: the rating parameters. *Water Resources Research*, 36 (9): 2747-2760.

Syvitski, J. P., Vörösmarty, C. J., Kettner, A. J. and Green, P. 2005. Impact of humans on the flux of terrestrial sediment to the global coastal ocean. *Science*, 308 (5720): 376-380.

Takagi, T. and Sugeno, M. 1985. Fuzzy identification of systems and its applications to modeling and control. *Systems, Man and Cybernetics, IEEE Transactions on Systems Man and Cybernetics*, (1): 116-132.

Takens, F. 1981. Detecting strange attractors in turbulence. In: *Dynamical systems and turbulence, Warwick 1980*. Springer, 366-381.

Taormina, R., Chau, K.-w. and Sethi, R. 2012. Artificial neural network simulation of hourly groundwater levels in a coastal aquifer system of the Venice lagoon. *Engineering Applications of Artificial Intelligence*, 25 (8): 1670-1676.

Tayfur, G. 2014. *Soft computing in water resources engineering: Artificial neural networks, fuzzy logic and genetic algorithms*. Ashurst Lodge, Ashurst, Southampton, UK WIT Press.

Thorsen, M., Refsgaard, J., Hansen, S., Pebesma, E., Jensen, J. and Kleeschulte, S. 2001. Assessment of uncertainty in simulation of nitrate leaching to aquifers at catchment scale. *Journal of Hydrology*, 242 (3): 210-227.

Tokar, A. S. and Johnson, P. A. 1999. Rainfall-runoff modeling using artificial neural networks. *Journal of Hydrologic Engineering*, 4 (3): 232-239.

Tollow, A. J. 2004. *An approach to resource management using umgeni water system as an example.* Available: http://www.hydrology.org.uk/Publications/imperial/6_20 (Accessed 15/12/2014).

Tongal, H. and Berndtsson, R. 2014. Phase-space reconstruction and self-exciting threshold modeling approach to forecast lake water levels. *Stochastic Environmental Research and Risk Assessment*, 28 (4): 955-971.

Turing, A. M. 1950. Computing machinery and intelligence. *Mind*, 59 (236): 433-460.

Turner, R. E., Baustian, J. J., Swenson, E. M. and Spicer, J. S. 2006. Wetland sedimentation from hurricanes Katrina and Rita. *Science*, 314 (5798): 449-452.

Vas, P. 1999. *Artificial-intelligence-based electrical machines and drives: application of fuzzy, neural, fuzzy-neural, and genetic-algorithm-based techniques*. Oxford University Press.

Walker, M. 2001. Introduction to genetic programming. *Tech. Np: University of Montana,USA*,

Walling, D. 1977. Limitations of the rating curve technique for estimating suspended sediment loads, with particular reference to British rivers. *Erosion and solid matter transport in inland waters*: 34-48.

Walling, D. and Webb, B. 1988. The reliability of rating curve estimates of suspended sediment yield: some further comments. IN: Sediment Budgets. In: Proceedings of *Porto Alegre Symposium*. December 1988. IAHS Publication,

Wang, C., Jiang, R., Mao, X., Sauvage, S., Sanchez-Perez, J.-M., Woli, K. P., Kuramochi, K., Hayakawa, A. and Hatano, R. 2015. Estimating sediment and particulate organic nitrogen and particulate organic phosphorous yields from a volcanic watershed characterized by forest and agriculture using SWAT model. *International Journal of Limnology*, 51 (1): 23-35.

Wang, H., Yang, Z., Wang, Y., Saito, Y. and Liu, J. P. 2008. Reconstruction of sediment flux from the Changjiang (Yangtze River) to the sea since the 1860s. *Journal of Hydrology*, 349 (3): 318-332.

Wang, S. and Huang, G. 2012. Identifying optimal water resources allocation strategies through an interactive multi-stage stochastic fuzzy programming approach. *Water resources management*, 26 (7): 2015-2038.

Wang, W.-C., Chau, K.-W., Cheng, C.-T. and Qiu, L. 2009. A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series. *Journal of Hydrology*, 374 (3): 294-306.

Waters, K. A. and Curran, J. C. 2015. Linking bed morphology changes of two sediment mixtures to sediment transport predictions in unsteady flows. *Water Resources Research*, 51 (4): 2724-2741.

Westerberg, C. H. and Levine, J. 2014. Optimising plans using genetic programming. In: Proceedings of *Sixth European Conference on Planning*. AAAI Publication,

Wheater, H., Jakeman, A. and Beven, K. 1993. Progress and directions in rainfall-runoff modelling. In Jakeman, A. J., Beck, M.B., and McAleer, M. J. eds. *Modelling change in environment systems*: 101-132.

Wijayaweera, W. and Karunananda, A. 2012. Framework for Discovery of Data Models Using Genetic Programming. Paper presented at the *Sri Lanka Association for Artificial Intelligence (SLAAI)Proceeding of the ninth Annual Sessions*. Sri Lanka, 18th December 2012. The Open University.

Winz, I., Brierley, G. and Trowsdale, S. 2009. The use of system dynamics simulation in water resources management. *Water resources management*, 23 (7): 1301-1323.

Wu, C. 2010. Hydrological predictions using data-driven models coupled with data preprocessing techniques. Doctoral Thesis, The Hong Kong Polytechnic University.

Wu, C. and Chau, K. 2011. Rainfall–runoff modeling using artificial neural network coupled with singular spectrum analysis. *Journal of Hydrology*, 399 (3): 394-409.

Xiao, C., Shao, D., Yang, F., Gu, W. and Wang, K. 2011. A new coupled chaos optimization-projection pursuit model for initial water rights allocation in the watershed. In: Proceedings of *2011 International Symposium on Water Resource and Environmental Protection (ISWREP)*. IEEE, 3100-3104.

Xu, K., Chen, Z., Zhao, Y., Wang, Z., Zhang, J., Hayashi, S., Murakami, S. and Watanabe, M. 2005. Simulated sediment flux during 1998 big-flood of the Yangtze (Changjiang) River, China. *Journal of Hydrology*, 313 (3): 221-233.

Xu, Q., Chen, Q. and Li, W. 2011. Application of genetic programming to modeling pipe failures in water distribution systems. *Journal of Hydroinformatics*, 13 (3): 419-428.

Yakowitz, S. 1987. NEAREST-NEIGHBOUR METHODS FOR TIME SERIES ANALYSIS. *Journal of time series analysis*, 8 (2): 235-247.

Yang, D., Herath, S. and Musiake, K. 1998. Development of a geomorphology-based hydrological model for large catchments. *Annual Journal of Hydraulic Engineering, JSCE*, 42: 169-174.

Yang, G., Chen, Z., Yu, F., Wang, Z., Zhao, Y. and Wang, Z. 2007. Sediment rating parameters and their implications: Yangtze River, China. *Geomorphology*, 85 (3): 166-175.

Yang, S.-l., Zhao, Q.-y. and Belkin, I. M. 2002. Temporal variation in the sediment load of the Yangtze River and the influences of human activities. *Journal of Hydrology*, 263 (1): 56-71.

Yegnanarayana, B. 2009. *Artificial neural networks*. New Delhi, India: PHI Learning Pvt. Ltd.

Yu, P.-S., Chen, S.-T. and Chang, I.-F. 2006. Support vector regression for real-time flood stage forecasting. *Journal of Hydrology*, 328 (3): 704-716.

Zeng, Y., Cai, Y., Jia, P. and Jee, H. 2012. Development of a web-based decision support system for supporting integrated water resources management in Daegu city, South Korea. *Expert Systems with Applications*, 39 (11): 10091-10102.

Zhang, L., B Jack, L. and Nandi, A. K. 2005. Fault detection using genetic programming. *Mechanical Systems and Signal Processing*, 19 (2): 271-289.

Zhao, Y., Chen, F., Shen, Q. and Zhang, L. 2014. Optimal design of graded refractive index profile for broadband omnidirectional antireflection coatings using genetic programming. *Progress In Electromagnetics Research*, 145: 39-48.

Zyserman, J. A. and Fredsøe, J. 1994. Data analysis of bed concentration of suspended sediment. *Journal of Hydraulic Engineering*, 120 (9): 1021-1042.