

CAPITAL UNIVERSITY OF SCIENCE AND
TECHNOLOGY, ISLAMABAD



Development of Model-Data Fusion System for Upper Indus Basin Stream-flow

by

Muhammad Hassan

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the

Faculty of Engineering

Department of Civil Engineering

2022

Development of Model-Data Fusion System for Upper Indus Basin Stream-flow

By

Muhammad Hassan

(DCE153009)

Dr. Assefa M. Melesse, Professor
Florida International University, USA
(Foreign Evaluator 1)

Dr. Shafiqul Islam, Professor
Tufts University, Medford, USA
(Foreign Evaluator 2)

Dr. Ishtiaq Hassan
(Thesis Supervisor)

Dr. Ishtiaq Hassan
(Head, Department of Civil Engineering)

Dr. Imtiaz Ahmad Taj
(Dean, Faculty of Engineering)

DEPARTMENT OF CIVIL ENGINEERING
CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY
ISLAMABAD

2022

Copyright © 2022 by Muhammad Hassan

All rights reserved. No part of this thesis may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, or other electronic or mechanical methods, by any information storage and retrieval system without the prior written permission of the author.

DEDICATED TO MY FAMILY



CAPITAL UNIVERSITY OF SCIENCE & TECHNOLOGY ISLAMABAD

Expressway, Kahuta Road, Zone-V, Islamabad
Phone: +92-51-111-555-666 Fax: +92-51-4486705
Email: info@cust.edu.pk Website: <https://www.cust.edu.pk>

CERTIFICATE OF APPROVAL

This is to certify that the research work presented in the thesis, entitled “**Development of Model-Data Fusion System for Upper Indus Basin Stream-Flow**” was conducted under the supervision of **Dr. Ishtiaq Hassan**. No part of this thesis has been submitted anywhere else for any other degree. This thesis is submitted to the **Department of Civil Engineering, Capital University of Science and Technology** in partial fulfillment of the requirements for the degree of Doctor in Philosophy in the field of **Civil Engineering**. The open defence of the thesis was conducted on **July 27, 2022**.

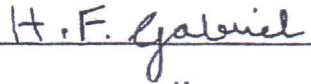
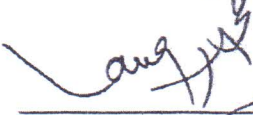
Student Name : Muhammad Hassan
(DCE-153009)



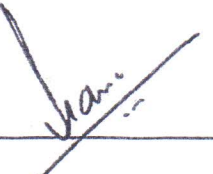
The Examination Committee unanimously agrees to award PhD degree in the mentioned field.

Examination Committee :

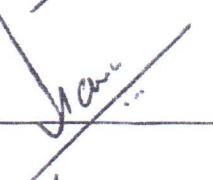
- (a) External Examiner 1: Dr. Hamza Farooq Gabriel,
Professor
NUST, Islamabad
- (b) External Examiner 2: Dr. Naeem Ejaz
Professor
UET Taxila
- (c) Internal Examiner : Dr. Muhammad Usman Farooqi
Assistant Professor
CUST, Islamabad


_____
_____

Supervisor Name : Dr. Ishtiaq Hassan
Professor
CUST, Islamabad



Name of HoD : Dr. Ishtiaq Hassan
Professor
CUST, Islamabad



Name of Dean : Dr. Imtiaz Ahmed Taj
Professor
CUST, Islamabad

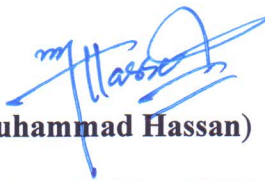


AUTHOR'S DECLARATION

I, **Muhammad Hassan (Registration No. DCE-153009)**, hereby state that my PhD thesis titled, '**Development of Model-Data Fusion System for Upper Indus Basin Stream-Flow**' is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/ world.

At any time, if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my PhD Degree.

Dated: July, 2021



(Muhammad Hassan)

Registration No : DCE153009


PLAGIARISM UNDERTAKING

I solemnly declare that research work presented in the thesis titled “**Development of Model-Data Fusion System for Upper Indus Basin Stream-Flow**” is solely my research work with no significant contribution from any other person. Small contribution/ help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/ cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of PhD Degree, the University reserves the right to withdraw/ revoke my PhD degree and that HEC and the University have the right to publish my name on the HEC/ University Website on which names of students are placed who submitted plagiarized thesis.

Dated: July, 2022


(Muhammad Hassan)
Registration No : DCE153009

List of Publications

It is certified that following publication(s) have been made out of the research work that has been carried out for this thesis:-

1. **Hassan, M.**, Hassan,I. “Improving ANN-based streamflow estimation models for the Upper Indus Basin using satellite-derived snow cover area.” *Acta Geophys.* 68, 17911801 (2020). <https://doi.org/10.1007/s11600-020-00491-4>.
2. **Hassan, M.**, Hassan, I. “Improving Artificial Neural Network based Streamflow Forecasting Models through Data preprocessing”. *KSCE Journal of Civil Engineering.* 25, 3583-3595 (2021). <https://doi.org/10.1007/s12205-021-1859-y>

(Muhammad Hassan)

Registration No: DCE153009

Acknowledgement

In the name of **ALLAH**, the most beneficent, the most Merciful. All praises to Him (Alhmdulillah). I am very thankful to Him for His help and countless blessings. May we stay blessed and keep praising to Him (Amen). I am very grateful to my supervisor **Engr. Dr. Ishtiaq Hassan** for his help and sincere guidance throughout my Ph.D. Without his help and continuous effort, this would not be possible. His helping attitude and encouraging behavior supports me to complete this task in time.

I am thankful to my parents, whose support, prayers and help is indefinite. I am grateful to my wife for believing in me and standing with me in my every decision with full support and trust. I am also thankful to my parents in-law for sacrificing their invalueable time to support my family. I want to thank all my friends for their help and support, especially I am grateful to Engr. Imran Mehmood for his brotherly care during my Ph.D.

I am also thankful to my teacher and mentor, late Dr. Muhammad Ali Shamim for his teachings and lifelong skills. At the end, I want to acknowledge the Surface Water Hydrology Department (SWHP), Water and Power Development Authority (WAPDA) for providing all the necessary data to carryout this researchwork.

(Muhammad Hassan)

Abstract

Hydrological Processes are complex and highly nonlinear because of their dependency upon multiple climate and hydrological variables. Modeling the complexity of these processes is quite challenging because of the many factors that may hinder the efficiency of models in capturing the relationship between these variables and response of the catchment. These factors may include; complex terrain, contrasting regimes, limited meteorological network, noise present in the data and confined economical resources. The Indus Basin is the main source of water for Pakistan. Almost 80% of water requirement of this basin is derived by the Upper Indus Basin (UIB). The UIB has many challenges: which include extreme complexity, varying hydro-meteo-cryospheric regimes, limited meteorological network, climate change pattern and the spread of UIB over political sensitive trans-boundary area. The contrasting regimes in different part of basin is the result that the response of the catchment is difficult to capture. Therefore, the trend discrepancies and model uncertainties in the UIB exist, which is reflected in the previous literature as well. The present research work is carried out by focusing; the larger part of the UIB, incorporating multi type/ source data, and applying data preprocessing techniques to minimize the uncertainties in the UIB streamflow measurement.

The aim of the research work is to develop Artificial Neural Network (ANN) based hydrological models that can efficiently estimate the streamflow in the Pakistani part of the UIB. A systematic approach is adopted to improve the different steps involved in the hydrological modeling process, which involves data improvement, data selection and data fusion. This ultimately leads to a development of model data-fusion system for the region that optimizes the performance of ANN based streamflow models. The research work is divided into three (03) parts with a main focus on improving ANN based streamflow estimation models for the UIB through; 1. Data preprocessing, 2. By incorporating satellite derived Snow Cover Area (SCA), and 3. Utilizing data fusion. Two-step data preprocessing is performed, which includes data transformation through Box-Cox transformation and input selection through Gamma Test. Satellite derived SCA is utilized in combination with the on-ground flow observations to enhance the performance efficiency

of the streamflow estimation models in the region. The ANN models are also developed using a variety of data combinations which are made either on the basis of type/nature of climate variable or through advanced input/feature selection methods.

The results indicated; the models developed through data preprocessing performed well as compared to the models developed with original data-set, with more than 90% correlation coefficient in both training and testing phases. The flow dependency on satellite derived SCA of UIB region is clearly evidenced with the improved average values of Nash Sutcliffe Efficiency (NSE) = 99.5/97.5 (training/testing), BIAS = -0.01/-6.6, Root Mean Squared Error (RMSE) = 251.4/532.3 and Variance (VAR) = 63218.0/286917.1 for the models developed using SCA in combination with the other on-ground observations, as compared to the NSE = 99.1/97.1 (training/testing), BIAS = 14.6/-26.1, RMSE = 327.6/531.4 and VAR = 106390.6/284363.4 for models developed using on-ground observations without SCA. The improvement in ANN based models through feature selection techniques including Genetic Algorithm (GA), Hill Climbing (HC) and Sequential Embedding (SE) has been observed with better values of statistical indices (NSE and $R^2 > 0.9$), as compared to the models developed through manual selection of input variables. However, the models developed utilizing multiple climate variables like Precipitation, Discharge, Solar Radiation and SCA also performed well. Only one feature selection technique, which is Full Embedding (FE) does not provide good results with low values for R^2 , NSE and high corresponding values of other errors. Overall, the models developed through SE outperformed with $R^2 = 93.7/91.4$ (training/ testing) and NSE= 97/96. The outcomes of this research could be used to establish a comprehensive linkage between the changing climate variables and their impact on the response of the UIB. The ANN based data fusion models could be applied confidently for the streamflow estimation in the region and ultimately for better management of flood mitigation and reservoir operation at downstream of Tarbela. The research work recommends the use of multi type/ source data coupled with data preprocessing to capture the non-linearity and complexity of catchments which observe contrasting regimes.

Keywords:

Artificial Neural Network (ANN), Snow Cover Area (SCA), Upper Indus Basin (UIB), Streamflow Estimation, Data Fusion.

Contents

Author’s Declaration	v
Plagiarism Undertaking	vi
List of Publications	vii
Acknowledgement	viii
Abstract	ix
List of Figures	xv
List of Tables	xviii
Abbreviations	xix
Symbols	xx
1 Introduction	1
1.1 Background	1
1.2 Water Resources Management in Pakistan	3
1.3 Indus Basin Irrigation System	5
1.4 Research Motives, Problem Statement and Research Questions . . .	6
1.4.1 Research Motives	6
1.4.2 Problem Statement	8
1.4.3 Research Questions	9
1.5 Novelty of the Research Work	9
1.6 Research Significance and Practical Applications	10
1.7 Research Objectives	11
1.8 Scope and Limitations of Research	12
1.8.1 Scope	12
1.8.2 Limitations	13
1.9 Brief Methodology	14
1.10 Thesis Organization	15

2	Literature Review	17
2.1	Background	17
2.2	Indus Basin: Climate Change Pattern and Trend Analysis	18
2.2.1	Summary	19
2.3	Upper Indus Basin Dependency on Multiple Climate Variables: Studies and Models	19
2.3.1	Importance of Satellite Derived Snow Cover Area (SCA)	22
2.3.2	Summary	24
2.4	Non-Linearity, Noise in Hydrological Data and Need for Data Pre-processing	24
2.4.1	Artificial Neural Networks (ANN)	25
2.4.2	Data Pre-processing Options and Practices for ANN	26
2.4.3	Summary	28
2.5	Data Fusion: Need and Practice in Hydrological Forecasting	29
2.5.1	Summary	34
2.6	Overall Summary	34
3	Study Area & Methodology	37
3.1	Background	37
3.2	Study Area	39
3.3	Improving ANN Based Hydrological Forecasting through Data Pre-processing	43
3.3.1	Data-set	45
3.3.2	Data Transformation	47
3.3.2.1	The Box-Cox Transformation	48
3.3.3	Input Selection through Gamma Test (GT)	50
3.3.3.1	Gamma Test	50
3.3.3.2	Working Principle of Gamma Test	51
3.3.4	Artificial Neural Networking (ANN)	54
3.3.4.1	BFGS Algorithm	55
3.3.4.2	Model Training	56
3.3.5	ANN Model Development	56
3.3.6	Performance Indicators	59
3.4	Improving ANN Based Hydrological Forecasting Through Satellite Derived Snow Cover Area (SCA)	60
3.4.1	Dataset	60
3.4.1.1	MODIS Snow Products for SCA	61
3.4.2	Input Combination and Data Length Selection	63
3.4.3	Model Training	63
3.5	Improving ANN based Hydrological Forecasting through Data Fusion	65
3.5.1	Dataset	65
3.5.2	Data Fusion Options	67
3.5.2.1	Type / Nature and Source of Data	67
3.5.2.2	Feature Selection Methods	68

3.5.2.3	Full Embedding	69
3.5.2.4	Sequential Embedding	70
3.5.2.5	Genetic Algorithm	70
3.5.2.6	Hill Climbing	71
3.5.3	ANN Model Development	71
4	Results & Discussion	73
4.1	Background	73
4.2	ANN Models developed through Data Preprocessing	73
4.2.1	Data Transformation Results	74
4.2.2	Gamma Test Results	81
4.2.3	ANN Model Results	82
4.2.4	Discussion	84
4.2.5	Summary	88
4.3	ANN Models developed using Satellite Derived Snow Cover Area	89
4.3.1	Gamma Test Results	89
4.3.2	ANN Model Results	91
4.3.3	Discussion	93
4.3.4	Summary	100
4.4	ANN Models Developed Through Data Fusion	101
4.4.1	Gamma Test Results	101
4.4.2	ANN Model Results & Discussion	106
4.4.3	Summary	117
4.5	Overall Discussion	118
5	Conclusions & Recommendations	121
5.1	General	121
5.2	Conclusions	121
5.3	Research Significance and Implications	123
5.4	Precincts of Techniques used in Study	124
5.5	Recommendations	125
	Bibliography	126
	Annex-4A	147
	Annex-4B	152
	Annex-4C	157

List of Figures

3.1	General Methodology Flow Chart	38
3.2	Upper Indus Basin: Gauging Stations (Source: WAPDA)	40
3.3	Study Area (UIB): Catchments delineation using Digital Elevation Model (DEM)	43
3.4	Methodology Flow Chart	44
3.5	Mean values of hydrological variables of different stations located in the UIB	46
3.6	Framework for ANN with 2 Hidden Layers	54
3.7	Time series of SCA for Astore, Gilgit and Bunji Catchments	62
4.1	Transformed Data Analysis for $\lambda = -1$. (a) Histogram, (b) Normal Probability Plot	76
4.2	Original Data Analysis for $\lambda = 1$. (a) Histogram, (b) Normal Probability Plot	77
4.3	Transformed Data Analysis for $\lambda = 0.1$. (a) Histogram, (b) Normal Probability Plot	78
4.4	Transformed Data Analysis for $\lambda = 0.01$. (a) Histogram, (b) Normal Probability Plot	78
4.5	Transformed Data Analysis for $\lambda = 0.005$. (a) Histogram, (b) Normal Probability Plot	79
4.6	Transformed Data Analysis for $\lambda = 0$, Log transformation (a) (Histogram) (b) (Normal Probability Plot)	80
4.7	Spread of Gamma values and V-ratio. (a) Original Data, (b) Transformed Data	81
4.8	Comparison of ANN models for Original and Transformed data	83
4.9	Model 1-1 for Original Data. (a) Training, (b) Testing	85
4.10	Model 2-2 for Transformed Data. (a) Training, (b) Testing	86
4.11	Time-series plot for models developed using original & transformed data-set (a) Original Data, (b) Transformed Data	87
4.12	Variation in Gamma Value with different masks of input variables	90
4.13	Stabilizing the Gamma Value with increasing data points for combination 010101	91
4.14	Stabilizing the Gamma Value with increasing data points for combination 111111	91
4.15	ANN Modeling Results for both set of input-combinations	92
4.16	Model No. 1 (Nodes: 1-1) developed with input combination 111111 (a) Training Phase, (b) Testing Phase	94

4.17	Model No. 7 (Nodes: 2-2) developed with input combination 010101 (a) Training Phase, (b) Testing Phase	95
4.18	Model No. 11 (Nodes: 4-4) developed for input combination 111111 (a) Training Phase, (b) Testing Phase	96
4.19	Model No. 21 (Nodes: 7-2) developed for input combination 010101 (a) Training Phase, (b) Testing Phase	97
4.20	Model No. 19 (Nodes: 6-6) developed for input combination 010101 (a) Training Phase, (b) Testing Phase	98
4.21	Time series plot for Besham Qila developed with combinationS. (a) 111111 , (b) 010101	99
4.22	Stabilizing the Gamma Value with increasing data points for com- bination no. 2	103
4.23	Stabilizing the Gamma Value with increasing data points for com- bination no. 8	103
4.24	Stabilizing the Gamma Value with increasing data points for com- bination no. 10	104
4.25	Stabilizing the Gamma Value with increasing data points for com- bination no. 12	104
4.26	Stabilizing the Gamma Value with increasing data points for com- bination no. 13	105
4.27	Stabilizing the Gamma Value with increasing data points for com- bination no. 15	106
4.28	Variation of R^2 for different input combinations	107
4.29	Variation of NSE for different input combinations	107
4.30	Variation of BIAS for different input combinations	108
4.31	Variation of RMSE for different input combinations	108
4.32	Model developed using only P (node combination 1-3). (a) Training Model, (b) Testing Model	110
4.33	Model developed using combination determined through GA (with node combination 1-1). (a) Training Model, (b) Testing Model . . .	111
4.34	Model developed using combination determined through SE (with node combination 1-1). (a) Training Model, (b) Testing Model . . .	112
4.35	Model developed using P+S+Q (with node combination 1-1). (a) Training Model, (b) Testing Model	113
4.36	Model developed using P+SCA+Q (with node combination 1-1). (a) Training Model, (b) Testing Model	114
4.37	Model developed using ALL input variables (with node combination 5-5). (a) Training Model, (b) Testing Model	115
4.38	Model developed using Q (with node combination 2-2). (a) Training Model, (b) Testing Model	116
4.39	(a) Time-series plot for models developed using combination of in- put variables which contains only Q. (b) Time-series plot for models developed using combination of input variables determined through SE	117
1	M-Test Result for combination 01 (Only P)	147
2	M-Test result for combination 03 (Only Q)	147

3	M-Test result for combination 04 (P+Q)	148
4	M-Test result for combination 05 (P+SR)	148
5	M-Test result for combination 06 (Only SCA)	149
6	M-Test result for combination 07 (SCA+Q*)	149
7	M-Test result for combination 09 (P+S+Q)	150
8	M-Test result for combination 11 (ALL)	150
9	M-Test result for combination 14 (HC)	151

List of Tables

3.1	Details of Gauging Stations and respective observations	45
3.2	Defining ANN Model Structure through a set of different node- arrangements for BFGS	58
3.3	Data set used for model development	62
3.4	Selection of nodes in hidden layers on the basis of MSE and R^2 . . .	64
3.5	Data set used for data fusion and model development	66
3.6	ANN model Architecture	72
4.1	Different values of λ against Co-variance and Histogram character- istics	74
4.2	Gamma & V_{ratio} values along with optimized data length for differ- ent Data Fusion options	101
1	R^2 Values for different architectures of ANN using multiple combi- nation options	153
2	NSE Values for different architectures of ANN using multiple com- bination options	154
3	RMSE Values for different architectures of ANN using multiple com- bination options	155
4	RBIAS Values for different architectures of ANN using multiple combination options	156
5	Results of Models developed without Solar Radiation	158
6	Results of models developed by input combinations selected through GA and Gamma Test	159

Abbreviations

ANN	Artificial Neural Networks
BFGS	Broyden Fletcher & Goldfrab Shano
EMD	Empirical Mode Decomposition
GA	Genetic Algorithm
GT	Gamma Test
HKH	Hindukash-Karakorum-Himalaya
IBIS	Indus Basin Irrigation System
IRSA	Indus River System Authority
LST	Land Surface Temperature
MODIS	Moderate Resolution Imaging Spectroradiometer
MSE	Mean Square Error
NSE	Nash Sutcliffe Efficiency
P	Precipitation
Q	Discharge
RMSE	Root Mean Square Error
SCA	Snow Cover Area
SSD	Single Spectrum Analysis
SR	Solar Radiation
SRM	Snow-melt Runoff Model
SWHP	Surface Water Hydrology Project
UIB	Upper Indus Basin
VAR	Variance
WA	Wavelet Analysis
WAPDA	Water and Power Development Authority

Symbols

X	Input Variable
Y	Output Variable
γ	Gamma Statistics/ Gamma Value
λ	Lambda, Power Factor
R^2	Coefficient of determination
f	Function
M	Points
θ & ϵ	Vectors associated with transformed value
r	Noise

Chapter 1

Introduction

1.1 Background

Predicting stream-flows is extremely important to meet water management challenges such as, flood mitigation, drought management, reservoir operation and planning of various water resource projects. Development of accurate stream-flow prediction models is a difficult task as the behavior of natural streams is complex and highly non-linear in both temporal and spatial aspects [1]. In recent past, numerous methodologies have been adopted by researchers to predict stream-flow. These can be divided in to two general types which are process based physical methods and data-driven modeling approach [2].

Process based physical models are driven by some physical process or processes, which is/are governed by the set of mathematical equations. These methods have the advantage of understanding the hydrological process with the help of physical features, however they are usually complex and often bear uncertainties due to the limited availability of physical information of the watershed [3]. They also bear the constraints of initial and boundary conditions while modeling the real world problems. According to [4], the use of such models is highly restricted for complex hydrological processes where the available data is not sufficient to explain the process. They are not data-driven, still they require data (observations) to evaluate model parameters.

On the other hand, data driven models are derived by data and there are no background equations involved for the description of data or any process. They are purely based upon the characterization of input-output data and has a limited knowledge of underlying physical mechanism [5]. The antecedent data condition is utilized to estimate the model parameters by capturing the relationship between input variable/s and targeted variable/s. Now a days, these methods are quite popular in stream-flow estimation due to their ease of development, real time implementation and minimum information requirement [6]. They are capable of predicting the stream-flow with reasonable accuracy, despite of the fact, that these methods have limited physical information of underlying hydrological process [7].

In past, traditional models including multiple regression and autoregressive moving average type linear models have been used as data driven based forecasting models. But, these models use the assumption of stationary data and provide good results only, when the data is linear and normally distributed. On the other hand, hydrological processes are highly non-linear and the observations are non-stationary. Similarly, the dependency of these processes on multiple climate factors make them even more difficult to understand. Therefore, non-linear models are required to capture the complex relationships of observations for better understanding of a hydrological process.

The nonlinear models are those which are able to capture the nonlinearity present among the data. In simple terms, the nonlinearity may be defined as the lack of direct relationship between inputs and outputs. i.e. when these variables are plotted graphically, the points do not fit in a straight line. The unequal change of one variable to the other variable is normally defined by the heteroscedasticity of the data. This is a simple interpretation of nonlinearity in hydrology and covers only one aspect of it. However, nonlinearity as a whole is a complicated phenomenon because the hydrological data is complex and normally observes a chaotic behavior.

Therefore, the nonlinear models are complex than the linear models due to the fact that there are multiple parameters involved to capture the different trends present

among the data, at the same time. The nonlinear models are usually described by a function which is developed through a series of iterations.

Similarly, the use of multiple input variables observed through different sources could be used to get a better picture of the hydrological phenomenon, especially in catchments which are complex, observe contrasting regimes, and has dependency upon multiple climate factors. The same is the case with UIB part on Pakistans side, which being a main source of water for the country, has many challenges to combat water shortage issues and maximize water control efficiency in the water management sector of Pakistan.

1.2 Water Resources Management in Pakistan

Pakistan is an agricultural country and its economy is highly dependent upon agricultural growth of the region. More than 70% of its population is attached to this sector. The contribution of this sector in Gross Domestic Product (GDP) of Pakistan is about 24% [8]. This is why agriculture is considered as one of the main contributor towards economic growth of the region.

Agricultural growth is mainly dependent upon the availability of water. More than 90% of Pakistan's available water is utilized for agriculture [9]. Therefore, the continuous availability of water for crops is crucial for the sustainability and stability of this sector. On the other hand, Pakistan is facing water availability issues and listed in one of the water scarce country with per capita water availability $< 1000 m^3/s$.

The main source of water required for Pakistan is The Mighty River Indus and is often called as the bread basket for Pakistan. It is one of the longest rivers of Asia with more than 3000 km length. Indus Basin Irrigation System (IBIS) is one of the world's largest canal network system. This integrated system irrigates a total area of 22 Million Hectares and contributes about 85% of all food production of country. It supports irrigated agriculture and in a result provides food security to

the people of Pakistan [10]. Not only this, Hydropower potential available in the form of water storage reservoirs also ensures economical energy production.

The management of IBIS has been mainly achieved through construction of two main storage dams Mangla and Tarbela along with a comprehensive inter-river canal system. This wide spread system covers a huge geographical area and established to distribute the water on equitable basis.

The irrigation intensity has increased enormously over time with increase in population and industrial growth. Thus creating desire for high efficiency systems and more need of water. But unfortunately the development in IBIS lags behind this increased water demand and created even more call for changing water use habits and moving towards sustainable approach [11].

Water resources play a fundamental role in the economic development of a country, specifically a developing country like Pakistan. The region's speedy population growth is a serious alarm and resulting more demands of limited water resources require proficient management of existing water resources rather than constructing new amenities to meet the challenge.

In the water management communities, it is well known that to combat water paucity issues, maximizing water management efficiency based on streamflow forecasting is crucial. Not only this, accurate streamflow forecasts has utmost importance while dealing with natural disasters, e.g. floods and droughts. For flood warning, small temporal scale i.e. hourly or daily is preferable while a relatively large scale forecasts are important for reservoir operation and irrigation scheduling. Reservoir operation further decides the amount of water release which is either surplus or required by downstream users.

In short, an efficient future planning and comprehensive management of water resources that are linked with the streams, is impossible without a reliable and accurate estimate of streamflow. The next section explains the existing irrigation system of Pakistan, its history along with the current practices of water sharing arrangements of Indus basin stream flow.

1.3 Indus Basin Irrigation System

Indus Basin Irrigation System (IBIS) is one of the oldest irrigation systems of the world and the farmers are using this largest integrated system since times immemorial. However the technical adaption in irrigation methods has brought a new revolution and the concept of irrigation has changed from open wells to mechanically operated tube wells and flood canals to perennial canals. Although, besides Indus Basin there exists two other basins namely Kharan Closed Basin and Makran Coastal Basin in Pakistan but there resources and application is very limited [12].

While discussing history of IBIS, there come two important milestones; Indus Water treaty 1960 and Indus Water Accord 1991. After Indo-Pak partition, the Indus water treaty was signed in 1960 between India and Pakistan under umbrella of World Bank to resolve the water issues of both the countries. The outcome of this treaty resulted in giving a full control of western rivers Indus, Jhelum and Chenab to Pakistan while an exclusive authority of eastern rivers Ravi, Sutlej and Beas to India. As a result of this treaty a system of link canals has been devised and a total of 12 link canals were constructed to feed the canals which were earlier on fed by the eastern rivers.

Before water apportionment accord 1991, the water was distributed between all provinces of Pakistan on basis of pre-partition arrangements (1947-1970) and decided by Federal Ministry of water and power Pakistan primarily based upon historic uses, after 1970. The accord not only balanced river supplies including flood supplies and future impounding but also came out with a plan to cope with the future shortages, problems in water sharing between the provinces and challenges in agricultural sector of Pakistan.

To execute this plan and implement the water accord, Indus River System Authority (IRSA) has been formulated which represents all the four provinces. Before start of the each crop season, IRSA prepares a water availability forecast and determines the share of each province in accordance with the water apportionment

accord. All provinces then prepare a thorough canal operation program depending upon the share as determined by IRSA for their province.

Historic water rights of the main canals are protected and adhoc water sharing arrangement was based on the 10-daily water allocations; Excess and shortages to be shared proportionate to the design allocations; and Historic surface water rights should not apply to groundwater. It also assured the escape of certain amount of water to the sea to lay down the procedure for sharing shortages and surpluses. The actual average system uses for the period 1977-82 could be used as a guideline for future regulation pattern [13]. According to [14] the Indus Basin Water Management Framework is based upon timely supplies, equitable distribution and water allocations for cultural land and non-perennial canals.

Being the crucial source of water for Pakistan, the Indus basin system is often called as the bread basket for Pakistan. To efficiently manage this precious source of water, accurate stream flow management in Indus basin is crucial. The management efficiency could be enhanced by capturing the challenges that exist in the complex terrain of Indus basin. These include the contrasting regimes, limited meteorological network and the dependency of the catchments response on multiple climate factors. The research motives to deal with these challenges are explained in the next section.

1.4 Research Motives, Problem Statement and Research Questions

1.4.1 Research Motives

The mountainous catchments, which are snow-fed are highly sensitive to the changes in temperature and precipitation. The response of such catchments is mostly derived by snowmelt, which is a function of respective SCA. Therefore, modeling the response of mountainous catchments often require a comprehensive data-set including information about temperature, precipitation and snow cover,

etc. Due to confined economical resources, Pakistan has a very limited network for on-field measurement of snow fall and snow cover. Not only this, the existing methods are old and may have flaws in the measurement process [15]. The extreme hydro-meteorological conditions in a complex terrain of UIB, make it even more difficult to get on-field measurements of many climate variables.

Previously, the researchers have utilized the concept of fusing/integrating on-ground (time series data) and satellite derived observations (spatial maps) to create better hydrological models but these studies are limited as compared to the overall importance of the subject; [16], [17], [18]. It is notable, that the existing data-fusion practices are more valid in the field of remote sensing and geophysics and somehow less valid in the field of hydrological modeling [19]. Similarly, the limits of fusion are presently not characterized and there are no specific guidelines that what amount of data should be entered into a fusion process and whether or not the use of multi-source data would be advantageous [20]. Therefore, it is essential to experiment the data-fusion approaches in the field of hydrological forecasting to establish the evidence of its importance.

Although, multiple data of different types and sources provide a better picture of catchment and its response. However, it is necessary to select or fuse only that particular information among the set of data, which is more correlated to our desired output. The data scrutiny is crucial in getting the optimal results from data fusion, which is often neglected by the modelers. In this case, advanced feature selection methods provide an opportunity for features/data selection in data driven models. Again these methods are not very common in the field of hydrological modeling.

Typically, feature selection methods work on a principle of selecting inputs among the larger set of inputs, which are creating noise while predicting the given output. The process of input variable selection could be facilitated if the noise present in the data is already known. The measure of noise present among the data, prior to model development, not only reduces computational effort but also eliminates uncertainty in the developed models. This also helps in creating models that are

more generalized to the unseen data with reduced over-fitting problems, which are common in data-driven hydrological models.

The present research work is carried out to reduce the uncertainty in data-driven models by strengthening the weak links present between the hydrological data and their respective models. These include the use of inappropriate data, relying on single type/ source of data and no information of input's noise/variation prior to modeling. The above discussion clearly indicates that there is a need of time to; 1. Improve data-driven modeling through creating a better data-set, 2. Incorporate other sources of data to get a better picture of the hydrological process and 3. Experiment data fusion options to improve the efficiency of hydrological models.

1.4.2 Problem Statement

The extreme complexity of the UIB terrain with a poorly gauged network makes it very difficult to accurately model the response of the catchment. On the other hand, contrasting regimes of the UIB requires multiple climate variables of different types and/or sources to cover the extent of the catchment. The data obtained from changing regimes of the UIB has trend discrepancies, noise and nonlinearity. The use of raw data may create uncertainty in hydrological models, especially in data-driven models, which fully rely upon the input/output data. Although, data preprocessing provides an opportunity for data improvement but investigation of a physical process through mathematical applications could be affected by boundary conditions and may create errors and uncertainties. Most of the previous studies on the UIB were carried out with the lumped hydrology of a portion of this large river basin and the models were developed for relatively high estimation interval. The shorter intervals for streamflow estimation like weekly or daily are more preferable for flood estimation and management in the region. The previous studies on UIB entirely missed the concept of data-fusion. The models covering a larger part of the UIB catchment and containing multiple data-sets from different sources and types is crucial to develop data fusion system for the UIB. This is essential to understand the impact of different climate variables, individually and in-combination on the model efficiency.

1.4.3 Research Questions

1. What is the impact of data preprocessing on the improvement of ANN based hydrological forecasting?
2. What is the effect of multi-source information on the improvement of streamflow forecasting models for complex catchment of UIB?
3. What are the best data fusion options to model the UIB streamflow through ANN?

1.5 Novelty of the Research Work

The research aims at the improvement of hydrological forecasting through focusing all aspects simultaneously including data transformation, data length selection, input selection, data fusion of different types of climate variables, rather focusing on one or two dimensions as most of the previous researchers did. Not only this, the methodology provided to achieve this goal is simple, innovative and different as compared to the past studies carried out on the same topic. It is evident that every catchment behaves differently, even the same catchment may have different behavioral phases. This comprehensive study is unique in a sense that it focuses on the development of data fusion models for the UIB, which was not explored for this type of research before. The main points describing the novelty of the research work are described in the following points:

1. The present work utilized larger part of the UIB for hydrological modeling with multiple input variables as compared to the previous researchers.
2. For the first time, the impact of number of climate variables on the response of the UIB is checked individually and in-combination.
3. A comparative assessment has been made on the performance of discharge estimation by changing the type/ source of input data.

4. The study is first of its kind to apply the Box-Cox transformation on hydrological data to improve performance of ANN based hydrological models by data preprocessing.
5. Conjunctive use of Box-Cox with Gamma test and ANN is also unique in hydrological modeling. The development of data fusion system for UIB streamflow through advanced feature selection techniques is distinctive.
6. It is also expected that the developed models will correct and supersede all the previous data-driven models established for Upper Indus Basin and provide an accurate model-data combination that can be further used with confidence.

1.6 Research Significance and Practical Applications

The study is significant in a sense that it provides a simple and understandable approach to improve data, add data, fuse data and ultimately provide a better data-set, which provides a better picture of hydrological process within a catchment. To get a better representation of the catchment's response, it is necessary to understand the different behavioral phases and regimes of the catchment. The UIB observes snow-melt as a dominant regime among the other regimes, which are glacial-melt and rainfall-fed. Therefore, it is necessary to impart the snow cover change along with the other climate variables to verify its impact on the catchment's response individually as well as in combination. The present study provides an opportunity for the researchers to get a clear picture of the catchment's behavior through changing the type/ source of multiple climate variables. These multiple climate variables are fused together to get a more informed data state for the hydrological modeling of the UIB.

A better data-set could be defined as a set of data that consists of a particular set of input variables among the many candidate input variables and that specific shape of data among the many other possible shapes, which can model the target output

in a best possible way. This is achieved through applying data preprocessing techniques which involve data transformation by a family of power transformation and combination selection through Gamma test. The change in power factor changes the shape of data in Box-cox transformation. The selection of this power factor is simply made through the histogram characteristics and probability plots. Whereas, the selection of input combination is achieved through estimating the variance of noise present in the data-set with the help of a gamma test. Not only this, it provides more clarity in understanding the role of each type of data in the model improvement, particularly ANN based hydrological forecasting models.

The study particularly focus on the improvement of streamflow estimation models in the Upper Indus Basin. However, it provides an initial set goal to the researchers to eliminate uncertainty in their hydrological models through data improvement and data fusion before the model development process. The developed models could be used confidently to predict the weekly stream flows at Tarbela. The predicted streamflows could be used for better reservoir operation, flood management and irrigation scheduling at downstream of Tarblea. It is also expected that the developed models will enhance the water management efficiency as the uncertainty in the models has been reduced by incorporating multi type/source data, data preprocessing, data fusion and calibration of models through advanced ANN based technique.

1.7 Research Objectives

The main objectives of this research work are;

1. Creating a better data-set to train ANN based streamflow forecasting models through Data Preprocessing. The preprocessing of data involves a two-step procedure; data improvement through a mathematical transformation and screening of inputs through Genetic Algorithm and Gamma Test. The objective is to use this improved data-set to develop ANN based streamflow models, which could be used with less uncertainty and more confidence.

2. Creating a better data-set by incorporating the satellite derived Snow Cover Area (SCA) with on-ground discharge observations, to effectively model the response of the mountainous catchment. The objective is to provide an evidence that use of multi-source data provide a better picture of the catchment;s response. Specifically, satellite derived SCA could be used successfully as a possible predictor to capture the response of mountainous catchments like UIB. Consequently, to develop ANN based streamflow models through this integrated data-set, which perform better as compared to traditional rainfall-runoff, or snowmelt-runoff models.
3. Combining different types of climate variables and checking the impact of each type of variable, on the performance of ANN based streamflow forecasting models, individually and in combination. The objective is to improve ANN based streamflow estimation models by adopting data fusion options on multi-type, nature and source of data.

1.8 Scope and Limitations of Research

1.8.1 Scope

The present research focused on developing stream flow estimation models for UIB at Tarbela. Four types of input variables have been considered as inputs to model the response at Tarbela, which include discharge at upland stations, information of precipitation, antecedent data condition of global solar radiation and snow cover area. The data length used to achieve the first objective is ten (11) years (1995-2005). Whereas, for the 2nd and 3rd objective the data length of 8 years have been utilized (2003-2010). The SCA is derived from MODIS satellite imageries for three sub-basins of UIB including Astore, Bunji and Gilgit. The ANN based stream flow estimation models are trained via BFGS algorithm with two fixed hidden layers and varying nodes. The ANN model training and data fusion process is performed in WinGamma environment. Only four (04) types of feature selection techniques have been adopted including full embedding, sequential embedding, hill climbing

and genetic algorithm. The limitation to carry out this research work are defined, which are provided in the following section.

1.8.2 Limitations

Modeling the real world problems always bear some limitations and constraints. Similarly the present research work also contains some limitations, which are listed below:

1. There are issues with the recent data availability and consistency due to confined economical resources and limited meteorological network of the UIB. Therefore, the data-set is carefully selected for the duration for which the data is found consistent for all the gauging stations. Furthermore, the data type (direct / global) entirely depends upon the availability constraints of the relevant department, e.g. the data used for solar radiation is only available as global measurements.
2. There are data transformation options other than power transformation, which could be applied and evaluated in comparison with this transformation. However, this study is limited to only application of the Box-Cox transformation due to the fact that it is not a single transformation, rather a family of power transformation. It is used because of its unique characteristics of simultaneously reducing the non-normality, non-linearity and heteroscedasticity in the data. This limitation is applied on a condition of satisfactory results.
3. The study utilizes the satellite derived snow cover area as one of the input variable for only three sub-basins of the UIB. The selection of these sub-basins depend upon the previous researches that demonstrated their role in contributing the flow derived by snowmelt.
4. The main focus of the present study is to improve the ANN based model performance through playing with the “data” only. Therefore, all the developed ANN models consider two fixed hidden layers (with varying nodes).

Also, the ANN models are trained only via BFSG algorithm, which is also fixed for the training of all models of the present study.

1.9 Brief Methodology

For the current research work, initial selection of climate variables has been performed on the basis of data availability by the relevant department, data consistency for the specific duration, sensitivity analysis performed by the previous researchers and already used variables to model the response of UIB. The flow dependency of UIB on snow-melt is captured by the temporal change in snow cover area of three main sub-catchments of the entire UIB along with the global solar radiation that plays an important role in defining the snow cover dynamics. Due to the scarce conditions of snow data in the complex terrain of UIB, the satellite derived snow cover area is utilized. In addition to this the other two key variables including rainfall and runoff are utilized, which are traditionally used in most of the catchment's flow estimation models. The antecedent data condition of these variables is collected for a number of stations located within the UIB, which constitute a comprehensive set of 25 inputs.

To improve the ANN modeling through data preprocessing which targets the 1st objective of the research work, a two-step procedure containing data transformation through BoxCox transformation and data screening through Gamma test has been adopted. To achieve the 2nd objective, satellite derived SCA has been utilized in combination of the on ground discharges of three sub-basins of UIB to develop ANN based stream flow estimation models at Besham Qila. To achieve the 3rd objective of this research work, the impact of multiple type/ source of input variables, which are described in the above paragraph, has been checked on the response of UIB at Tarbela. For this purpose, the two considerations of data fusion have been adopted, which include the data fusion on the basis of type/ source of data and data fusion through advanced feature selection methods. The feature selection methods utilized in this study are; Full Embedding, Sequential Embedding, Hill Climbing and Genetic Algorithm.

1.10 Thesis Organization

This chapter starts with the information about the water resources management practices in Pakistan, followed by the Indus Basin Irrigation System and the impact of climate change in the UIB. After that, the research background and motives are discussed in detail. At the end, research objectives are outlined.

The 2nd Chapter is dedicated for the literature review carried out for this research work. The literature review includes the data preprocessing options and practices for ANN, role of snow cover in mountainous catchment and previous studies which have established the UIB flow dependency over snow-melt. The literature review further includes the importance of SCA in many climate studies, the importance of hydrological data in data-driven models, the importance and the practice of data fusion options by the researchers in hydrological forecasting. The chapter is concluded by the problem statement with particular motives to carry out the research work.

The 3rd chapter presents the methodology adopted to carry out the research work. The whole methodology is divided in three (03) main sections. The first section is dedicated for the methods to achieve the 1st objective of the research work, which includes development of improved ANN models through data-preprocessing. This section starts with the study area & datasets followed by the Box-Cox transformation. Further, the input selection procedure through Gamma test is described in detail, which is also adopted in two other main sections targeting other objectives of the research work. The model development procedure adopted for ANN based streamflow estimation models is described in detail. The section also contains the detail of performance indicators which have been used to evaluate the performance of models throughout this research work. The second main section of the methodology covers the 2nd part of the research work that includes the development of ANN based streamflow models by incorporating the satellite derived Snow Cover Area (SCA) as one of the input variable along with the on-ground flow observations. The methodology includes the study area and data-set, MODIS snow product for SCA, input combination & data length selection and ANN model development.

The third main section of this chapter is dedicated for the 3rd part of the research work that uses the multi-type/nature of climate variables in different combinations to check the impact of each input combination for the UIB streamflow estimation models. The impact of data fusion is checked in the performance evaluation of ANN based streamflow estimation models. The methodology includes the study area & dataset, data fusion options and ANN model development.

The 4th Chapter presents all the results obtained for a variety of ANN models developed on different datasets/ conditions, as described in the chapter 3. The chapter covers the results obtained through the Box-Cox transformation, Gamma Test, M-Test and ANN Modeling. The chapter also includes a comprehensive discussion on all types of results obtained through different tests and models. The chapter also contains summaries of results at the end of each main section.

The 5th chapter is dedicated for overall conclusions of the research work. The chapter is concluded by the future possibilities and recommendations.

Chapter 2

Literature Review

2.1 Background

The importance of accurate stream flow estimation to increase the water management efficiency along with the significance of the UIB for the water management sector of Pakistan has been explained in Chapter 1. The challenges for the UIB has been identified, which include; climate change and nonlinearity, catchments complexity, varying regimes, limited network and the dependency of response on multiple climate variables. This chapter provides a detailed literature review confirming the above mentioned challenges along with the previous efforts, to model the response of the UIB, as well as the catchments with similar complexities and problems.

The chapter provides a detailed literature review about the options available to capture hydrological process in general and ANN based hydrological models in particular. The motives of the research work utilizing multi type/ source data to get a better input data set are explained with the help of literature review as the UIB flow dependency on multiple climate variables and data fusion. Similarly, the novelty of the research work is highlighted with the help of previous literature by identifying the research pockets in different aspects of the literature, as explained in following headings, supporting the objectives of this research work. This is

achieved by providing a short summary at the end of each section with overall summary at the end of this chapter.

2.2 Indus Basin: Climate Change Pattern and Trend Analysis

The Indus River originates from the Tibetan Plateau located in China, runs through Kashmir and enters the Pakistan from Gilgit Baltistan. Passing through the northern areas of Pakistan, it descends from the mountains after Tarbela reservoir, runs through the entire country and discharged into the Arabian Sea. In jurisdiction of Pakistan, the Indus Basin is divided into two main parts; the Upper Indus Basin (UIB) and the Lower Indus Basin (LIB). The Indus Basin up to Tarbela reservoir is termed as the UIB, after which the basin is termed as the LIB.

The runoff originating from the HindukushKarakoramHimalaya (HKH) ranges, is mainly generated through melting of snow and glaciers [21], [22] and contributes up to 80% of mean annual flows of Upper Indus Basin (UIB) [23], [24]. The mountainous region with high altitude and low temperature, receives most part of its precipitation as snow. Therefore, the most part of the UIB remained covered by snow in maximum time of the year and snow cover may reach up to 90% [25].

Due to the complexity of the catchment, the previous studies on the basin have varying observations. Parsad [26] and Liu and Chan [27] reported that global warming instigated glacial recession and created significant changes on hydrology and water resources over HKH region. According to them, the glacier cover over HKH ranges is considered as one of the fastest retreating cover in the world. According to Wang [28], glacier mass of the region is shrinking, resulting in increased melt-water contribution to river flows downstream of the area, specifically during summer season. The reduction in snow covered areas within the UIB has also been confirmed by [29] who reported a decline of about 2.15% during period 1992

to 2010. On the other hand, [30] claimed that snow cover in UIB has slightly increased in south (Western Himalaya) and in North (Central Karakorum). Archer and Fowler [31] have also reported the same trend of increased ice mass over UIB for the last two decades.

The discrepancies in previous studies have created uncertainties in climate estimation models [24]. It has been confirmed by previous researchers that UIB has contrasting hydrometeo-cryospheric regimes because of extreme complexity of HKH terrain [32]. Being a complex terrain, it has a limited meteorological network that is unable to cover the extent of this basin in both horizontal and vertical directions [22].

2.2.1 Summary

The complex terrain of UIB is poorly gauged. The limited on-ground network is unable to cover the extent of the watershed. The discrepancies in previous studies exist due to the fact that UIB observe contrasting regimes in different parts of the basin. Therefore, it is difficult to capture the trend variation of climate to catchment's response.

2.3 Upper Indus Basin Dependency on Multiple Climate Variables: Studies and Models

Due to its geographical location and continental climatic effects, UIB has been used as a key area for variety of climate related studies [21] and captured interest of many researchers during past, e.g., [33] investigated different parameters of River Jhelum and found strong correlation between snowpack and water storage. De Scally [34] performed sensitivity analysis using climate variables of UIB to stream-flow using MODIS satellite product. Hewitt [35] proved that precipitation over UIB is highly affected by orographic barriers. Not only this, [36], [37] and [38] have also evidenced the dependency of stream-flow of the region on

meteorological and climatic variables. According to [39], the hydrology of UIB is poorly understood because the quantification of water balance is highly variable temporally as well as spatially due to complex terrain of the basin. He estimated the high altitude precipitation through glacier mass balance and found this far beyond than the observed or estimated gridded precipitation. Further [40], used this corrected precipitation data set and developed model that is calibrated using river runoff, snow cover and geodetic glacier mass balance. He concluded that the future climate of UIB is highly uncertain and there is a projected decrease in glacier volume which will ultimately results in decreased river flow. He utilized climate simulations to forecast a change in the region and the main focus was on the analysis of precipitation change signals using General Circulation Model (GCM). Mukhopadhyay [29] developed a distributed model for flow estimation in UIB utilizing both spatial and temporal data.

Most of the studies carried on the UIB are more likely climate assessment studies [21], [22], [25], [30], [40], [41], [34], [35], [36], [39], [42], [43], [44], [45] and a few of them focused on climate modeling, specifically hydrological modeling, e.g. [29], [46], [47], [48], [49] and [50]. However, these studies were basically carried out with the lumped hydrology of a portion of this large river basin, e.g. [48] focused only on Shigar river and its catchment, [49] performed hydrological modeling by focusing only Hunza catchment, [51] and [15] on Gilgit catchment, [46] developed Snowmelt runoff models for two sub catchments of UIB including Hunza and Astore. Moreover, the estimation interval for the models developed for UIB is usually high, e.g. [47] developed seasonal stream flow estimation models, whereas monthly models are developed by [29] and [52] for stream flow estimation in the region. The shorter intervals for stream flow estimation like weekly or daily are more preferable for flood estimation and management in the region. Similarly, the most of the models developed for the UIB are either purely snow-melt runoff models [46] or rainfall runoff models [5] and they entirely missed the concept of data-fusion. The hybrid models covering a larger part of the UIB catchment and containing multiple datasets from different sources and types, is the need of time to efficiently model the stream flow in the region.

Previously, many researchers have carried climate assessment of UIB and determined the variables, which are important to capture the response of this complex catchment e.g. [36] stated that despite of the basins contrasting regimes, the response of the UIB could be forecasted through precipitation measurements, taken at different valley stations located within the catchment. He also concluded that the flow originating in higher altitudes is dependent upon the area of catchment that is covered by snow. Further, [34] performed a sensitivity analysis for the assessment of flow in the UIB and found that the meteorological point observations like precipitation and temperature have the predictive relationship to the flow of the region. Similarly [53] indicated that the temperature plays an important role in the snow-melt models developed for the UIB. Also, confirmed by [30] that the UIB regimes are susceptible to change due to rapidly changing precipitation and temperature patterns. Further, many researchers pointed that snow cover dynamics play an important role in defining the hydrological regimes of the area [22] [46]. The snow and glacier melt are the major contributors to the flow generated in the UIB [49], [41]. Charles [47] mentioned the snow-melt as the dominant source of flow in the UIB and glacier melt as the second.

The snow-melt is accelerated when the more solar radiations are absorbed at the surface. Bilal [25] carried a snow cover variation analysis on UIB using different set of climate variables such as temperature, precipitation, relative humidity and solar radiation. He observed that SCA of UIB has the inverse relationship to the intensity of solar radiation. Similarly, [54] indicated that increasing values of solar radiations can shrink the snow cover area and accelerate the melting process. The higher value of solar radiation cause a rapid snow melt in the eastern part of the basin [29]. Remesan [55] pointed out the importance of solar radiation in hydrological modeling. [56] used solar radiation as an input variable for flow estimation in upper Ticino River Basin. [57] mentioned that the solar radiation is a crucial variable in estimating snow melt estimation and neglecting it may lead to higher errors. Further, [58] demonstrated that for UIB, significant correlation exist between the solar radiation, temperature and precipitation measurements. Previously, many researchers have utilized solar radiation in hydrological modeling

e.g. Singh used solar radiation as one of the input variable for hydrological modeling of the UIB [49]; [59] for discharge estimation at Waiokura catchment, New Zealand; [60] for mountainous catchment in China; [61] for stream flow forecast at catchment located in east Australia and [57] for snow glacier melt estimation in Andean glaciers, Bolivia.

The limited on-ground network present in the UIB is unable to cover the extent and complexity of the basin [22]. The dependency of flow on contrasting regimes of the basin urges researchers to use another sources of data to develop a better understanding of the catchments behavior. Literature indicated that remotely sensed observations such as snow cover area and land surface temperature obtained through MODIS provide significant analogues for on-ground observations [41].

Previous studies on the UIB evidenced that a predictive relationship exists between satellite observations and on ground observations of the basin. As [41] reported that remotely sensed spatial data products (MODIS SCA and LST) can provide adequate analogues for these point observations. He also suggested that the fusion of these two types of data may improve the assessment of the hydrological impact of the UIB. Similarly, [46] has shown the UIB flow dependency upon satellite-derived SCA and simulated the runoff using Snowmelt-Runoff Model (SRM). The importance of SCA derived through satellite images is increased exponentially when a snow fed catchment is complex and poorly gauged.

2.3.1 Importance of Satellite Derived Snow Cover Area (SCA)

The mountainous catchments located at higher altitudes, receive their most part of the precipitation as snow. Therefore, in such catchments, snow cover dynamics plays an important role as change in snow cover area directly relates to the response, which is in the form of flow in the streams. With the constraints of topographical complexities and confined economical resources, it is often practically impossible to physically measure the snow cover changes in such catchments. However, in catchments where most part of the flow is derived by the snow-melt,

the snow cover dynamics plays a crucial role in estimating the response of the catchment and could not be neglected.

Snow cover area (SCA) is considered as an important factor for many climate change and water management challenges [62]. It plays a vital role in estimation of stream flows for mountainous areas where the flow is mostly generated through melting of glacial masses [63]. The balance of these masses define the contribution of snowmelt to runoff [64]. The magnitude of snowmelt could be obtained through calculation of changing SCA of a typical region.

Remote sensing offers a wide range of options through a set of satellites to check the spatial and temporal variation of snow cover extent. Most of these satellite observations which are essentially the gridded data-sets, are easy to use and freely available, e.g. [40], [65], [30]. Although, these gridded datasets have the capacity of observing multiple parameters at the same time with more Ariel coverage but in some cases the use of this data alone, may create erroneous results. For example, in the complex terrain like UIB, where the grids are often larger than the spatial variability of precipitation and the adopted interpolation schemes may add up and lead to uncertain outcomes. Similarly, [66] observed that the satellite observations underestimate the precipitation in areas where significant snowfall occurs. So, the use of gridded datasets alone, for hydrological estimation, questionable.

To overcome this problem, remotely sensed data could be used in addition to on-ground observations for hydrological forecasting, which is easy to use and freely available in most of the cases. The change in snow cover area could be assessed through satellite images, specifically downloaded for this purpose. The fusion of these two types of data provides a better picture and it is expected that more information about the catchment results in better hydrological forecasting models.

The main drive of utilizing snow cover area is to improve the real time stream flow forecasting for a catchment which is complex and has a limited meteorological network by combining the two sources of data (on-ground discharge data & satellite derived SCA). It is expected that the fusion of this multi-source data could create a better initial state with less uncertainty, which ultimately results in better stream

flow forecasts for a difficult topographical catchment such as UIB. The fusion of these input variables has been carried out with the help of Gamma Test that provides an initial estimate of the Mean Square Error (MSE), prior to modeling for each set of input combination/mask. Moreover, this study has also explored the use of ANN for stream-flow estimation in a mountainous catchment, where greater part of the flow is derived through melting of snow.

2.3.2 Summary

The existing models developed for Pakistani part of the UIB has mainly focused the lumped portion of the large basin with limited inputs and missing data constraints. The trend studies on the climate of UIB exist but with a little focus on hydrological modeling. The flow of the UIB is mainly derived by snow-melt but also observe variation to multiple climate variables. SCA is an important factor, which should be considered as one of the main input variable for the catchments, which observe dominating snow-melt regimes. Satellite derived SCA could be used for the estimation of stream flow in the UIB as the previous studies found strong correlation between the remotely sensed SCA and catchment's response in the region.

2.4 Non-Linearity, Noise in Hydrological Data and Need for Data Preprocessing

The real time hydrological data may contain noise, missing information and deviation from its original scale due to complex and nonlinear nature of hydrological processes. The data when used as it is in hydrological forecasting may create uncertainty in hydrological models, especially in data-driven models which fully rely upon the input-output data.

The data based models for hydrological estimations are becoming popular day by day owing to increase in data availability and increased computational ability

with the development in computer techniques and applications [6]. However, the accuracy of these hydrological models entirely depends upon the quality of the data which are essentially the hydrometric observations. The uncertainty in these observations may lead to the uncertainty in the models themselves [67]. Similarly, the uncertainty may occur with limited availability of stream flow data due to inadequate observational network as compared to the extent of the watershed [68], [69]. Which is the case in most of the developing countries where watersheds are ungauged or poorly gauged [70].

It is evident that the trust in hydrological models could only be achieved with the surety that the hydrometric observations are correct and verified. Because these observations serve as the foundation for any empirical or statistical model and are extensively used in the process of calibration. Therefore, the hydrometric observations should be as accurate as possible as their accuracy will be ultimately reflected in the performance of hydrological models. Besides the quality of data, the appropriate selection of climate variables is equally important and plays a crucial role in developing efficient climate estimation models.

2.4.1 Artificial Neural Networks (ANN)

Advancement in computational approaches and methodologies directed researchers to adopt innovative methods to model the real world problems. However, these sophisticated methods always bear limitations, boundary conditions and sometimes uncertainties. The same is true in case of hydrological modeling, where modeling options often bear uncertainties due to the complex nature of hydrological processes. These processes are usually described by empirical laws and rely on catchments characteristics, climate variables and hydrological observations. The data-driven hydrological models entirely depend upon observations/data with a limited or no knowledge of underlying physical mechanism. The nonlinearity involved in hydrological data may require non-linear data-driven models to capture the complex relationships between input output. These models usually rely upon methods of computational intelligence and machine learning.

During recent past, the use of Artificial Neural Networking (ANN) techniques in hydrologic forecasting remains the focus of many researchers [71] and its capacity of performing well in hydrologic modeling has also been accepted by ASCE Task committee [72], [73]. Recently, Artificial Intelligence (AI) based data-driven modeling have become quite popular among hydrologists due to their ability of dealing nonlinear hydrologic data, especially in rainfall-runoff modeling; [74] and [75]. Although, Artificial Neural Networks (ANNs) are self-adaptive in nature and considered capable to deal with non-linearity of the data [73], however, their performance could be affected by the quality of input data [76]. The presence of noise in data may hinder the performance of ANN models because an inappropriate input data may lead to an inappropriate learning map [77]. Owing to the nonlinear nature of hydrological process, the resulting hydrological observations may contain undulations, missing information, skewness and large deviations from its original scale. Therefore, the use of original time series data for hydrological forecasting may affect the prediction accuracy of data-driven models [50].

2.4.2 Data Pre-processing Options and Practices for ANN

Artificial Neural Networks (ANNs) are capable of capturing the complex and non-linear relationships between inputs and outputs and they do not require detailed knowledge about catchment and underlying physical processes [74], [78]. However, an inappropriate input data to ANN models may create inappropriate learning maps, that ultimately results in reduced efficiency of hydrological models [77]. Similarly, the accuracy of these models depends upon a careful selection of elements that are user-defined such as model structure, data length selection, parameter optimization and data normalization techniques etc. [79].

Apart from the selection of suitable network-type and its architecture, a suitable method to reduce over-fitting is required for the successful application of ANN models [80]. The conventional methods to reduce over fitting in ANN are; regularization [81], early stopping [82] reducing complexity [83] and noise injection [84].

Preprocessing of data plays an important role in development of ANN models, especially when applied in a real time hydrological forecasting [85]. Data preprocessing for an ANN type model may include transformation of data by applying different mathematical functions which are essentially called “data transformations”. These transformations provide ease of description, vibrant understanding and enables to perform further operation that is more acceptable and useful for data driven modeling.

Data preprocessing could also be used to accelerate the learning rate of an ANN model through eliminating the irrelevant data [86]. Preprocessing of data provides high accuracy with less computational effort in improving the training capability of ANN models [87]. Other researcher have also reported that the performance of ANN models could significantly be increased using transformed data-set instead of using original data as input to ANN models [88], [89], [90], [91], [92], [93], [94]. In addition to data transformation, the optimal selection of inputs also plays a crucial role in the accuracy of ANN based stream flow forecasting models. Afan [95] reported a significant improvement in the efficiency of ANN based stream flow forecasting models with the process of input selection through Genetic Algorithm (GA). Similarly, [96] highlighted the importance of selection of climate variables for stream flow forecasting in Upper Senegal River.

Previously, the types of preprocessing techniques that have been applied to improve ANN models in water resources are Single Spectrum Analysis (SSA); [97] Wavelet Analysis (WA); [98], [99], [3], [100] and Empirical Mode Decomposition (EMD); [101], [102]. Investigation of a physical process through these mathematical transformations may cause some errors that must be taken in account [103]. Both, WA and EMD transformations are greatly affected by the boundary effects and may result in poor modeling quality and lower prediction performance in stream flow estimation models [104] and [3]. Therefore, selecting a right transformation is quite critical in development of hydrological models as uncertainty in the input data may result in uncertainty in the models themselves [67]. Although, no constraint of normality is considered in ANN type models still their performance could significantly be increased using simple normalization techniques [105].

Most commonly used transformations in data-driven modeling are log, inverse and square root transformations [106]. Dirk [107] used square root, cube root and logarithmic transformation for rainfall data analysis. Log normalization was used to standardized the input data for ANN based stream flow estimation models [5] and for reservoir level estimation models [108]. Hassan [109] applied the Box-Cox transformation in order to make data more convenient for the development of ANN based sediment load estimation models.

The practice of trying different types of transformations on input data and checking their impact individually, on the model improvement is not practicable. The hydrological data is complex and may have varying trends. So, it is quite possible that one type of transformation is suitable for a set of hydrologic data and the same transformation is highly unsuitable for another set of hydrologic data. Same is the case with the modeling options as a specific model could provide better results when the input data to this model is transformed by a specific type of transformation.

In this case, the Box-Cox transformation as provided by [110] provides an opportunity for researchers to find the optimal normalizing transformation that fits to their data as it simultaneously corrects non-normality, nonlinearity and heteroscedasticity in the data. The Box-Cox transformation is not a single transformation rather it is a family of power transformation that transforms the data through a power factor, Lambda (λ). While applying the Box-Cox transformation, selecting the suitable value of λ that makes the data more normal is quite crucial. As changing values of λ , may change the shape of the data and ultimately may affect the modeling performance of a data-driven model.

2.4.3 Summary

The presence of noise and non-normality in hydrological data hinders the performance accuracy of models despite of their defined ability to capture non-linearity. The appropriate pre-processing procedure is required to improve models by improving data. During recent past, Artificial Neural Networks (ANN) have been

successfully used in real time hydrological forecasting as they have advantage over traditional forecasting models with no constraint of data normality and nonlinearity. However, the studies indicate that the training capability of ANN could be improved through data preprocessing options. The preprocessing of raw data provides an opportunity to improve the data quality through reducing noise, creating equal spread and eliminating those inputs which are creating hindrance in the development of smooth modeling process.

The motive of applying data preprocessing is to improve the accuracy of ANN based stream flow estimation models through; 1. Scaling of data to proportionate with the transfer function in output layer and normalization of data to create better learning maps; and 2. Screening of input variables to eliminate the inputs that are creating hindrance in the process of developing smooth models.

2.5 Data Fusion: Need and Practice in Hydrological Forecasting

Advancement in computer applications, innovations in data sampling and improved methods of modeling has not only provided facilitation to hydrologic models but also created complexities and sophistications [20]. There are number of linear and non-linear modeling options available to deal with the complexity of catchment characteristics including process based physical models and data based black box models [2]. The process based models reflect the underlying physical process and essentially are knowledge driven models, while the data driven models deals and play with “data only” without considering the detailed underlying physical process [5].

The choice of modeling option for hydrological estimation depends upon availability of data observations, physical information and utilization of purpose [59]. Development of models for hydrological estimation, which is not only accurate but also reliable, is an important issue in water management communities [70]. Previously, the researchers have used various types of flow estimation models, ranging

from lumped empirical to physical distributed models, with diverse mathematical equations and relationships [111], [112]. Both the data-driven and physical distributed models have their advantages and disadvantages while being used in hydrological estimation. Although, the data driven models often possess higher performance efficiency but their applicability is valid only within the limits of the boundary of the data.

Similarly, the physical models are better at representing the spatial variability of hydrological parameters but they require extensive amount of data for physical interpretation and rigorous computational effort [113]. Besides these two options, conceptual lumped models provide another opportunity in the field of hydrological modeling with an advantage of simple structure and low computational cost. These models are constructed by inter-linked conceptual elements, in which each element represents a specific hydrological component [114].

Comparison of models also provide an opportunity to quantify the uncertainty present in the model, through multi-model development process [115], [116] and [117]. Previously, many researchers have used this comparison to analyze the performance of different types of models developed for stream flow estimation [118], [119], [120], [121] [122], [123], [124], [125], [60] and [126]. However, a very few of them have focused on the mountainous catchments which are not only complex but also observe data scarce conditions. The UIB has the similar background and receive limited attention in terms of data selection, data preprocessing and data fusion in hydrological modeling.

In developing countries, most of the watersheds are ungauged or poorly gauged due to confined economical resources [70]. Thus, creating a problem of limited availability of the data for hydrological models, especially in the case of data-driven models. As, the stream flow data is highly nonlinear in both temporal and spatial aspects, therefore, the uncertainty in such data may create uncertainty in the hydrological models themselves [67].

The uncertainty in data-driven hydrological models could be reduced using multiple source/type information instead of using single source/type information. Azmi

[127] reported that the hydrological models could only be improved through “observations” and not by increasing the complexity of the models. This is the reason that despite of choosing advanced modeling options to model the response of complex mountainous catchments, it is difficult to eliminate the “uncertainty” because the available data is limited due to little coverage over the terrain [128]. Dijk [129] suggested to use the new breed of observations, in combination with the on-ground observations, containing remotely sensed data, airborne data and in-situ sensor’s data to improve the performance of hydrological models.

Dasarathy [130] reported that the use of limited or single source data may hinder the performance of hydrological models. Most of the studies on multi-source information fusion assumed that use of more information results in better outcomes [131]. According to [20], data fusion is defined as the integration of data from different natures and different sources but it encourages the amalgamation of original as well as processed information to produce an output that is more useful and acceptable.

Normally, data fusion is considered as the process of combining/fusing information, which are essentially measurements or features collected from a different set of sensors to create a comprehensive data-set or a picture. This integrated data-set or picture could be used for the estimation of parameters and/or for problem classification purpose. This type of datafusion is common in the applications of remote sensing and geophysics [19].

While dealing with real time stream flow estimation, it is difficult to obtain hydrological data from a set of variable sensors or sources as spatially dense in-situ networks are limited due to confined resources, especially in developing countries. Similarly, most of the existing data fusion methods are either able to work with information from a same mode/family (image at a particular time) or a temporal information at a specific spatial point or object [132], [133] and [134]. Although some of them are able to fuse spatial data with point observations [135], but still they are unable to integrate spatial data with distributed temporal data, efficiently; which is an essential requirement to create a better initial state in many hydrological models.

Therefore, data-fusion approaches should be re-invented or re-explored to make its applications more practical in the field hydrological estimation. According to [129], it is not only of a scientific interest but a bare necessity to combine new observations like insitu sensors data, airborne data and remotely sensed data with the traditional on ground observations to create hydrological models that behave better. The same is reported by [127] that operational hydrological forecasting may be improved through the use of multiple source information as compared to single source information.

Earlier, it has been reported by [136] that the on ground hydrological observations provide a real information about the hydrological process at a point and these observations could be used to calibrate the model inputs. However, their spatial and temporal distribution could be affected by the limited data availability [137]. [138] reported that if the rain gauges are installed uniformly over the catchment area despite of their limited numbers, they still can provide reasonable accuracy for the measured observations, however in the remote areas the distribution of gauges is irregular and the data collected through these stations doesnt represent the true picture of hydrological process. In addition to the on-ground observations, the satellite derived observations are now being used to enhance the efficiency of hydrological models developed on complex terrains, where the ground observations are limited due to poor site accessibility [32]. The change in snow cover area for such in-accessible complex terrains could be determined using MODIS products, which provides the high resolution imageries with accuracy over 90% [139], [140] and [141].

The previous work on hydrological estimation through data-fusion approach has focused on many aspects to improve hydrological forecasting through imparting multiple source information (data integration/selection), model-data coupling and model-model coupling, e.g. [17] proposed an algorithm to fuse the radar and gauge precipitation time-series to capture the better response of the watershed.

The results showed an increase in $R^2 = 0.74$ for models developed with fused data set as compared to the models utilizing only gauge observations ($R^2 = 0.54$). Data fusion approach is used to improve river-flow forecasting, by combining the

multiple sources data, with NSE values more than 95% [20]. [142] proposed a cluster fusion framework to model the important features of reservoir inflow. [112] performed model-data fusion through coupling of Xinanjiang Rainfall-runoff (RR) model with optimization algorithms and data assimilation techniques, to simultaneously improve the uncertainty in stream flow data and the model structure. The models developed with data assimilation techniques showed better $R^2 > 90\%$ as compared to the models developed without data assimilation ($R^2 = 86\%$). [143] improved discharge estimation models through fusing remotely sensed SCA with MOHYSE hydrological model. The results indicated an increase of NSE from 72% to 85% and decrease of RMSE up to 22% for the models developed with integrated data set. Similarly, a multi-sensor satellite data fusion methodology is adopted by [18] to produce ET maps over Choptank River watershed with an improvement in errors up to 27%. [144] proposed coupling of global climate models with hydrological models to improve stream flow forecasting that exhibited correlation coefficient values more than 90%.

Although the data fusion within ANN is not new and it has been successfully utilized in many fields [145], [146], [147], [148], [149] but its application in hydrological forecasting received a limited attention as compared to its overall significance. Previously, [20] evaluated a variety of data fusion strategies in hydrological modeling and concluded that the data fusion by ANN outperformed. Shu and Burn [150] combined individual ANN models to enhance the estimation of pooled flood frequency. Azmi et al. [127] performed model data fusion by incorporating ANN to improve hydrological forecasting. However, the previous studies are inadequate to exploit the full benefits of data-fusion despite of its overall significance.

ANN have become quite popular in runoff modeling, specifically in areas where catchment response is mostly dependent upon rainfall. There are many studies which have utilized ANN to develop hydrologic models on rainfall dominant catchments [151] and [152]. But a very few have been reported for mountainous catchments due to the fact that measuring on ground precipitation and other climate variables is quite challenging for such catchments [75]. Although there are many past studies on the application of ANN for snow estimation such as [57],

[153], [154], [155] and snowmelt- runoff modeling [156], but a very few have been reported which utilized SCA as one of the input variables for ANN based flow estimation models [156], [157], [75]. Whereas the satellite-derived SCA in addition to gauge-observations could be used to create a better input data-set and ultimately better stream flow estimation models.

The utilization of ANN based models for stream flow estimation in complex catchments, has received limited attention as compared to their overall potential. The dependency of such catchments on multiple climate factors often make it difficult to select the appropriate set of input variables for stream flow estimation in the region. For this, a careful selection of input variables is crucial to make the data-set appropriate for model development. One of the important climate variable that has a direct impact on most of the complex terrains located in high altitudes, is snow. The same is the case with the UIB, where flow is the main derivative of snow melt and should be incorporated in development of stream flow estimation models in the region.

2.5.1 Summary

Multiple type/source information is required to capture the complexity of a catchment which observe contrasting regimes. The application of fusing multiple type/source data in hydrology is limited, despite of the overall significance of the data fusion approaches.

2.6 Overall Summary

The effectiveness of data-driven hydrological models depends upon many others factors besides the data itself, like model architecture, calibration and model configuration. Still the importance of “data” comes at first, because an inappropriate hydrological data may alter the performance of carefully selected advanced model. Therefore, the first attempt to improve hydrological models should include the use of appropriate data that describe the hydrological processes more accurately

with less noise and uncertainty involved in it, rather improving the complexity of the models, itself. But the quality of hydrological data obtained through already installed gauges is itself questionable as most of the watersheds in the developing countries are either ungauged or poorly gauged due to confined financial resources. The mountainous catchments suffer more because they bear both, the lack of coverage and the complexity of the terrain. These catchments often observe contrasting regimes and the limited available meteorological network creates uncertainty not only in the hydrological data but also in hydrological models. On the other hand, the complex catchments with varying regimes require multi-type and multi-source data to effectively model their response. The response is generated in the form of tributaries that carry streamflows from different part of the basin and drain into the main stream.

Although, many previous researchers have evidenced the importance of data improvement in hydrological forecasting. However their use is neglected as compared to the overall importance of the subject. The uncertainty in hydrological estimation could be significantly reduced by providing a better data input-state to the hydrological models, because the presence of uncertainty in hydrological data may affect the accuracy of hydrological models [158], [159], [160].

A better input state may include; improving hydrological data through data transformation, finding best input combinations, appropriate data length selection and use of advanced data fusion options etc. Similarly, the use multiple input variables observed through different sources could be used to get a better picture of the hydrological phenomenon, especially in catchments which are complex, observe contrasting regimes, and has dependency upon multiple climate factors. The same is the case with Pakistani part of the UIB, which being a main source of water for Pakistan, has many challenges to combat water shortage issues and maximize water control efficiency in the water management sector of Pakistan.

Likewise, the limits of fusion are presently not characterized and there is no clear guidance that what amount of data should be entered into a fusion process and whether or not the use of multi-source data would be advantageous, 2. The errors in data may cause problems and it is quite possible that the use of erroneous data

in fusion process produce even less accurate results than using non-erroneous data from a single source.

Therefore, it is expected that the use of multi-type data would be beneficial for complex watersheds that observe contrasting regimes as the response of such catchments could be a function of multiple parameters such as snow melt, snow depth, temperature, rainfall, solar radiation, humidity, etc. Similarly, it is anticipated that the use of multi-source data provides a better picture for such catchments which are not only complex but also bear a problem of limited availability of on-ground observations.

The similar conditions apply on the Upper Indus Basin (UIB) that faces both, a contrasting hydro-meteo-cryospheric regime and the limited meteorological network [32]. The motivation of the current research is to reduce the uncertainty by creating a better initial state for hydrological forecasting models through using multi-type (hydrological, meteorological and cryospheric) and multi-source (on ground and satellite derived) data.

To make the integrated/fused data set that can accurately represent the catchments picture, the choice of input variables that affect the catchments response of the UIB is crucial. For the current study, the selection of variables among the many candidate input variables are made with the help of literature review. The variables are chosen on the basis of their role in affecting the response of UIB with the help of various forecasting models developed and evaluated by previous researchers. The current study utilizes four (4) types of variables (on ground observations) including discharge (Q), precipitation (P), global solar radiation (SR) and satellite derived snow cover area (SCA). Out of which the first three variables (Q, P and SR) have been used for data transformation, whereas all the four (Q, P, SR and SCA) have been utilized for data-fusion purpose. The data observations from multiple stations and snow cover area of 3 catchments help to make a comprehensive data-set, which sums up to a total 25 inputs.

Chapter 3

Study Area & Methodology

3.1 Background

The research work is divided into three (03) main parts and methodology is adopted to achieve the research objectives as defined in Chapter 1. The first part of the methodology targets the 1st objective (Section), the second part (Section) targets the 2nd and the third part (Section) targets the 3rd and last objective of this research work.

The methodology of the first part of the research work provides the methods to improve ANN based stream flow estimation through data preprocessing. The need of data preprocessing is discussed in detail with the help of literature review in Chapter 2. The second part targets the 2nd objective through adding satellite derived SCA as an additional variable to model the response of the UIB. The previous chapter showed the dependency of UIB stream flow on changing SCA and the importance of satellite derived SCA for complex catchments. Whereas, the third and final part includes a more comprehensive dataset with data fusion option to develop a data fusion system for the UIB stream flow measurement. The importance of data fusion for catchments observing contrasting regimes is highlighted in both Chapters 1 and 2. The current methodology differs the previous practices as already defined in the novelty of this research work.

The study area is same for all three parts, which is the Pakistani part of the Upper Indus Basin. Similarly, some of the procedures as the Gamma Test, M-test and ANN model development are consistently used throughout this research work. However the ANN model architecture may vary because of the different data-sets, used for all the three parts according to the focus and requirement of the studies. The general flow chart of the research work is presented in the Fig. 3.1.

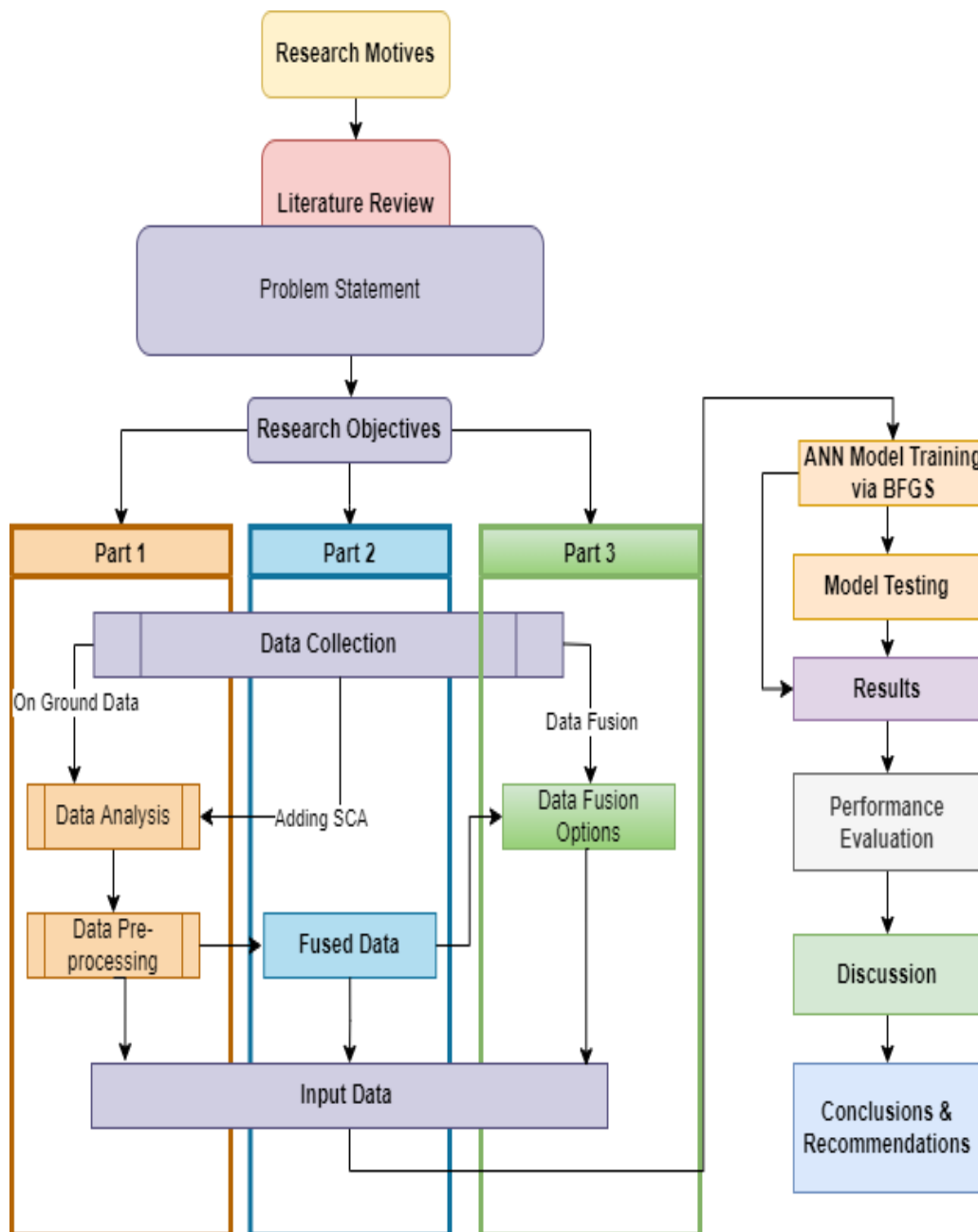


FIGURE 3.1: General Methodology Flow Chart

3.2 Study Area

The catchment of the Indus River begins from the mountains of Tibet. It originates from the mountains of the Hindukash and the Himalaya and flows on southwesterly course largely through Pakistan but also through India, Afghanistan and China [40]. The total length of the Indus is 3160 km and it covers an entire area of 900,930 km^2 , out of which 528,156 km^2 lie in Pakistan [161]. Major portion of this mountainous catchment is covered with snow and glaciers, which contributes more than 80% of the flow for the Indus Basin Irrigation System (IBIS) [22].

The Upper Indus Basin (UIB) covers a huge geological area ($> 200,000 km^2$) and it observes large altitudinal variations with a mean elevation of 4000 masl (meters above sea level) [46]. The UIB terrain consists of high elevated mountains with extreme roughness and complexity. With several peaks having elevation more than 6000m, the basin is a widespread belt of ridges and high valleys. The main source of water resources is derived by the Karokaram range and covers the major part of the basin with huge glaciers and snow peaks [44].

Before the Indus reaches the plains, it is impounded behind Pakistan's largest dam at Tarbela, termed as 'rim station' where water is stored, measured, and diverted into an extensive network of canals in the province of Punjab. The Indus basin area, starting from its origin, up to Tarbela dam is known as the Upper Indus Basin (UIB) which is used as an area of interest for this part of study as shown in Fig. 3.2.

Climate change has a significant impact on water resources worldwide and may result in change in flow magnitude, variability and frequency of extreme events [162]. It becomes even more significant for the hydrology of mountainous catchments, which are located in high altitudes and are normally covered with snow and glaciers [163].

The upper part of the Indus Basin is also a mountainous catchment and this is why the response generated through this catchment is very sensitive to the changes occur in different climate factors.

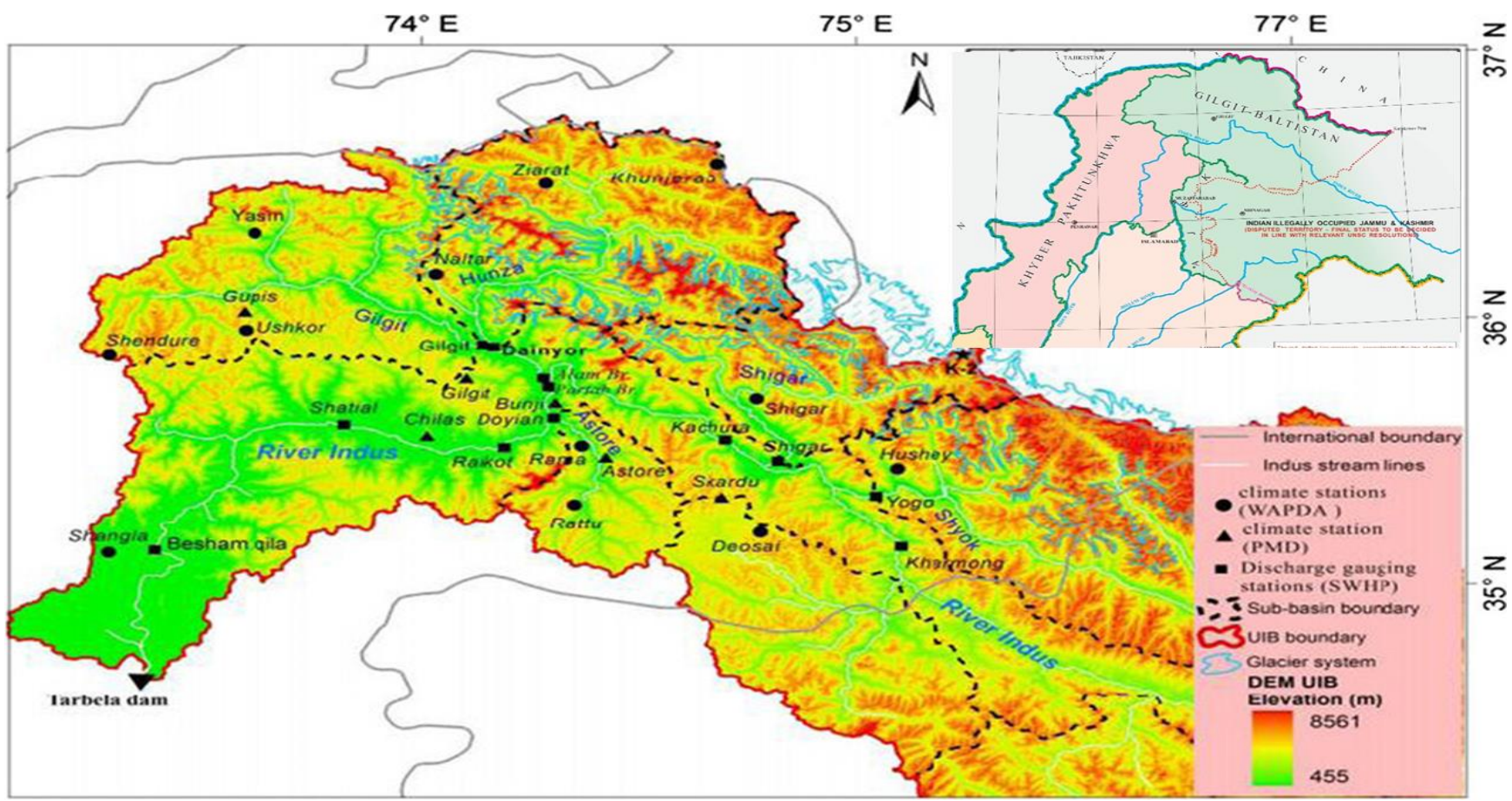


FIGURE 3.2: Upper Indus Basin: Gauging Stations (Source: WAPDA)

Climate changes directly affect the shape of flood hydrograph and could create a risk of food security of the people of the Hindu Kush, the Karakoram and the Himalaya (HKH) mountain ranges [48], as the economy of this region is highly dependent upon water availability [164]. Impact of climate change on the Indus basin includes rise in temperature that ultimately resulted in decreased river flow [11].

The evidence of this change is studied by [165] who reported the depletion of glacier volume over the mountainous ranges of Hindukush and Himalaya. Fowler and Archer [21] reported a rise in summer temperature that is the basic cause for rapid glacier melt in the region. Similarly, [166] indicated a change in temperature in pre-monsoon and monsoon period for this region. It was also noticed by [31] that there is a significant change in precipitation in the area of Upper Indus Basin (UIB) in both summer and winter seasons during period of 1961 to 1999.

Due to its geographical location and continental climatic effects, UIB has been used as a key area for variety of climate related studies [21] and captured interest of many researchers during past, e.g., [33] investigated different parameters of River Jhelum and found strong correlation between snowpack and water storage. A weather generator had been developed by [58] for the estimation of different environmental parameters in the periphery of The UIB. Forsythe [34] performed sensitivity analysis using climate variables of UIB to streamflow using MODIS satellite product. Hewitt [35] proved that precipitation over Upper Indus basin is highly affected by orographic barriers. Not only this, [36], [37] and [34] have also evidenced the dependency of stream-flow of the region on meteorological and climatic variables.

According to [39], the hydrology of UIB is poorly understood because the quantification of water balance is highly variable temporally as well as spatially due to complex terrain of the basin. He estimated the high altitude precipitation through glacier mass balance and found this far beyond than the observed or estimated gridded precipitation. Further [40] used this corrected precipitation data set and developed model that is calibrated using river runoff, snow cover and geodetic

glacier mass balance. He concluded that the future climate of UIB is highly uncertain and there is a projected decrease in glacier volume which will ultimately result in decreased river flow.

The above studies were basically carried out with the lumped hydrology of a portion of this large river basin. For better understanding of different behavioral phases of the catchment, it is necessary to gather as much information about the catchment, as possible, because the catchments with varying regimes observe complex hydrology and often require multi-type and/or multi-source data to effectively model their response. The hindrances in the development of accurate hydrological models in the UIB include; politically sensitive trans-boundary area, complex terrain and poorly gauged catchment. The uncertainty in hydrological models for such catchments is often generated due to limited catchment information, inappropriate variable selection and improper model calibration.

For the second part of the study, which considered addition of satellite derived SCA in the data-set, three sub-basins of UIB located in the northern part of Pakistan, namely Gilgit, Astore and Bunji. Both, Astore and Gilgit rivers are tributaries of the Indus River while Bunji basin is directly draining into the River Indus from north to west and into River Astore from the eastern side. Astore River originates from Burzil Pass, runs through Astore valley, drains the Deosai Plateau and joins River Gilgit at 35°32'N, 74°42'E. The boundary of Gilgit basin is defined by the location of Alam Bridge Gilgit (74°18'E; 35°55'N), where a stream gauge is installed by Water and Power Development Authority (WAPDA).

The Gilgit River joins the River Indus at Juglot near Bunji where three (03) mountain ranges, Hindu Kush, Karakorum and Himalaya (HKH) meet. The total catchment areas for Astore and Gilgit are 4040 km^2 and 12095 km^2 , respectively. The peaks in Astore catchment are even higher than 8000 m.a.s.l. and remain covered with snow especially during winter season. The upper part of Gilgit basin also has persistent snow cover [167]. The upper part of Gilgit basin also has persistent snow cover [51]. The snowmelt response of these three basins is observed at Bisham Qilla (72°52'E, 34°55'N) located upstream of Tarbela Reservoir, at 580 m.a.s.l. The study area is presented in Fig 3.3.

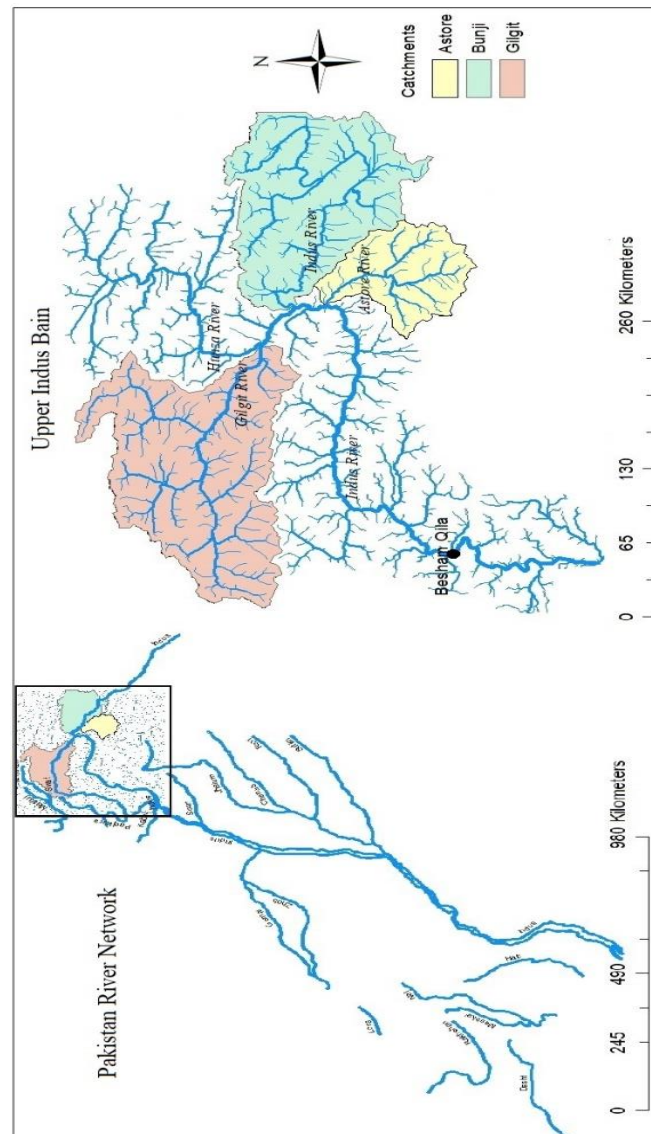


FIGURE 3.3: Study Area (UIB): Catchments delineation using Digital Elevation Model (DEM)

3.3 Improving ANN Based Hydrological Forecasting through Data Preprocessing

The methodology of this part of research work starts with the collection of catchment information of the UIB, continues with the analysis of data and formation of data set, followed by a two-step preprocessing procedure and ANN based model development, which is presented in Fig. 3.4.

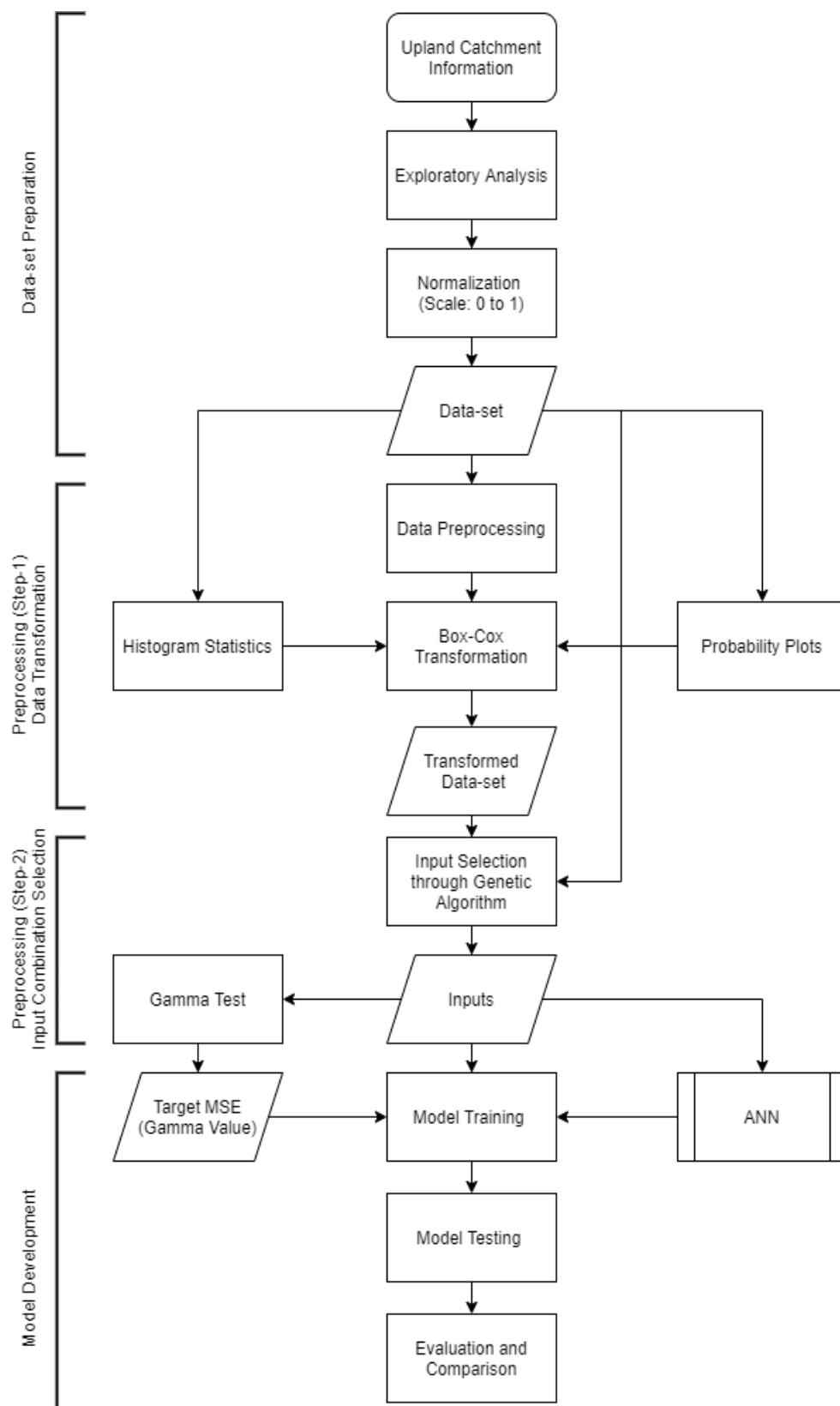


FIGURE 3.4: Methodology Flow Chart

3.3.1 Data-set

The antecedent data condition for the study is obtained from Surface Water Hydrology Project (SWHP), Water and Power Development Authority (WAPDA). The available data-set consists of weekly records of Precipitation (P), Global Solar Radiation (SR) and Discharge (Q) at a total of thirteen (13) stations located within UIB, as presented in Table 3.1. The data length spans over 575 weeks (1995-2005). The Table 3.1 lists station name, its elevation (masl i.e. meters above sea level), and the variables for which the data is available. Out of total twenty one (21) variables, twenty (20) variables are considered as inputs and one (1) variable as output, which is discharge (Q) at Tarbela.

TABLE 3.1: Details of Gauging Stations and respective observations

Sr. No.	River	Station	Elevation (masl)	Variables	No of Variables
1	Indus	Deosai	4142	P*, SR*	2
2	Indus	Rama	3300	P, SR	2
3	Indus	Hushey	2850	P, SR	2
4	Indus	Rattu	2745	P, SR	2
5	Astore	Astore	2546	P	1
6	Kachura Lake	Kachura	2341	P, Q*	2
7	Indus	Sakardu	2228	P	1
8	Gilgit	Gilgit	1430	P	1
9	Indus	Bunji	1403	P, Q	2
10	Indus	Chilas	1265	P	1

Sr. No.	River	Station	Elevation (masl)	Variables	No of Variables
11	Indus	Shatial	1040	Q	1
12	Indus	Basham Qila	580	P, Q	2
13	Indus	Tarbela	450	P, Q	2
Total			13	3	21

* P = Total Weekly Precipitation, SR = Average Global Solar Radiation,

Q = Average Weekly Discharge

The raw data is comprised of daily records of Precipitation in mm, Global Solar Radiation in Watt/m^2 and Discharge in m^3/s . The weekly average data of Solar Radiation and Discharge is achieved by simply taking average of a daily data in a week, whereas the weekly precipitation data is obtained through adding the daily precipitation in a week. The conversion of daily data into weekly data is performed to make the each entity of data is unique and independent. The mean values of the selected variables for the duration 1995 to 2005 are presented in Fig. 3.5.

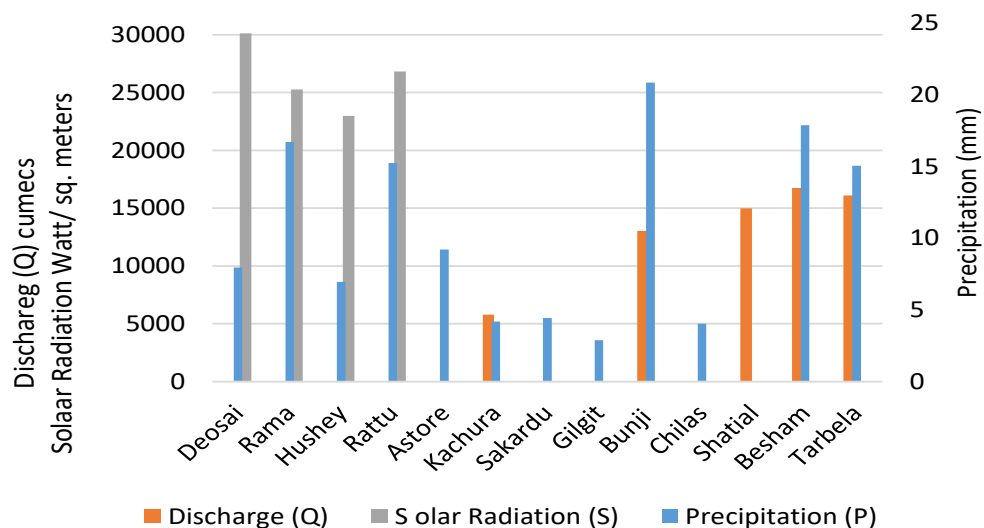


FIGURE 3.5: Mean values of hydrological variables of different stations located in the UIB

The variation in the mean values in the variables for different stations is a depiction, that the UIB observes contrasting hydro-meteorological regimes due to its spread over an elevation range of 200 masl to 8500 masl [78].

In order to ensure the quality of hydrological data, the gauging stations and the data duration is carefully selected for which the available data is consistent and homogeneous. There are no missing values in the data. The descriptive statistics of original hydrological data is presented in the next section while making a comparison with the Box-Cox transformed data-set.

3.3.2 Data Transformation

Data transformation is applying a mathematical function to change the format, structure and/or values of data. Data transformation is applied on the data to make it more efficient and easy to use for data representation and information extraction. In case of data-driven analysis and modeling, data transformation techniques are applied to change the data in a form which is more acceptable to model training process and produce desirable results or outputs. Most of the traditional estimation techniques require the data to follow normal distribution. Although the assumption of normality does not apply necessarily on some of the advance data driven techniques including Artificial Intelligence based methods. However, the past studies indicate (Literature Review/ Chapter 2) that the efficiency of these models could be increased significantly when the data is normally distributed as it facilitate the better learning maps in neural networks. Moreover, the scaling of data make the input data to proportionate with the transfer function in output layer. Hence, enabling the researchers to train models with desirable efficacy.

The selection of data transformation that makes the data to follow the desired distribution could be made on the basis of hit and trail. However, checking the impact of each transformation on the model development is not feasible. Therefore, in order to make it more practical one could achieve the desired shape/format of the data by targeting the specific characteristics of the given distribution. For

example, the requirement for a data to follow normal distribution could be defined by histogram characteristics and probability plots. The same approach is adopted to select the appropriate data transformation to transform the data in a form that it follows normal distribution. For, this purpose a wide spectrum of power transformation is adopted that is often called as the Box-Cox transformation. It must be noted that this transformation is not a single transformation, rather it contains many other know transformation in its umbrella, e.g. Square Root Transformation, Cube Root Transformation, Inverse Transformation and even the Logarithmic Transformation is the special type of Box-Cox Transformation [168].

3.3.2.1 The Box-Cox Transformation

Tukey [169] was the first researcher with the idea that a variety of power transformations could be thought of as a class of same mathematical function. In mathematical form, the idea was indexed as presented by Eq. 3.1 to transform the variable y into y^λ . The formula was improved by Box-Cox in 1964 to take into account the discontinuity at $\lambda = 0$ and later on got the name as Box-Cox Transformation, presented by Eq. 3.2.

$$Y_i^{(\lambda)} = \begin{cases} Y_i^{(\lambda)}; & \lambda \neq 0 \\ \log Y_i; & \lambda = 0 \end{cases} \quad Y_i > 0, \quad (3.1)$$

$$Y_i^{(\lambda)} = \begin{cases} (Y_i^{(\lambda)} - 1)/\lambda; & \lambda \neq 0 \\ \log Y_i; & \lambda = 0 \end{cases} \quad Y_i > 0, \quad (3.2)$$

Where, $Y_i = i_{th}$ value of the data () which is to be transformed, and $\lambda =$ Power Factor.

For unknown power factor (λ), Box-Cox suggested an expression presented by Eq. 3.3

$$Y^{(\lambda)} = (Y_1^{(\lambda)}, Y_2^{(\lambda)}, \dots, Y_n^{(\lambda)}) = A\theta + \epsilon, \quad (3.3)$$

Where, \mathbf{A} represents the matrix of known constants, θ and ϵ both represent vectors associated with the transformed values, of “unknown parameters” and “random errors”, respectively. Although, λ can take infinite number of values but the optimum value could be selected theoretically that minimizes the error and transforms the variables towards normality [106]. This value is often selected on the basis of so called confidence interval that contains the value of power factor, which returns the data towards normality.

For this, one approach is to determine the upper and lower confidence levels for a specified confidence value (often used as 95%) and check if the power factor “1” lies within this range. If so, the transformation is not necessary and if not, the transformation is necessary to make the data normal. The other option is that we can select a range of power values between -5 to 5, as provided by the Box-Cox to select the best value, which provides the best approximation of the normality. As the formula presented in Eq. 3.2 is valid only for positive value of variables, that’s why many researchers including [170], [171] and [172] came up with minor variations to take account of negative values and to be applied for special conditions.

In this study, the confidence interval came out to be -0.037 to 0.425, which indicate that the transformation is required to make the data normal because the power factor = 1 does not lie in this range. However besides this range, the other values of power factor (λ) have been tried to include other known transformations like square, cube root, inverse and logarithmic to make a comparison of results achieved in form of histogram characteristics and probability plots. For this case, the optimum value ($\lambda= 0.005$) lies with the confidence interval and selected on the basis of histogram characteristics and relative standard deviation. The results of which are shown in the results & discussion portion. Before applying the power factor, the whole data is scaled between 0 and 1.

As there are no negative values in the available data-set, so the scaling is simply achieved by dividing each entity in a data column with the maximum value in that column. The transformed data-set with $\lambda= 0.005$, is further used for the next step of preprocessing, which is the selection of inputs through Gamma Test.

3.3.3 Input Selection through Gamma Test (GT)

3.3.3.1 Gamma Test

In order to develop smooth and reliable models for a given set of inputs, it is often necessary to find the unique combination of inputs that provide minimum noise and variance while predicting the desired output [173]. The idea of Gamma test was first reported by [174], which states that the variation of noise on an output could be estimated directly instead of using trial and error procedure. This estimate is called the Gamma Statistics, which is essentially considered as the best value of MSE (Mean Square Error) for the development of a smooth model [175]. The Gamma statistics or gamma value is defined as the noise estimated from the training dataset which is being used for model development. A smooth model is often considered as the model that only captures the systematic behavior and neglects other aspects, which are created by the noise present in the data. WinGamma (A nonlinear modeling tool/software) assumes those models smooth in which outputs can be determined smoothly from the inputs with only limiting factor of noise in the data [176]. The ability of a model to capture only systematic behavior and consequently performs well for the unseen data (testing data) is called, generalization [177].

The generalization capability of a model could be affected by various factors that may include, insufficient training data, irrelevant input variables and improper parameter adjustments, etc. Early stopping is a very common method which is used to facilitate good generalization by dividing the whole data set into three subsets: training, validation and testing data sets. The model trained via training data set is periodically checked against validation data set. The training process is continued if the errors from the both sets reduce, otherwise the training is stopped to avoid the so called over-fitting problems, which are very common in data-driven models.

The other methods to improve the generalization ability of models may include parameter/ weight adjustments, regularization and noise injection methods. Model regularization works by altering the function through adding an extra penalty to

the error function. This enhances the generalization ability of models by controlling the coefficient to take extreme values. In noise injection method, a random noise is injected to the training data-set to enhance the learning capability of models. All these technique used to improve the generalization of models, work without any prior knowledge of noise present in the data. However, GT provides us an opportunity to enhance the generalization capability of models with the known information of noise present in the data, even before the model development process. The use of GT is very advantageous as:

1. Gamma value provides the measure of statistical noise present in the training dataset.
2. The expected model performance is known prior to model development.
3. The estimated noise (Gamma Value) can be used as a stopping criteria for model training.
4. The significance of input variables can be evaluated.
5. The need for a separate validation data set is eliminated.

Therefore, in addition to achieve the best input combination, the Gamma value could also be used as a stopping criteria to reduce over fitting problems in ANN modeling. As compared to the conventional approaches used for the generalization, the Gamma test is superior in the context that the noise present in the data is already known and could be used to access the model performance, prior to the model building [178]. Thus reducing the need of separate validation data-set which is usually required for conventional early stopping methods. Elshorbagy [179] used Gamma Test as an assistive tool to select the appropriate modeling option for hydrological predictions among the various non-linear modeling techniques.

3.3.3.2 Working Principle of Gamma Test

The Gamma Test (GT) provides us the value of Gamma Statistics or best Mean Square Error (MSE) which is the measure of variance of noise on our desired

output [173]. The Gamma test works on an argument that if X and Y are two parameters and Y is a function of X , then this function can be divided into two parts; “smooth” and a “noisy”. By considering the mean of this noise as zero, a constant bias could be added to this function. Although, the function between X and Y is unknown, Gamma test enables to estimate this error prior to the model development process.

On the basis of an initial data set $(x_i, y_i), 1 \leq i \leq M$, an algorithm can be developed between two variables x and y to capture a relationship between them. The process involves decomposition of these variables into smooth and noisy parts while having an assumption that y is a function of x . If f is a smooth function between x and y and r is the part of noise which cannot be considered for by any even model, then their relationship can be shown by Eq. 3.4.

$$y = f(x) + r, \quad (3.4)$$

If the mean of this random variable “ r ” is zero then a constant bias can be engaged into this unknown function f . Despite of the fact that f is unknown, this tool enables us to calculate the value of noise on an output on a certain condition which states that “As the number of data samples increase, the gamma value becomes equal to an asymptotic value which represents the variance of a noise on an output” [174].

This could be achieved by developing a regression line ($Y = A\delta + \tau$) between gamma function of outputs ($Y_m(u)$) and delta function of inputs ($\delta_m(u)$) as given in Eq. (3.5) and Eq. (3.6). The vertical intercept of this regression line is the measure of Gamma statistics (τ), whereas, A is the slope that defines the model complexity; higher the value of A , higher will be the complexity involve in the model. For input data set $X = x_1, x_2, \dots, x_n$, the k th nearest neighbor of an input vector x_i is $x_{m[i,k]}$ and its associated vector output will be $\gamma_{m[i,k]}$.

$$\delta_m(u) = \frac{1}{N} \sum_{i=1}^N |x_{m[i,k]} - x_i|^2 \quad (3.5)$$

$$\gamma_m(u) = \frac{1}{2N} \sum_{i=1}^N |Y_{m[i,k]} - Y_i|^2 \quad (3.6)$$

In order to give stability to this asymptotic value, the gamma statistics should be determined for increasing number of data points (M). This process, performed through an algorithmic program to check the convergence of an infinite number of functions, is called the M-test. This test helps in the selection of the most suitable data length for model training which ultimately helps in providing the best goodness of fit in the ANN models [5], [180] and [181].

It is also noted that not in all cases, gamma test provides best results as it follows an assumption that, non-smoothness in data is only due to the presence of statistical noise in the data. Whereas, this is not true when the outcome which is being predicted is of a probabilistic nature. Therefore, a scale invariant noise $\sigma^2(y)$ is used to standardize the gamma statistics as a deciding factor that how well an output could be modelled by a smooth function. This is called V_{ratio} (Eq. 3.7) and its value normally ranges from 0 to 1. The value closer to 0 means the gamma test could be used as a prediction tool for the best MSE while the value closer to 1 means that there is a low predictability of a given output.

$$V_{ratio} = \frac{\tau}{\sigma^2(y)}, \quad (3.7)$$

For this research-work gamma test has been used in conjunction with a model identification technique in a Win-Gamma environment. Model Identification (MI) is an option in the software that provides a computational facility to make the combination making procedure easy and fast, with the help of advanced algorithms. In this case, Genetic Algorithm (GA) has been utilized as a model identification tool that uses the Darwins idea of natural evolution. For this purpose, the values for population size, mutation rate, crossover rate, gradient fitness, intercept fitness and length fitness have been used as; 100 for original and 10 for transformed data, 0.01, 0.5, 0.1, 0.8, and 0.1, respectively. The input combinations with a minimum gamma values, obtained through Gamma-Test are further used for the development of BFGS neural network models.

3.3.4 Artificial Neural Networking (ANN)

The idea of ANN is inspired by the biological neuron system which has millions of neurons that are interconnected with each other. These neurons carry signals to our brain that acts as a processing unit and gives feedback on upcoming signals. ANN has a similar network of interconnected nodes that work in a same fashion. Each node acts as an artificial neuron and carries some input signals [181]. An arrow denotes a connection from the output of one node to the input of the other. These connecting links receive the input signals and multiply them with the corresponding weights. Signals are then transferred, based upon the type of transfer function opted for a particular type of ANN model.

A typical ANN model must have a minimum of three layers, including an input layer, intermediate layer and an output layer. Intermediate layer which is also called as a hidden layer could be one or more depending upon the structure of an ANN model [55]. A conceptual framework of ANN with two hidden layers is shown in Fig 3.6.

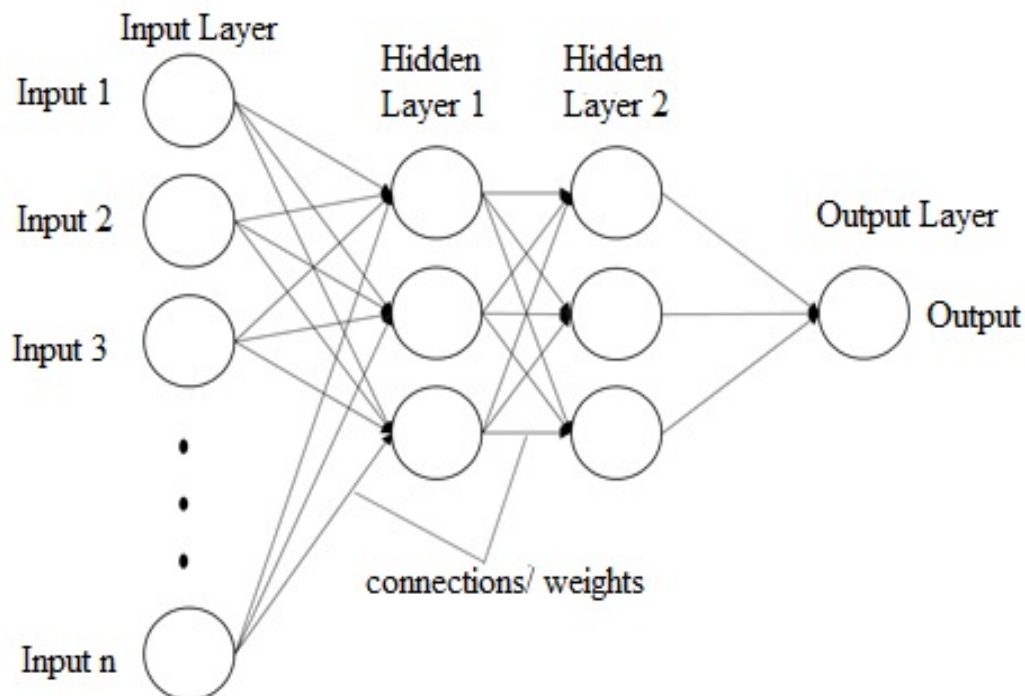


FIGURE 3.6: Framework for ANN with 2 Hidden Layers

ANN type models could be trained through two types of learning approaches which are supervised and unsupervised learnings. The unsupervised learning, also known as self-guided approach, works in the absence of output and is used to find the buried patterns in the data. On the other hand, supervised learning approach is used when a specific output/ target is present. Both inputs and outputs are fed into the network and weights are adjusted to reduce the difference between inputs and outputs.

Selection of the training algorithm is very important while developing an ANN type model because each training algorithm has its own specific properties. Back propagation algorithms are quite popular but are slow as their stable learning requires small learning rates. However, Levenberg-Marquardt (LM), conjugate gradient (CG) and quasi Newton have pretty fast processing as they use the standard optimization techniques [180].

3.3.4.1 BFGS Algorithm

The choice of training algorithm in ANN plays an important role in model training process as different algorithm works on different principles. The Back Propagation (BP) algorithms are quite popular and are commonly used for training of feed forward neural networks. However they are quite slow as they require small learning rates for stable learning. On the other hand, Conjugate Gradient (CG) and BFGS are fast in processing as they use standard optimization techniques.

BFGS is named against four persons; Broyden, Fletcher, Goldfarb and Shanno, who discovered it in 1970 and considered to be the most effective Quasi-Newton Method. Quasi Newton methods are used for the root finding algorithm in k variables. This method is developed for solving equation $f(x) = 0$ at only first iteration as compared to conventionally used Jacobian Matrix comprising difficult and expensive applications through multiple iterations. It is a type of gradient descent function which are used for optimization of nonlinear functions without any constraints. It is superior to other gradient descent functions as it overcomes the limitation of plain gradient descent by seeking a second derivative of the cost

function. In BFGS environment, the descent direction is determined by preconditioning the gradient with the help of known curvature information.

3.3.4.2 Model Training

In this study, a feed forward neural network with two hidden layers trained via Broyden Fletcher Goldfarb Shanno (BFGS) algorithm is used for model development process. Normally, one layer is used to represent the functions which are linearly separable. For complex problems, multi layers are preferred as empirically, deep learning seem to result in better generalization for a wide variety of tasks [182]. As the behavior of natural streams is very complex and the mountainous catchments (like Upper Indus Basin) often observes contrasting regimes. For the better understanding of this complex phenomenon, a MLP neural network is used with 2-hidden layers. The reason behind using two hidden layers in a neural network is due to their ability of solving nonlinear problems as reported by [183] and [184]. Previously the 2-layer BFGS is successfully used to train ANN models in the hydrological estimation e.g. [78], [181], [109] and [5].

The optimum number of nodes in hidden layers of a multi-layer ANN model may vary from problem to problem due to the difference in complexity level, number of inputs and outputs. The less number of nodes may create under-fitting problem while too many neurons may cause over-fitting and takes more time to train a model. It is suggested by [185] that the best practice of finding the optimum number of nodes is to experiment them for the given set of data.

3.3.5 ANN Model Development

After performing the input combination selection on the basis of minimum gamma value (as explained under heading, Input combination selection through Gamma Test), the whole data is divided into two sets; training data-set and testing data-set. The training data-set is employed for model training purpose which is basically adjustment of weights in a typical ANN model. Whereas, the testing data is a set

of unseen data, which is used to evaluate the performance of models trained using training data-set.

Different researchers have used different divisions for training and testing data lengths, such as; [186] used 67% data for training and 37% for testing in ANN based thunderstorm estimation models, [187] used 80-20% combo for ANN based rainfall estimation models, [5] developed streamflow estimation models and used 72.72% data for training and 27.27% data for model testing purpose. In the current research, 70% of the data was used to train models and the rest 30% of the data was employed for model testing purpose. This division of data length was found to be the best as this ratio gives the minimum variance in developed models for both phases.

In this case, feed forward Neural Networks are trained via BFGS and the model structure is finalized in WinGamma environment that uses the default value of five (5) nodes in each (1st and 2nd) layer. The number of nodes is altered on a random basis and the change in coefficient of determination for both the training and testing models is observed, separately for Original and Transformed data-sets as shown in table 3.2. The number of hidden layers for BFGS is fixed as 2, as previous researchers have successfully trained ANN models via 2-layer BFGS algorithm e.g.; [108] and [78]. The number of trials with different value of nodes have been performed e.g. increasing nodes in first layer, decreasing nodes in 2nd) layer and decreasing nodes in both layers.

It should be noted that the best input combinations for original (10110010111111110001) and transformed (10101110100110111011) data-sets are different, despite of the fact that all these combinations are made using same tool (Gamma Test) with same model identification technique (GA). This shows that, both the data-sets behave differently when considered as inputs for the development of ANN models.

It is clear from Table 3.2, that almost all the models which are trained using original data gave significantly high values of R^2 (>90%). However, when these models are tested against unseen data (testing data), they don't perform well with a very less values of R^2 (<70%). On the other hand, the models developed using

transformed data-sets produce stable and satisfactory results in both phases. In this case, the value of R^2 is above 90% in training phase while in testing phase it ranges 78-90% for most of the models.

TABLE 3.2: Defining ANN Model Structure through a set of different node-arrangements for BFGS

Trial No.	No. of Nodes in 1 st hidden layer	No. of Nodes in 2 nd hidden layer	Original Data ($\lambda=0$)			Transformed Data ($\lambda=0.005$)		
			Target MSE	MASK	1011001011111110001*	Target MSE	MASK	10101110100110111011
			MSE	R^2	R^2	MSE	R^2	R^2
			Achieved (Training) (Testing)			Achieved (Training) (Testing)		
			$(\times 10^{-7})$					
1	5	5	0.0012973	97.14	70.20	5.2212	89.5	79.1
2	6	3	0.0012973	97.14	67.79	6.8693	94.2	78.7
3	4	6	0.0013688	96.99	18.19	7.2796	94.3	63.1
4	3	3	0.0021849	95.19	63.15	7.9406	93.2	87.9
5	2	2	0.0053806	88.39	64.18	1.1050	91.2	90.4
6	1	1	0.0039756	91.25	73.10	1.3463	89.5	94.1
7	6	2	0.0012974	97.14	54.79	6.1856	94.7	70.6
8	8	3	0.0012941	97.15	69.24	4.3006	96.2	55.4
9	3	8	0.0015192	96.66	56.19	7.1418	94.9	80.4

*1 means an input is included, 0 means an input is excluded

The sequence of inputs is : P_{Deosia} , SR_{Deosia} , P_{Rama} , SR_{Rama} , P_{Hushey} , SR_{Hushey} , P_{Rattu} , SR_{Rattu} , P_{Astore} , $P_{Kachura}$, $Q_{Kachura}$, P_{Skardu} , P_{Gilgit} , P_{Bunji} , Q_{Bunji} , P_{Chilas} , $Q_{Shatial}$, P_{Basham} , Q_{Basham} , $P_{Tarbela}$

It is also clear from Table 3.2 that closer the value of Achieved MSE to the Targeted MSE, greater will be the value of R^2 in training phase. On the basis of R^2 values, the best selection of node arrangement came out to be 1-1 and 2-2 for original and transformed data-sets. Though, alone R^2 could not be used as a model efficiency parameter in all circumstance, yet in this case it has been considered as an initial criteria for the selection of nodes in hidden layers.

The performance of all the developed models is evaluated on the basis of set of performance indicators as explained the next section.

3.3.6 Performance Indicators

The performance of ANN based streamflow estimation models is assessed using a set of statistical indices as described below:

1. Coefficient of determination (R^2)

It is the measure of variance of an output which is determined by the input/s. mathematically, it is expressed as Eq. (3.8):

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}, \quad (3.8)$$

SS_{res} the sum of squares of residuals between observed and predicted values, whereas SS_{tot} represents the total sum of squares between observed and mean of observed data. The best value of R^2 is when the modelled output exactly matches to the model inputs, i.e. $SS_{res}=0$, and $R^2=1$. However, normally the R^2 value more than 0.70 is considered good as it depicts that the observed and modelled outputs are more than 70% correlated.

2. Nash Sutcliffe Efficiency (NSE)

Similar to (R^2), is also used as a model efficiency parameter and is determined by the mathematical transformation, as presented in Eq. (3.9):

$$NSE = 1 - \frac{\sum_{i=1}^n (X_p^i - X_o^i)}{\sum_{i=1}^n (X_o^i - \overline{X_o})}, \quad (3.9)$$

X_p^i and X_o^i are the i^{th} values of predicted and observed data, whereas $\overline{X_o}$ represents the mean value of the observed data.

3. Root Mean Square Error

It is the square root of the mean square error between predicted and observed data. It is mathematically expressed in Eq. (3.10):

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (X_p^i - X_o^i)^2}{n}}, \quad (3.10)$$

4. BIAS

BIAS is the measure of difference between the mean values of the predicted and observed data. Its formula is expressed in Eq. (3.11):

$$BIAS = \overline{X_p} - \overline{X_o} \quad (3.11)$$

Where, $(\overline{X_p})$ and $(\overline{X_o})$ represent mean values of predicted data and observed data, respectively.

3.4 Improving ANN Based Hydrological Forecasting Through Satellite Derived Snow Cover Area (SCA)

In this section, the performance of ANN models has been optimized through the process of data normalization, input combination selection, selection of nodes in hidden layers and over fitting reduction through Gamma and M-Test. The models are developed and trained using two layer BFGS algorithm in WinGamma environment (see detail of WinGamma application at [178]) for stream-flow estimation at BeshamQila, utilizing SCA of three sub basins of UIB and their respective discharges as model inputs. Afzal [50] and [188] found that these three sub-basins namely, Astore, Gilgit and Bunji are mainly snow fed basins. The comparative assessment of models developed with and without satellite-derived SCA has also been made with the help of performance indicators to show the importance of data-fusion.

3.4.1 Dataset

The input variables for output (discharge at BishamQila) are considered as weekly flow observations (Q) at Astore, Gilgit, Bunji and BishamQila gauging stations and SCA of Astore, Gilgit and Bunji basins. The details of climate dataset used in this study is presented in Table 3.3. Streamflow data is collected from WAPDA,

whereas snow cover areas for these basins have been extracted from MODIS snow cover products in ArcGIS environment. This snow cover product was previously used by researchers for extraction of SCA such as; [189], [163] and [46].

3.4.1.1 MODIS Snow Products for SCA

The MODIS/Terra Snow Cover product with specification, 8-Day L3 Global 500m is used for the measurement of snow cover area over three catchments, as mentioned above in Fig. 3.3. This data product provides the maximum extent of snow cover over 8-day period within $10^{\circ} \times 10^{\circ}$ MODIS sinusoidal grid tiles. These tiles are generated by combining 500 m observations from the data set. A bit flag index is used to track the eight-day snow/no-snow chronology for each 500 m cell. The index defines the probability of clear sky or cloudy sky, and takes the snow reading at least once out of 8 days, when the sky is clear. Thus minimizing the cloud cover and maximize the snow extent, because a cell will only be labeled as the cloud if it is covered by clouds for all 8-day period. This limitation is exceptional and could be neglected.

Effective use of MODIS snow products in climate estimation models is linked to their ability of estimating the snow cover accurately. Therefore, their accuracy should be assessed and validated before utilizing them for climate related studies [25]. However in practice, the MODIS products are validated regularly on a global scale, not only with available on-field measurements but also with other higher resolution products. Many studies have reported the accuracy of MODIS snow products as more than 90% [190], [191], [192], [193] and [194].

In this case, the data-set of MODIS images in HDF format was downloaded from the <https://earthdata.nasa.gov/> for the period of 2003 to 2010. The images are then converted into .tiff format, in order to project the downloaded tiles using WGS 1984 projection system with UTM zone 43N. The respective SCA for each catchment is extracted by masking the respective delineated catchment map of each basin in ArcGIS environment. Due to the fact that all other input variables have the temporal frequency of 7 days (week), the 8-day satellite derived SCA is

converted into weekly data by simply making a graph of obtained values of SCA with time. The weekly SCA data is then extracted from the graph and used for further processing. The variation of snow cover area for the selected catchments is shown in Fig. 3.7.

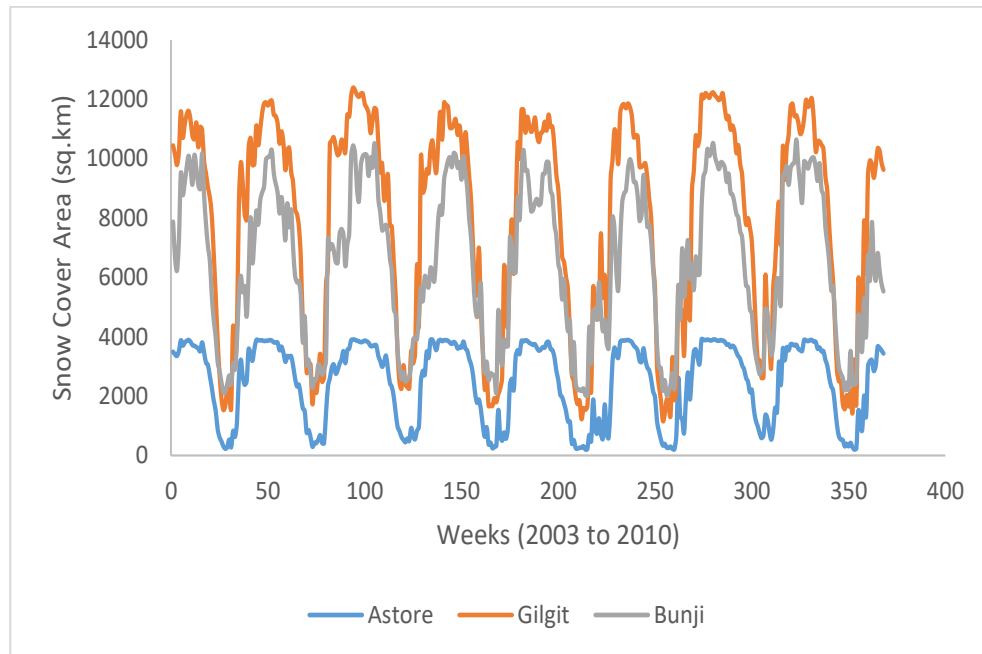


FIGURE 3.7: Time series of SCA for Astore, Gilgit and Bunji Catchments

TABLE 3.3: Data set used for model development

Sr. No.	Station	River	Elevation (masl)	Location	Data sets	Data Length (2003-2010)
1	Astore	Astore	2394	74°42'E, 35°32'N	SCA, Q	368 weeks
2	Gilgit	Gilgit	1460	74°18'E, 35°55'N	SCA, Q	368 weeks
3	Bunji	Indus	1372	74°36'E, 35°42'N	SCA, Q	368 weeks
4	Bisham Qilla	Indus	480	72°52'E, 34°55'N	Q	368 weeks

The whole data is transformed using log normalization (on a scale of 0 to 1) to get a unit-less data set as presented in Eq. 3.12.

$$\text{Normalized } X_i = (X_i - X_{min}) / (X_{max} - X_{min}) \quad (3.12)$$

Where;

X_i is the i^{th} value of X .

X_{max} is the maximum value of variable X .

X_{min} is the minimum value of variable X .

3.4.2 Input Combination and Data Length Selection

In this part of study, the combination selection procedure for input variables is carried out using a mathematical tool, Gamma Test which is already explained in detail in section 3.2.3. GT is applied in WinGamma environment and a model identification tool is used to make number of combinations for a given set of input variables. Model Identification tool provides help in performing GT through a set of algorithms for a given set of inputs. The data length for model training has been optimized using M-test. For this purpose, the gamma value is calculated for increasing number of inputs and the value for which standard error becomes stable against unique data points (inputs), is considered as the data length for model training.

3.4.3 Model Training

In this part of study, the models are trained via two layer BFGS algorithm which is already explained in Section 3.2.4. Different number of nodes in both the hidden layers is tried and value of Mean Square Error (MSE) and Correlation Coefficient (R^2) is computed for every set of nodes in training and testing phases, as shown in

Table 3.4. It is quite clear from the Table 3.4 that there is no significant difference in the values of MSE and R^2 values for different combination of nodes in hidden layers. However, a minute improvement in the R^2 values is noted in testing phase with same number of nodes in both the hidden layers.

The combination (111111) represents the input variables as: SCA_{Astore} , Q_{Astore} , SCA_{Gilgit} , Q_{Gilgit} , SCA_{Bunji} , Q_{Bunji} . “1” means an input is included and “0” means an input is excluded.

TABLE 3.4: Selection of nodes in hidden layers on the basis of MSE and R^2

Network		Combination (111111), with Combination (010101), without									
Architecture		SCA Target MSE 0.000536					SCA Target MSE 0.001118				
Sr. No	Nodes in 1 st Layer	Nodes in 2 nd Layer	MSE (Train)	MSE (Test)	R^2 (Train)	R^2 (Test)	MSE (Train)	MSE (Test)	R^2 (Train)	R^2 (Test)	
			Achieved	Achieved	%	%	Achieved	Achieved	%	%	
1	1	1	0.000612	0.00325	98.9	92.8	0.00104	0.00379	98.1	92.0	
2	1	2	0.000534	0.00341	99.0	92.4	0.00101	0.00418	98.1	90.9	
3	1	5	0.000526	0.00342	99.0	93.5	0.00108	0.0029	98.0	93.5	
4	1	6	0.000533	0.00332	99.0	92.7	0.00108	0.00254	98.9	94.2	
5	1	7	0.000535	0.00342	99.0	92.5	0.00109	0.00382	98.1	91.4	
6	2	1	0.000532	0.00268	99.0	93.8	0.00101	0.00372	98.1	91.9	
7	2	2	0.000531	0.00266	99.0	93.9	0.00099	0.00293	98.2	93.1	
8	3	3	0.000534	0.00292	99.0	93.4	0.00109	0.00369	97.9	92.0	
9	3	5	0.00053	0.00337	99.5	92.4	0.00092	0.00426	98.3	90.2	
10	3	6	0.000525	0.00324	99.0	93.2	0.00108	0.00468	98.0	89.2	
11	4	4	0.000529	0.00267	99.0	93.8	0.00078	0.00349	98.5	92.3	
12	4	7	0.000531	0.00343	99.0	92.2	0.00109	0.00293	98.0	93.4	

Network		Combination (111111), with Combination (010101), without								
Architecture		SCA Target MSE 0.000536				SCA Target MSE 0.001118				
Sr. No	Nodes in 1 st Layer	Nodes in 2 nd Layer	MSE (Train)	MSE (Test)	R ² (Train)	R ² (Test)	MSE (Train)	MSE (Test)	R ² (Train)	R ² (Test)
13	5	1	0.000527	0.00271	99.0	93.7	0.00112	0.00319	97.9	92.7
14	5	3	0.000536	0.00265	99.0	93.8	0.00072	0.00311	98.6	93.0
15	5	5	0.000536	0.00282	99.1	93.5	0.00103	0.00457	98.1	90.1
16	5	7	0.000529	0.00305	99.0	93.2	0.00082	0.00395	98.4	90.8
17	6	1	0.000522	0.00339	99.0	92.3	0.00103	0.00419	98.1	90.8
18	6	3	0.00053	0.00281	99.0	93.6	0.00086	0.00284	98.4	93.4
19	6	6	0.000525	0.00276	99.0	93.9	0.00077	0.00339	98.5	92.2
20	7	1	0.000512	0.00347	99.1	92.2	0.00111	0.00411	97.9	90.3
21	7	2	0.000517	0.00289	99.1	93.3	0.00103	0.00275	98.1	93.3
22	7	7	0.000534	0.00305	99.0	92.9	0.00107	0.00299	98.0	93.5

The models are further evaluated with the help of statistical indices including Nash-Sutcliffe coefficient (NSE), Variance (VAR) and BIAS.

3.5 Improving ANN based Hydrological Forecasting through Data Fusion

3.5.1 Dataset

The data-set is also same as presented in Table 3.1, except the Snow Cover Area (SCA) and discharges of three sub catchments of UIB is added in the data-set (the details and

data collection for SCA of these 3 catchments is presented earlier in Fig 3.3 and Section 3.2.1).

The integrated data-set used in this study is presented in Table 3.5, as follows;

TABLE 3.5: Data set used for data fusion and model development

Sr. No.	River	Station	Elevation (masl)	Variables	No of Variables
1	Indus	Deosai	4142	P, SR	2
2	Indus	Rama	3300	P, SR	2
3	Indus	Hushey	2850	P, SR	2
4	Indus	Rattu	2745	P, SR	2
5	Astore	Astore	2546	P, Q, SCA	3
6	Kachura Lake	Kachura	2341	P, Q*	2
7	Indus	Sakardu	2228	P	1
8	Gilgit	Gilgit	1430	P, Q, SCA	3
9	Indus	Bunji	1403	P, Q, SCA	3
10	Indus	Chilas	1265	P	1
11	Indus	Shatial	1040	Q	1
12	Indus	Basham Qila	580	P, Q	2
13	Indus	Tarbela	450	P, Q	2
Total		13		4	26

* P = Total Weekly Precipitation, SR = Average Global Solar Radiation,

SCA = satellite derived snow cover area (7 days frequency).

Sr. No.	River	Station	Elevation (masl)	Variables	No of Variables
------------	-------	---------	---------------------	-----------	-----------------------

Q = Average weekly discharge.

Q^* = discharge of sub-catchments for which SCA is calculated

The whole data-set has been scaled between 0 and 1 using the formula presented in Eq. 3.12.

3.5.2 Data Fusion Options

Data fusion of antecedent condition of different climate variables including precipitation, solar radiation and discharge collected from the stations located in different part of the UIB, along with the satellite derived SCA of 3 sub basins of UIB, is performed based upon the following conditions:

1. Type / nature and source of data
2. Feature Selection Methods

3.5.2.1 Type / Nature and Source of Data

Hydrological processes are complex and depends upon multiple climate factors. The complexity increases when a catchment area has difficult terrain with varying regimes. In order to capture the response of such catchments, multi-type, multi natured and/or multi source data is required to know more about the different behavioral phases of the catchment area. The variety of information from different sources provide a better picture of catchment, which results in a better correlation to the catchment's response.

The type of data includes the nature of climate variable, like meteorological (e.g. Precipitation and Solar radiation), hydrological (Discharge) and cryospheric (Snow cover Area). The source of data is the source from where the data is collected e.g. on-ground or satellite derived. The source of data may also be considered different if it is collected from different satellites, sensors or product. Similarly, the same variable measured on-ground but from different type of gauges / instruments could be considered as multi-source data.

However, in this chapter the multi-source data means the multi type of data collected from both on-ground gauges and satellite products.

3.5.2.2 Feature Selection Methods

The feature selection methods are the set of techniques that can be utilized for the selection of useful inputs/features from a larger set of inputs [195]. The usefulness of an input or feature can be defined as the maximum relevancy of that input or feature while minimizing the redundancy of other candidate inputs [196]. The inputs which are irrelevant and creating noise in the process of smooth model development should be excluded.

For a given set of inputs, the variance of a noise on an output is determined through a mathematical test, called Gamma Test (GT). The GT enables us to calculate the MSE that is present among the data, prior to model development [197]. The input selection through gamma test is explained previously under sections 3.3.3. With the help of GT, only those features (input variables) are selected that contribute to lower the MSE present between the input variables and output data.

The feature selection methods usually work on the basis of a criterion function that defines how good a particular set of features is? With this, a search criteria is used to decide which set of features is to try next. The search criteria should include all possible input variables in different combinations which are being tested on the basis of already defined criterion function. For a number of inputs n , total possible combinations could be calculated using relation, 2^n-1 . The manual application of this relation on a larger set of inputs requires a huge computational effort and time. For example, a complete search of possible combinations for 26 inputs (as used in this chapter) requires a total of 67108863 tests to be performed. In case of applying GT, for each combination, gamma value will be calculated and checked that if it is minimum enough to consider that combination of inputs for model development purpose.

Therefore, instead of finding the gamma value individually for each combination of inputs (except for input combination made on the basis of type/nature of data), the search process for the input combinations which have minimum gamma value is facilitated with the help of advanced feature selection techniques, which are described below.

3.5.2.3 Full Embedding

Full embedding employs the idea of a selection of inputs selected from all the candidate inputs [78]. To carry out the full search from the n number of inputs, there are $2^n - 1$ unique subsets or embeddings. Each subset has a unique binary code in form of “0” and “1” that represents a mask/combination of input variables [199]. For example, if there are 5 inputs; x_1, x_2, x_3, x_4 and x_5 . A random mask for these inputs 11010 represents that x_1, x_2 and x_4 are included to compute the desired output, whereas x_3 and x_5 are excluded.

Full embedding is a comprehensive research that calculates gamma value for each possible set of input combination. The values obtained for each set of inputs are then arranged in an ascending order. A gamma histogram is then plotted in order to represent the results obtained from the full embedding search, where the range of gamma values are divided into classes (along x-axis) and the frequency of gamma value for each class is plotted on y-axis. The input masks, which create higher gamma values lie in the higher region of gamma value, whereas the masks with low gamma values lie in the lower region of the gamma value. Now the full embedding search could be used to find the suitable mask of inputs through following procedure (Durrant’s Method [178]):

- A. Define **L** matrix with $(mL \times n)$, where mL is the number of input masks that belongs to low region of gamma value, whereas the n defines the total number of inputs. Each row in the matrix represents a unique mask of inputs.
- B. Define **H** matrix with $(mH \times n)$, where mH is the number of input masks that belongs to higher region of gamma value, whereas the n defines the total number of candidate inputs. Each row in the matrix represents a unique mask of inputs.
- C. Count the number of inclusions (1’s) from each column of matrix **L**.
- D. Count the number of exclusions (0’s) from each column of matrix **H**.
- E. With the help of the frequency analysis, we can find the suitable inputs on the basis of following principle;

“Only those inputs are relevant to the models, which are included in majority of the embeddings with small gamma values, whereas those inputs are irrelevant, which are excluded in majority of the embeddings with large gamma values.”

3.5.2.4 Sequential Embedding

Sequential embedding as the name shows, executed sequentially to select inputs among the all candidate inputs. Similar to the full embedding, it provides an optimal solution for embedding dimension, except it follows a pre-defined sequence for inclusion or exclusion of inputs [78].

The procedure is carried out from start to end for a given set \mathbf{X} of inputs $(x_1, x_2, x_3 \dots x_n)$, with a total number of points “ n ”. For each unique mask (or embedding dimension), n ranging from 1 to some predefined maximum value, a set with $n - 1$ dimensional delay vectors could be made.

Mathematically, these vectors can be defined as Eq. 3.13;

$$d_i = x_i, \dots, x_{i+n-1} \quad (3.13)$$

The Gamma test is then utilized to measure how smoothly the d_i could be used to determine the next point of the time series, which is x_{i+n} . This sequence of increasing embedding is carried out until the value of n for which the gamma value is the minimum (closest to 0), is found. The mask of inputs for which, the gamma value is minimum is considered as the optimal or best embedding dimension for model development.

3.5.2.5 Genetic Algorithm

Genetic algorithm (GA) mimics the biological evolution of the species by survival of the fittest, as described by Charles Darwin. The algorithm adjusts a population of individual solutions repeatedly until no further improvement is left.

The heuristic techniques for feature selection requires less time for computation as compared to other optimization (i.e. full embedding search) approaches [95]. Therefore, GA heuristic technique is more useful and rapid, especially when the dimension space is high, as in our case ($n = 26$). Even some consider it is not feasible to find all the subsets through the full embedding search when the $n > 20$, because it can take days to carry out full search for all possible subsets [199]. Although, the heuristic approaches do not provide an optimal solution, still they are good to provide the solution closer to the optimal with reduced computation effort and time.

GA maintains a population “P” of the probable individual solution, which is a ‘mask of inputs’ in this case. A repeated process that mimics the genetic evolution is applied, which modifies the population P. The P could be defined by Eq. ??, as below:

$$P(t) = x_1^t, x_2^t, x_2^t, \dots, x_n^t. \quad (3.14)$$

The initial population at $t=0$ is created on random basis. The individual solutions $x_i^{(t-1)}$ from $P(t-1)$ is selected on the basis of probability. The individual solutions (or masks of inputs) with low gamma value have more probability of being selected for the next generation. The GA alters $P(t)$ on the basis of mutation, that contains a unary and crossover genetic operators [201]. The first one is used to modify an individual solution, whereas the later one is used to create a new solution from the two parent solutions. In order to maintain a constant population size, the solution that does not perform well is rejected.

3.5.2.6 Hill Climbing

Similar to GA, Hill Climbing is also a heuristic approach. However, it is not as sophisticated as GA with comparatively less parameters involved in it. It only contains a mutation operator, whereas the GA has many parameters like population size, cross over probability and mutation probability [201].

Hill climbing starts with a random solution or mask of inputs. Every bit of this mask is flipped to calculate the gamma value for each combination till we reach the end of the mask. The process of flipping continues until no further improvement in gamma value is noted. Gamma test is performed on the total of 15 combinations of inputs which are made either on the basis of type/nature of data or through feature selection methods. The Gamma value for all the developed combinations, along with v_{ratio} is presented in the next chapter. For a given set of input combination, the closer the gamma value to zero the minimum will be the noise of variance on an output.

3.5.3 ANN Model Development

The Gamma value for each set of input combination, as presented in Table 3.6, is selected as a targeted MSE to train ANN based stream flow estimation models. The ANN models

are trained via two layered feed forward BFGS algorithm. The reason behind using two hidden layers in a neural network is due to their ability of solving nonlinear problems as reported by [183] and [184]. The BFGS algorithm is explained in section 3.3.4.1. The number of nodes in hidden layers are selected on random basis as it is suggested by [185] that the best practice of finding the optimum number of nodes is to experiment them for the given set of data. The architecture of ANN models is explained under Table 3.6.

TABLE 3.6: ANN model Architecture

Inputs	15 combinations as presented in Table 4.2								
Nodes in 1 st layer	1	1	1	2	3	3	4	5	5
Nodes in 2 nd Layer	1	3	5	2	1	3	4	1	5
Output	Discharge at Tarbela								
<i>No of Models = 15 × 9 = 135 models</i>									
<i>2 Phases = Training and Testing</i>									
<i>Total Models = 135 training models + 135 testing models = 270 models</i>									

Each combination of input is used to develop ANN based models using 2 hidden layers with a node arrangement as presented in Table 3.6. The training models are developed using training data length optimized through M-Test. The testing models are developed for the rest of the data length to evaluate the performance of models for the unseen data. Both the training and testing models are evaluated on the basis of variety of performance indicators to assess the efficiency of models developed for stream flow estimation at Tarbela. The performance evaluation indicators used to evaluate the performance of developed models are defined and explained in Section 3.3.6.

Chapter 4

Results & Discussion

4.1 Background

Various Stream flow estimation models have been developed targeting the objectives of this research work. The methodology adopted to develop these models is explained in detail in Chapter 3. This chapter contains the detailed results of all the developed models including Box-cox transformation results, Gamma Test results, and ANN model results etc. The procedure of these tests and the detail of performance indicators is explained in the previous chapter. The sections of this chapter describe the associated results with corresponding section of the methodology in Chapter 3, which mainly include the results for ANN models developed through data preprocessing (Section), ANN models developed with satellite derived SCA, and ANN models developed through different data fusion techniques. The results of each section is followed by the detailed discussion and summary.

4.2 ANN Models developed through Data Pre-processing

This section contains the results targeting the first objective of the research work which is carried out to improve ANN based streamflow estimation models through data preprocessing. The section contains results for ANN models developed through original

and transformed datasets, their comparison, a comprehensive discussion followed by the summary.

4.2.1 Data Transformation Results

The power factor (λ) has been selected on the basis of histogram characteristics, covariance and Normal Probability Plots, as shown in Table 4.1 and Fig. 4.1 to 4.5. The graphical representation of data in form of histogram shows center, spread, skewness and outliers present in the data. It also helps to identify the data that from which population distribution it belongs to.

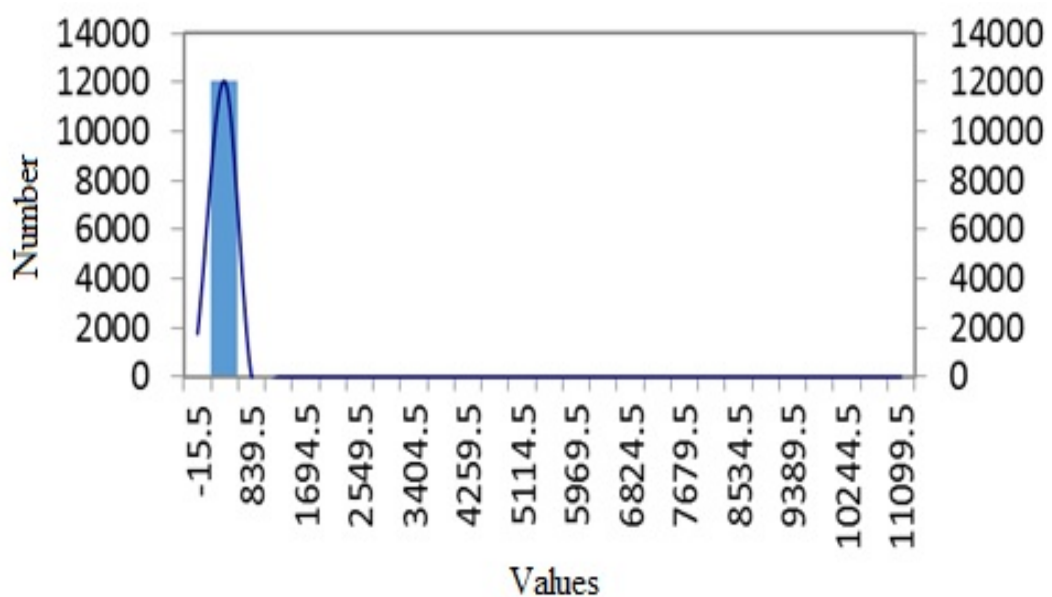
In addition to histograms, normal probability plots are also used to assess graphically that whether or not the data comprising of precipitation, solar radiation and discharges at various stations (detail in Table 3.1) is normally distributed. These plots also provide a measure of how much the data is close to follow the normal distribution with the help of correlation coefficient, R^2 . It is clear from table 4.1 that when the data is transformed using negative values of λ as -1 and -2, the respective values of standard deviation (1037399 and 101.31), skewness (109.79 and 6.23) and covariance (10.2 and 3.30) are very high as compared to other tried values of λ .

TABLE 4.1: Different values of λ against Co-variance and Histogram characteristics

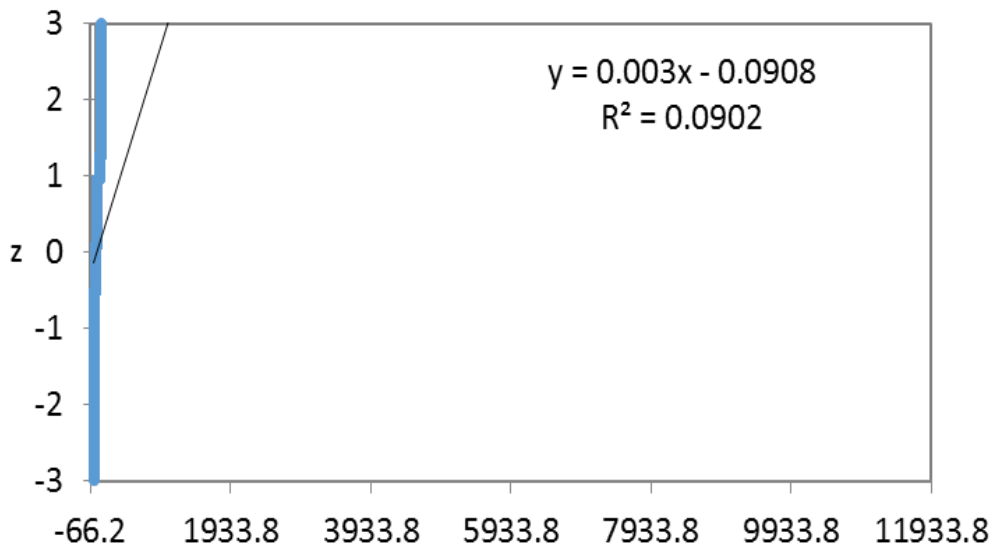
Sr. No	Lambda (λ)	Mean	Standard Deviation	Covariance	Skewness
1	-2	101201	1037339	10.2	109.79
2	-1 (inverse)	30.63	101.31	3.30	6.23
3	0.005	0.9870	0.0072	0.0073	0.41
4	0.01	0.9742	0.0143	0.0148	0.42
5	0.05	0.8795	0.0653	0.0742	0.48
6	0.1	0.7779	0.1161	0.149	0.54

Sr. No	Lambda (λ)	Mean	Standard Deviation	Covariance	Skewness
7	0.2	0.6142	0.1872	0.30	0.70
8	0.5 (sq. root)	0.3558	0.2738	0.77	0.96
9	0.8	0.2429	0.2799	1.15	1.20
10	3 (Cube)	0.0760	0.1761	2.31	2.95
11	2 (square)	0.1091	0.2125	1.94	2.16
12	1 (original)	0.1944	0.2672	1.37	1.42
13	0 (log)	-2.621	1.464	-.055	0.41

The graphical representation for $\lambda = -1$, in form of histogram and probability plot is shown in Fig. 4.1 (a) and Fig. 4.2 (b), respectively. The histogram shows that the data is concentrated on one side with mean value = 30.63, and positively skewed as the right tail is longer than the left tail. On the other hand, the normal probability plot does not show any linear relationship with a very low value of $R^2 = 0.0902$, which means that the data transformed using -1 as a power factor, is not normally distributed.



(a)

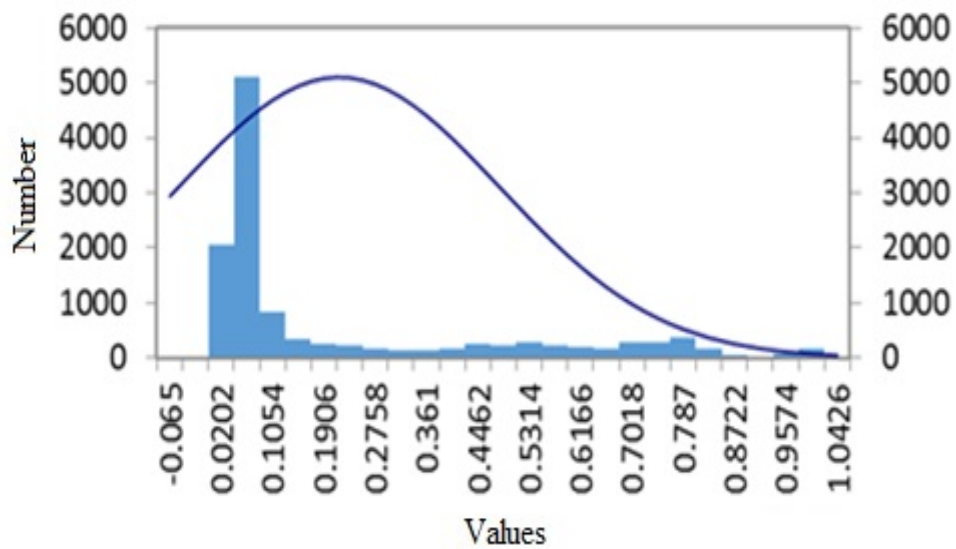


(b)

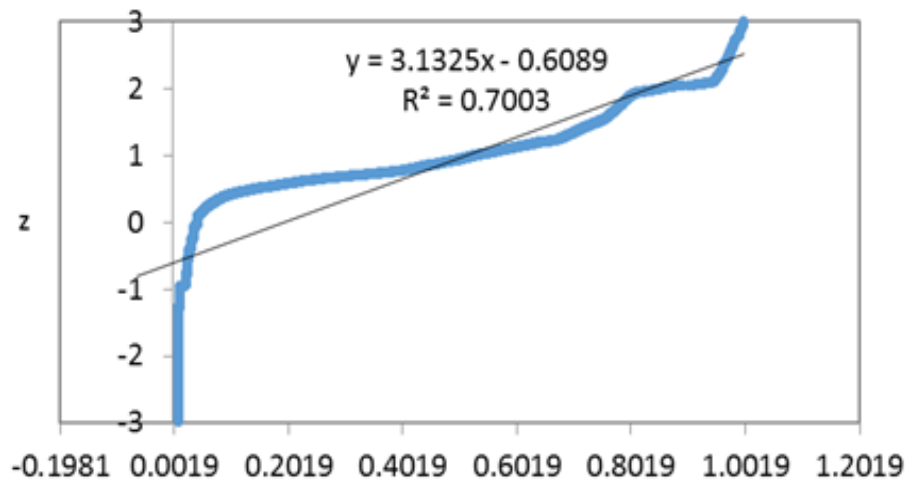
FIGURE 4.1: Transformed Data Analysis for $\lambda = -1$.

(a) Histogram, (b) Normal Probability Plot

For original data ($\lambda=1$), the values of R^2 , Standard Deviation, co-variance and skewness are 70%, 0.267, 1.37, 1.42, respectively. The departure of the data points from the linear line of the normal probability plot indicates the departure from the normality. In the case of original data, the value of correlation coefficient is 70%, which is far better than the negative values of λ . Therefore, the negative values of λ seem unsuitable unsuitable for this data-type as the transformed data comes out even more abnormal than the original data itself.



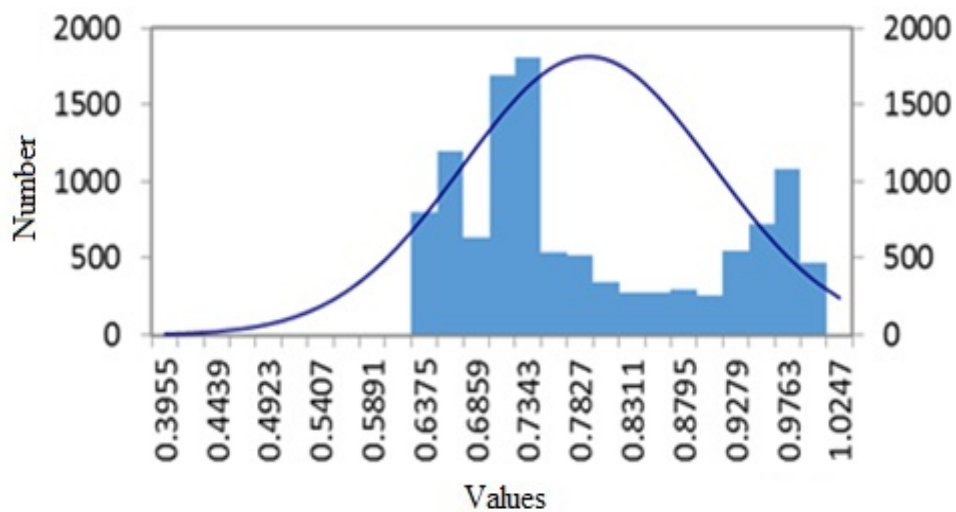
(a)



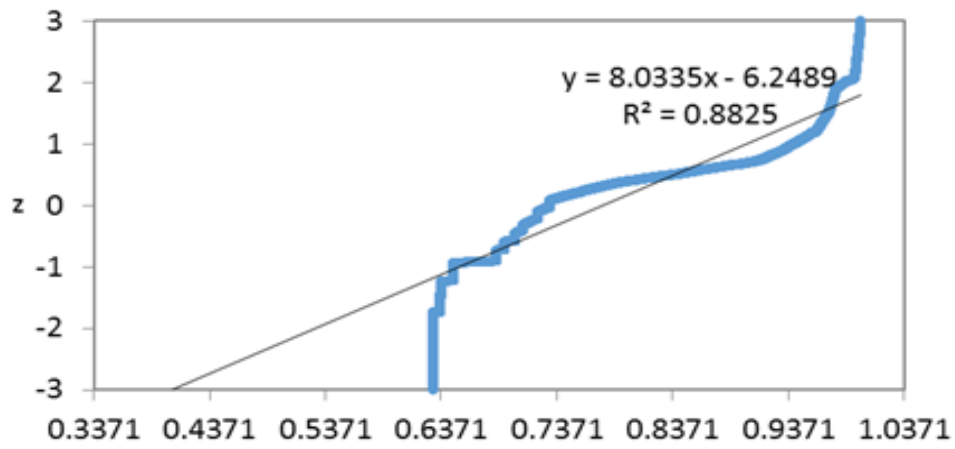
(b)

FIGURE 4.2: Original Data Analysis for $\lambda=1$.
 (a) Histogram, (b) Normal Probability Plot

However, the other values of λ , such as 0.1, 0.01 and 0.005 showed better fitness to the straight line of normal probability plots with significant high values of $R^2 > 88\%$ as compared to the original data set. The normal probability plots for these power factors are shown in Figs. 4.3(b), 4.4(b) and 4.5(b), respectively. The same is reflected in the histograms for these values of λ , which are shown in Fig. 4.3(a), 4.4(a) and 4.5(a). These histograms are bell shaped and almost seems graphically symmetrical about the mean values. This indicates that the transformed data using these power factors, approximately follows normal distribution. It is also noted that the change of λ values from 0.01 to less than 0.005 doesn't have any significant impact on R^2 value of probability plot.

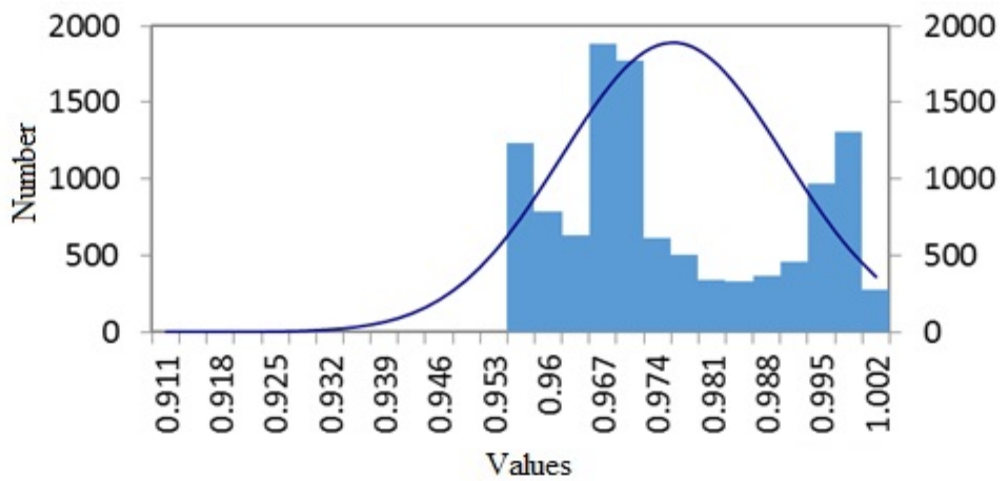


(a)

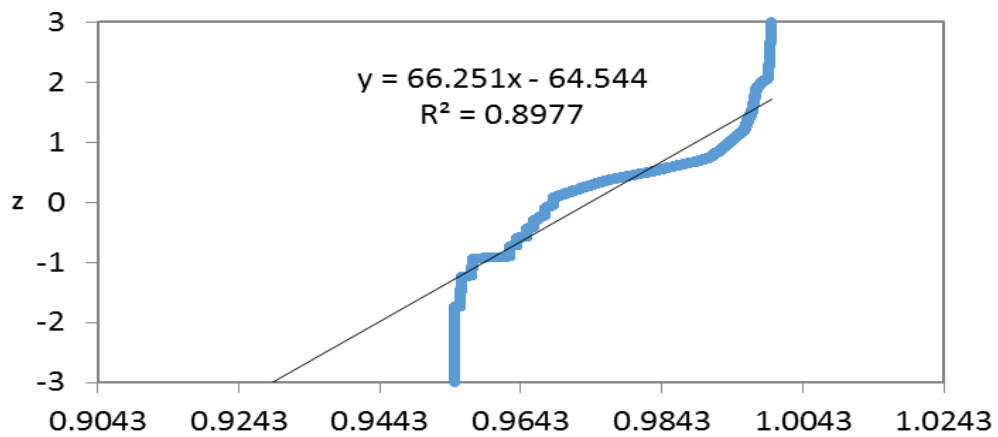


(b)

FIGURE 4.3: Transformed Data Analysis for $\lambda=0.1$.
 (a) Histogram, (b) Normal Probability Plot

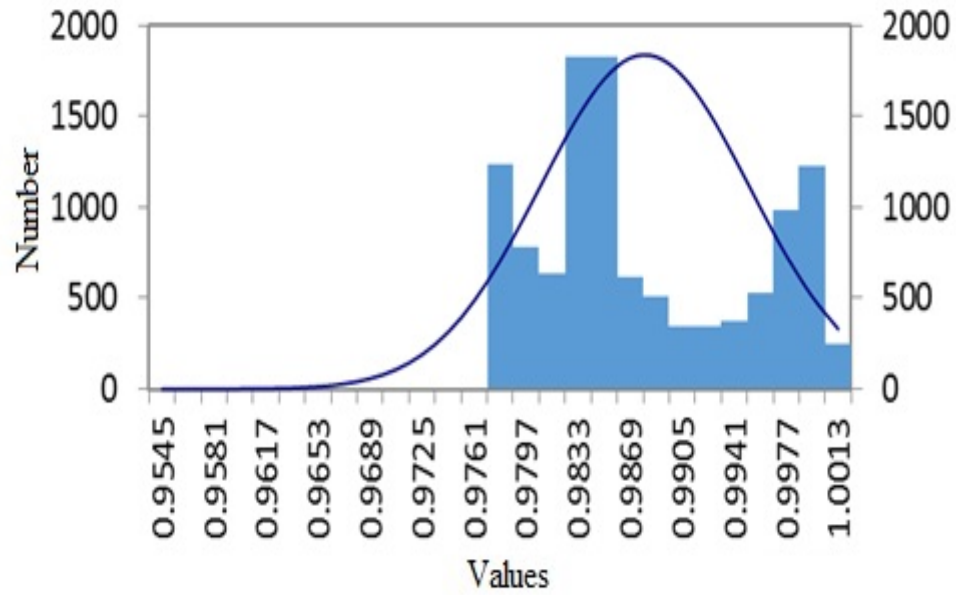


(a)

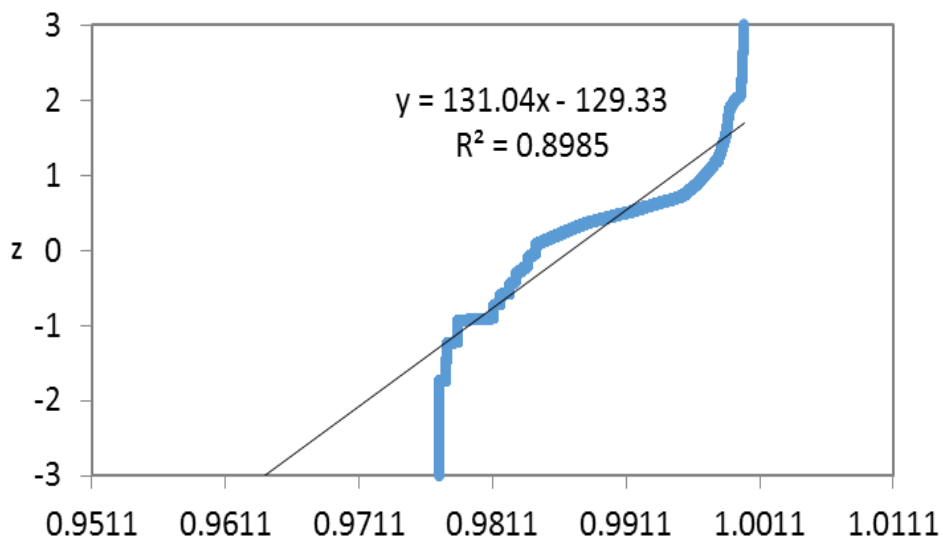


(b)

FIGURE 4.4: Transformed Data Analysis for $\lambda=0.01$.
 (a) Histogram, (b) Normal Probability Plot



(a)



(b)

FIGURE 4.5: Transformed Data Analysis for $\lambda=0.005$.

(a) Histogram, (b) Normal Probability Plot

When the value of λ is increased to 2 (square root transformation) and 3 (cube root transformation), a decreasing trend in Standard Deviation is observed. But, increased skewness for $\lambda = 2$ and 3, shows that the transformed data becomes more asymmetrical and even shifted away from the normal distribution curve. However, $1 > \lambda > 0$, makes the data more normal with reduced skewness and increased R^2 values in the probability plots.

For $\lambda=0$, the formula for Box-Cox transformation returns the data towards log transformation. The result of log transformations are shown in Fig. 4.6 (a) and 4.6 (b). The log transformation transforms the data towards normality better than square, cube or inverse transformation with 89% correlation coefficient in probability plot. The reason of log transformation performing well than them is because the value of $\lambda=0$ lies within the range of confidence interval determined for the data for a given confidence value. However, even better histogram characteristics are obtained with the other values of power factor, such as 0.001 and 0.005.

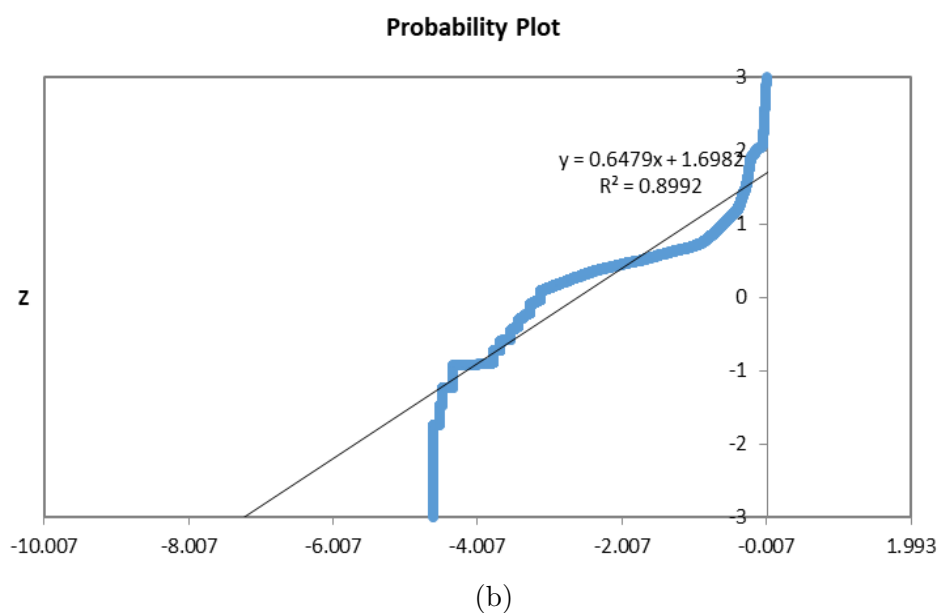
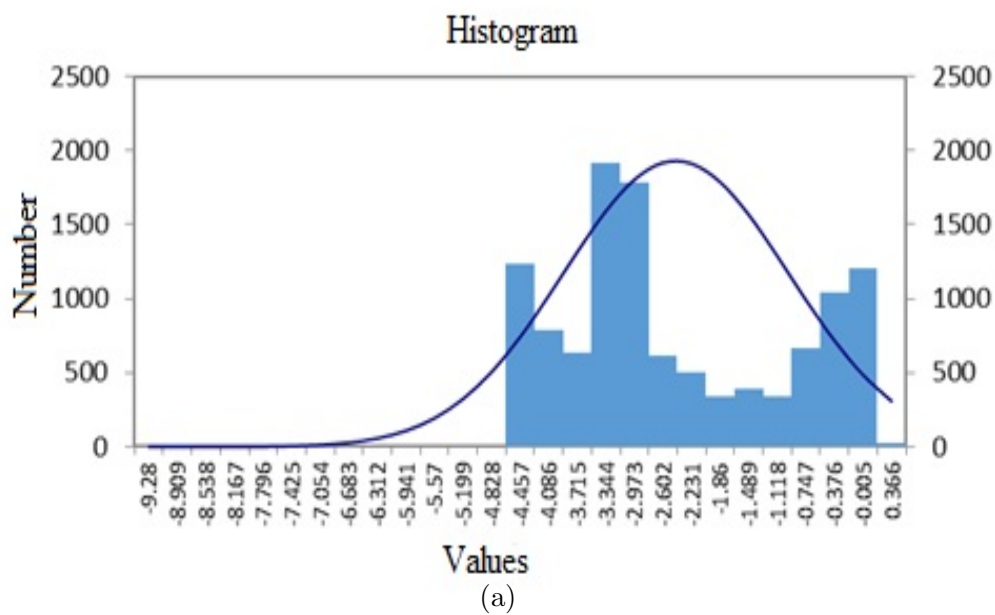
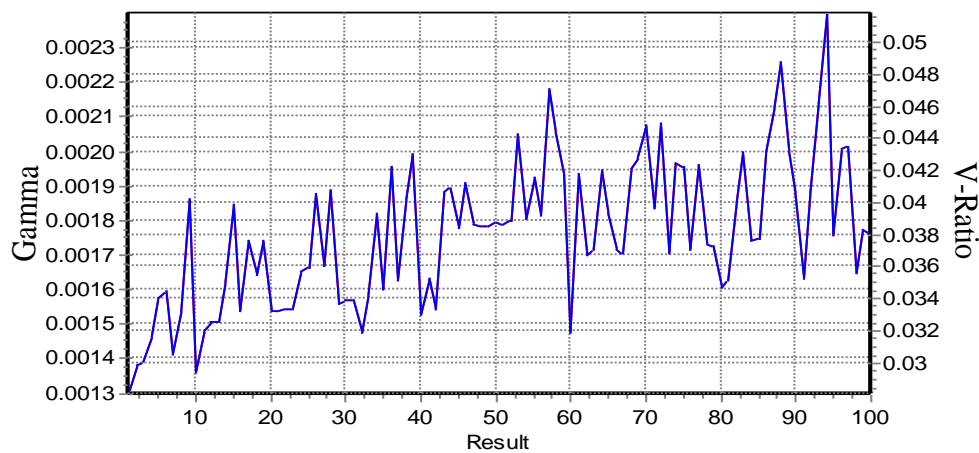


FIGURE 4.6: Transformed Data Analysis for $\lambda = 0$, Log transformation
(a) (Histogram) (b) (Normal Probability Plot)

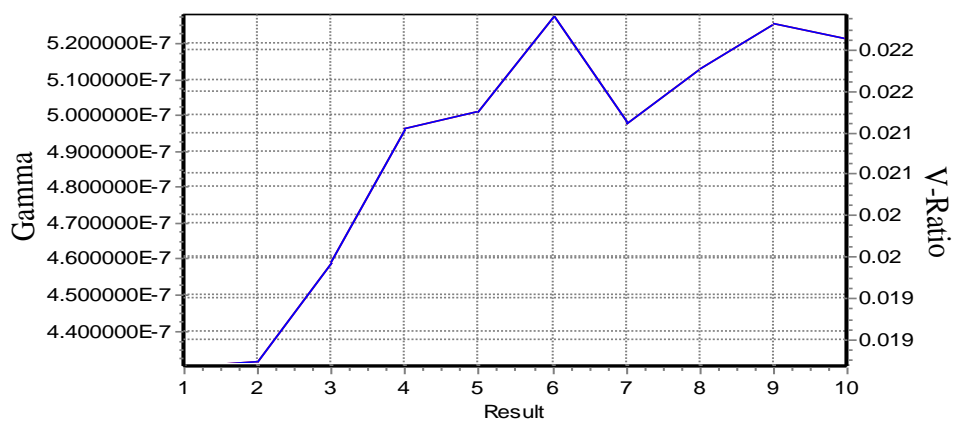
With the above discussion, it is concluded that the value of $\lambda = 0.01$ and 0.005 could be used confidently as these power factors are making the data more symmetrical about the respective mean values of the transformed data-set. In this case, $\lambda = 0.005$ is chosen to transform the data with minimum covariance (0.0073) and maximum value of R^2 (89.85).

4.2.2 Gamma Test Results

Genetic algorithm has been used as a model identification tool for making variety of combination of inputs and obtained gamma values along with V_{ratio} values that are presented graphically for original (Fig. 4.7(a)) and transformed data-sets (Fig. 4.7(b)). The result in x-axis denote the population size which is a crucial parameter in finding the optimum solution for GA.



(a)



(b)

FIGURE 4.7: Spread of Gamma values and V-ratio.
(a) Original Data, (b) Transformed Data

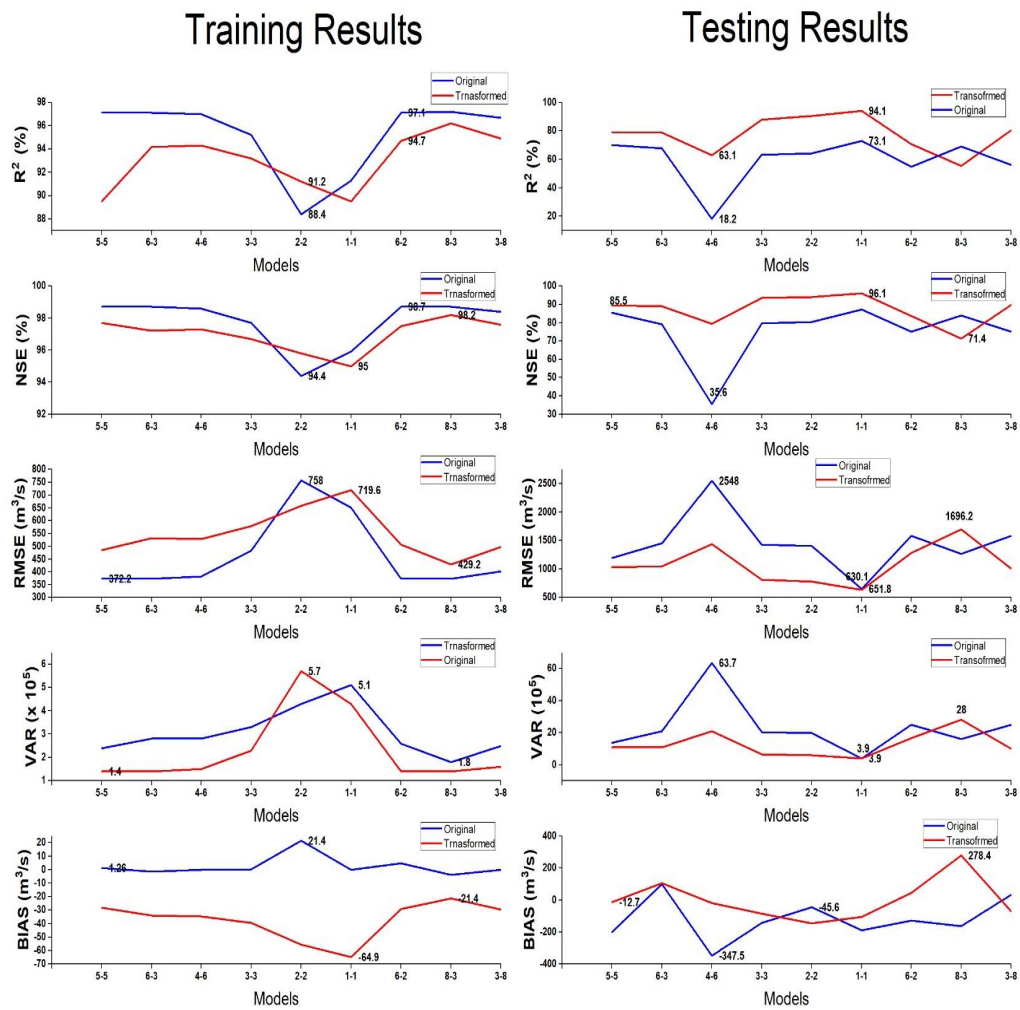
It can be seen that population sizes for GA in original data is 100 while in transformed data it is selected as 10. This is due to the reason that for transformed data, the data has been scaled down and it behaves differently than the original one. This has also been reflected in previous research [202] that population size depends upon number of factors that may include problem size, problem difficulty and or empirical evidence of improvement for a specific population size.

Although, selecting population size 100, for transformed data-set, results in very less gamma value but the developed models seem unable to achieve this unrealistically less targeted MSE (3.2×10^{-9}). Contrarily for population size 10, the achieved MSE values are quite closer to the targeted MSE values as shown in Table 4.1. The final gamma values that have been used as targeted MSE for training of ANN models are 0.0012976 and 4.3006×10^{-7} , respectively for original and transformed data. It should be noted that the MSE for transformed data-set is far low than the MSE for the original data. The less value, in case of transformed data, is due to scaling of data between 0 and 1 and might be the result of reduced heteroscedasticity in data through Box-Cox transformation. It is already discussed in Chapter 3, under section 3.2.3.2 that the V_{ratio} is used to standardize the gamma value as the GT bears the assumption that the noise present in data is only due to the statistical noise. It is clear from the Fig. 4.7(a) and 4.7(b) that for the targeted values of MSE for both original and transformed data-sets, the value of V_{ratio} is also close to zero. Therefore, it could be stated that in this case, the outcome which is being predicted is not of probabilistic nature and the gamma value is the true reflection of noise present in the data and could be used confidently as the targeted MSE to train models.

4.2.3 ANN Model Results

Initially only R^2 is used to define the architecture of ANN models, but because there is no significant difference in the R^2 values of different tried architectures of ANN. Therefore, other statistical parameters including NSE, RMSE, VARIANCE and BIAS have also been calculated for all set of ANN model structures which are presented in Fig. 4.8, in order to verify the initial selection and to choose the most appropriate model(s) for predicting the discharge at Tarbela. It is mentioned earlier that node arrangement 1-1 and 2-2 gave the highest R^2 values, respectively for original and transformed data-sets.

The same selection has been justified with the help of NSE, RMSE, Variance and BIAS values which are presented in Fig. 4.8. It is clear from the Fig. 4.8 that approximately all errors have been reduced for the models developed through transformed data specifically in testing phase, as compared to the original data-set but the difference in values of R^2 is more significant. In case of original data, the developed models have performed well only in training phase; with NSE and $R^2 > 95\%$ and low values of RMSE (< 400) in most of the cases. The low values of NSE and R^2 and high values RMSE and BIAS in testing phase clearly depict the inefficiency of models developed through original data. On the other hand, the models, developed through transformed data, not only resulted in high values of NSE (more than 89%) and R^2 (more than 90%) in training phase but also resulted in reasonably high values of NSE (more than 97%) and R^2 (more than 94%) in testing phase for most of the cases.



(b)

FIGURE 4.8: Comparison of ANN models for Original and Transformed data

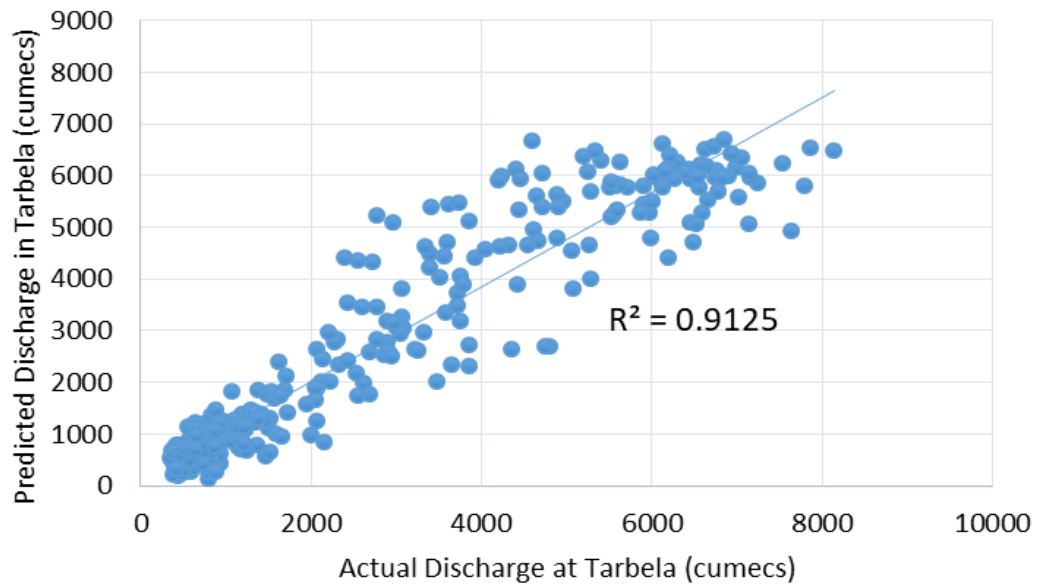
4.2.4 Discussion

In case of original data, the models with node arrangement 1-1 and 2-2 performs better with $R^2 > 70\%$ in testing phase. However, all the other models, developed through original data are unable to perform well with weak correlation and reduced model efficiency (Fig.4.8) specifically in the testing phase with high values of RMSE, BIAS and Variance. In training phase these combination of nodes (1-1 and 2-2) did not perform as well as compared to the combination of more nodes in both the layers (5-5, 6-3, 4-6, 6-2 and 6-3) with relatively less values of R^2 , NSE and high values of BIAS, VAR and RMSE. Whereas, for transformed data, the arrangement 1-1, 2-2 and 3-3 could be used with high values of correlation coefficient in both phases which are 89.5 and 94.5, 91.2 and 90.4, 93.2 and 87.9, respectively. However, the model with node arrangement 2-2 is considered as the best model because the corresponding statistical parameters show that it can be used for the prediction of discharge at Tarbela with a reasonable accuracy.

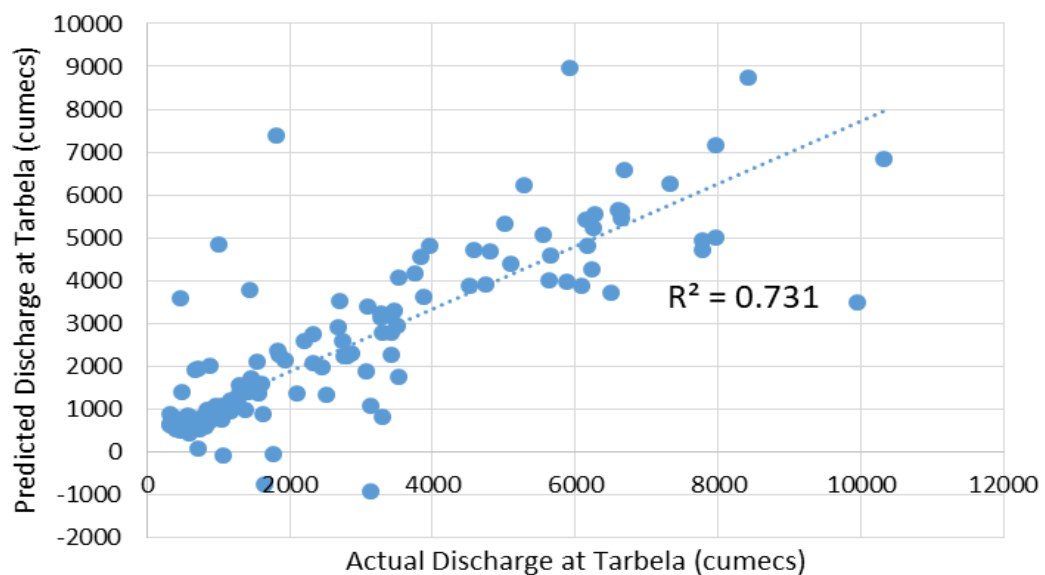
The value of R^2 for network with node combination 8-3 in testing phase is low because of the reason that this particular architecture of ANN containing this combination of nodes in both layers, unable to perform well in testing phase. This raises a question why some networks of ANN performs well and some not? Well, the best practice to find the appropriate network is to experiment them, as did in this case. However, one particular reason behind this low value of R^2 is that the network has more neurons than it should be, as it is clear from Fig. 4.8 that the same model is over-fitting (not performing well in testing phase, the BIAS value for the same model is also very high which are shown in Fig. 4.8. The results of table 4.1 and Fig. 4.8 clearly shows that the less number of nodes perform well in ANN modeling process for our data.

Although a single model from original data with a node combination (1-1) perform reasonably well but the overall trend of performance values for models developed through original and transformed data-set suggests that the transformed data-set provided a better initial state for model development process. The negative BIAS error for most of the models shows that the models are predicting less than the actual values and these could be used efficiently for water management purposes as it adds a factor of safety for the water storage and regulation. However, these models are not recommended to use for flood estimation purposes without correcting the BIAS error. The best models for original data set with node combination (1-1) is presented graphically in Figs. 4.9(a),

4.9(b). The value of R^2 in training phase is more than 90% which depicts that the predicted output is quite well correlated with the observed output. However, in testing phase the value of R^2 is 70% which is in acceptable range but not comparable to the R^2 value in the testing phase. The goodness of fit in both phases must be good for a reliable and efficient estimation model, which is lacking in the models developed with the original data-set.



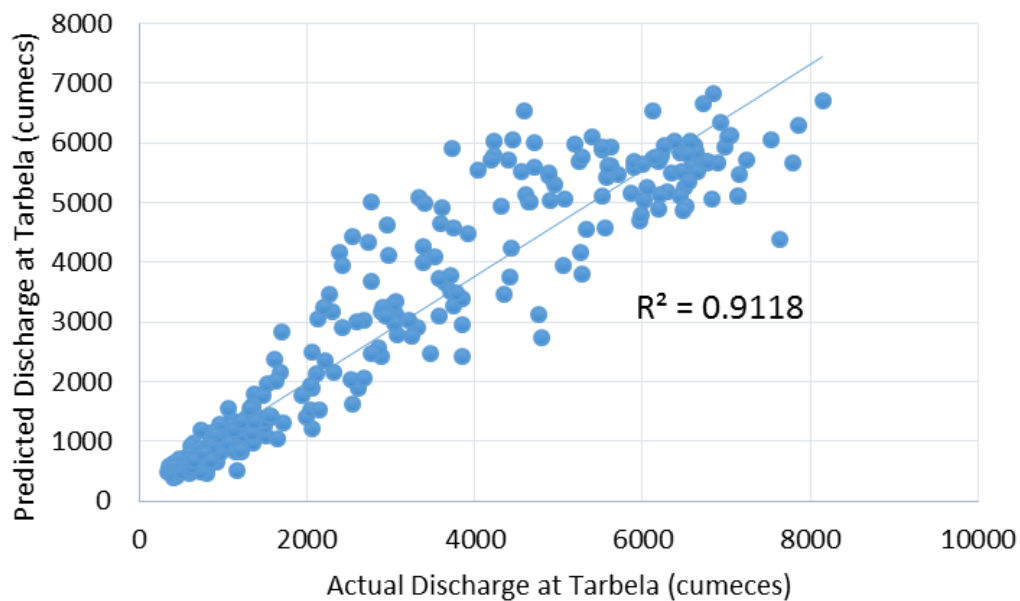
(a)



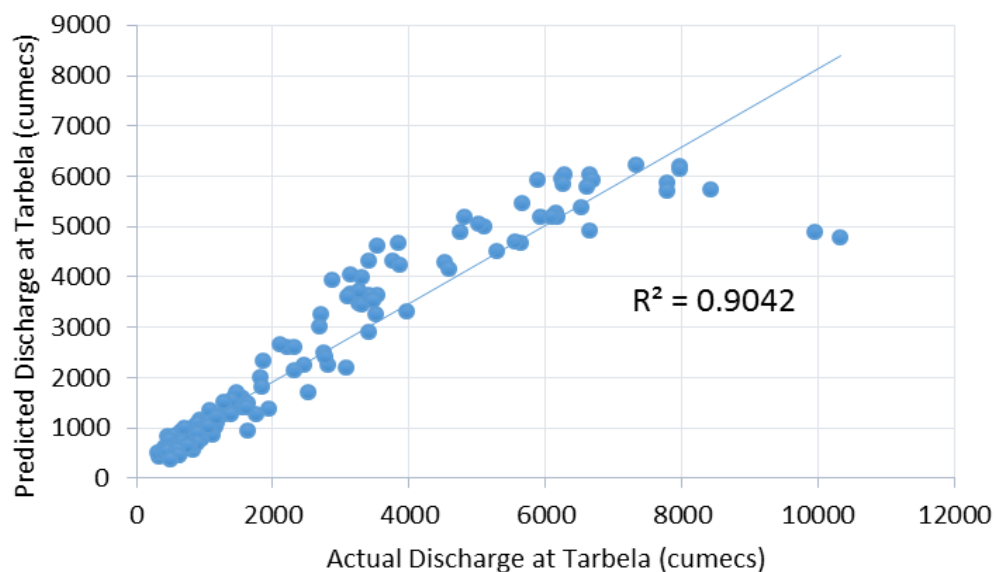
(b)

FIGURE 4.9: Model 1-1 for Original Data.
(a) Training, (b) Testing

The best models for transformed data-sets (2-2) are presented graphically in Figs. 4.10 (a), 4.10 (b). It is clear from the figures that the value of R^2 is more than 90% in training as well as in testing phase. This clearly depicts that how well the predicted output is correlated with the observed output in both phases. The same trend has been observed in most of the models developed through the transformed data-set, showing the reliability and efficiency of discharge estimation models. This clearly evidenced the importance of data transformation in hydrological data-driven model development.



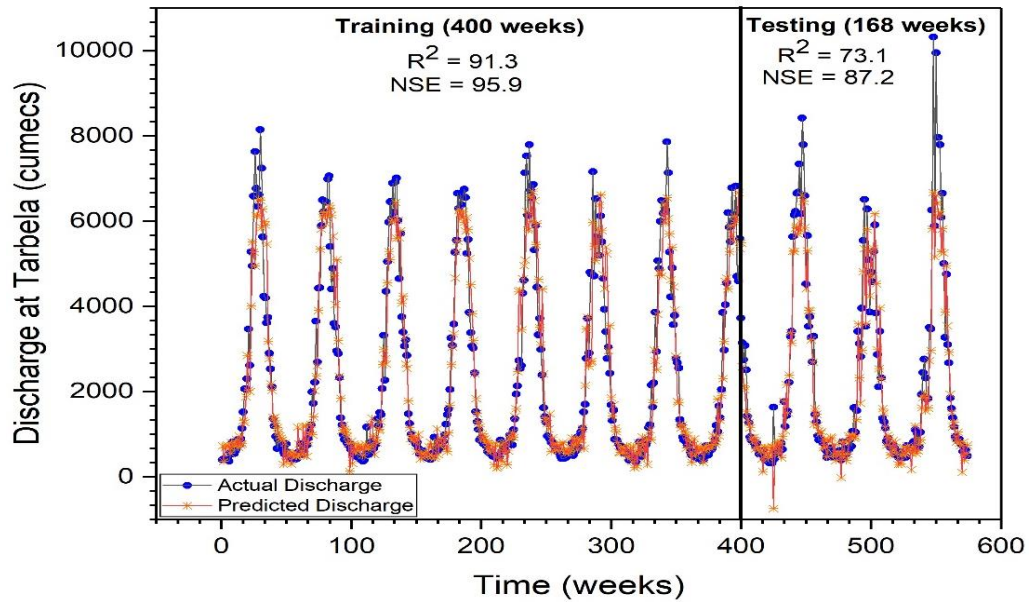
(a)



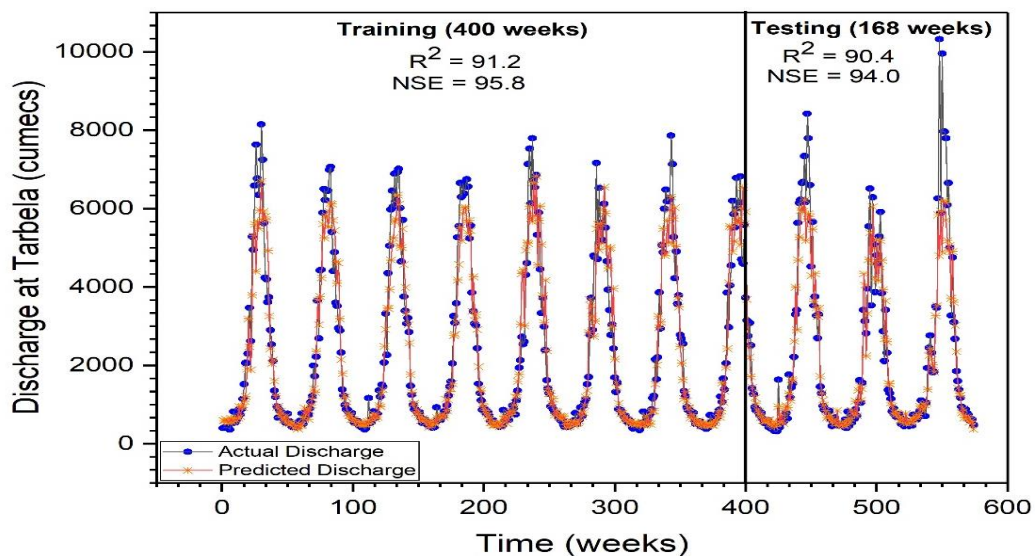
(b)

FIGURE 4.10: Model 2-2 for Transformed Data.
(a) Training, (b) Testing

The time series plots for models presented respectively in Figs. 4.9 (best model developed through original data-set) and 4.10 (best model developed through transformed data-set) are presented in Figs. 4.11 (a) and 4.11 (b). The time series plots show the real time variation of predicted discharge with respect to the actual discharge.



(a)



(b)

FIGURE 4.11: Time-series plot for models developed using original & transformed data-set

(a) Original Data, (b) Transformed Data

The time series plot for the transformed data set (Fig. 4.11 (b)) showed less variation of predicted discharge to actual discharge as compared to the time series plot developed for original data-set.

The ANN models developed through original and transformed hydrological data-sets, are both compared on the basis of variety of statistical indices. The results clearly indicate that the models developed with preprocessed data performed better with high values of correlation coefficient and less value of other statistical errors.

The comparison of models indicate that the processed hydrological data through a linear transformation (The Box-Cox transformation) provides a better initial state to the training of ANN models and could be used to improve the performance of ANN models. However, this transformation doesn't always provide optimal solution of correcting data and in some cases, the complex transformations are unavoidable [203]. The models developed through original data, although performed well in the training phase but failed in producing better results in the testing phase with low values of correlation coefficient and high values of other statistical errors.

4.2.5 Summary

In this research work, the efficiency of ANN based hydrological forecasting models have been improved through two types of data preprocessing options; data scaling through Box-cox transformation, and input selection through the Gamma test. For this purpose, a case study of UIB has been considered and streamflow forecasting models are developed for Tarbela Reservoir. Antecedent upland catchment information including precipitation, solar radiation and discharge has been considered as input variables.

The original data-set has been transformed using the Box-cox transformation after finding a suitable power factor through histogram statistics and probability plots. For given data-set, the value of the power factor (λ) with least skewness, standard deviation and maximum R^2 comes out to be 0.005. The input screening procedure is carried out with the help of a mathematical tool (Gamma-Test) and facilitated by a model identification tool, Genetic Algorithm (*GA*).

The best input combination for original and transformed data-sets are 1011001011111110001 and 10101110100110111011 with minimum Gamma values of 0.0012976 and

4.3006×10^{-7} , respectively. These values are considered as targeted MSE values to train ANN models, via two layer BFGS algorithm. Multiple node arrangements have been tried for both the hidden layers and the best models for the original and transformed data-sets, which came out “1-1” and “2-2” respectively. The evaluation of models is made on the basis of set of performance indicators including NSE, R^2 , RMSE, Variance and BIAS.

The results indicate that data preprocessing provides an opportunity to enhance the model efficiency through data transformation and input variable selection. Even simple data transformation techniques can improve the initial data state to the data driven models. This is achieved through reducing the statistical noise in the data and making it more normal (symmetrical around mean). Despite of the fact that ANN dont have any preliminary requirement of data normality to perform well, still there performance is improved. Especially, the models developed with preprocessing are more generalized as compared to the models developed with the original data-set. i.e. their capacity of performing well for the unseen data (testing data) is significantly improved.

4.3 ANN Models developed using Satellite Derived Snow Cover Area

This section covers the results targeting the second objective of the research work which is achieved to improve ANN based streamflow estimation models through satellite derived snow cover area for a mountainous catchment. The section contains results for ANN models developed using two types of dataset comprising of; on-ground flow observations and flow observations with SCA, their comparison, a comprehensive discussion followed by the summary.

4.3.1 Gamma Test Results

The Gamma test is performed on a set of six (6) input variables (SCA of Astore, Gilgit and Bunji and Q of Astore, Gilgit and Bunji) for output which is Q at BeshamQila. A total of 63 realistic combinations are made using a model identification tool and Gamma

value is computed for each, as shown in Fig (4.12), which is later on considered as the targeted MSE value for model training process (Table 4.2).

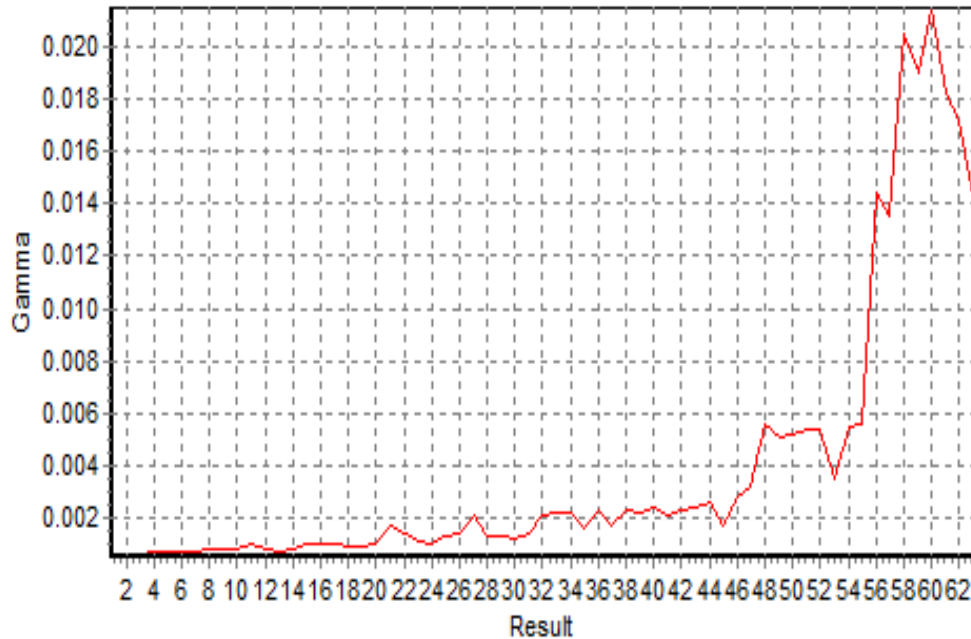


FIGURE 4.12: Variation in Gamma Value with different masks of input variables

It is found that input mask 111111 (SCA of Astore, Q of Astore, SCA of Gilgit, Q of Gilgit, SCA of Bunji, Q of Bunji) give minimum gamma value / targeted MSE that is 0.000536. “1” means that a particular input is included and “0” means a particular input is excluded. All 1’s in this combination showed that gamma value is minimum with all inputs included. So, the combination containing all input variables with minimum targeted MSE is further used for model development process.

The combination that includes only on-ground observations 010101 (Q of Astore, Gilgit and Bunji) is also used to develop models in order to compare the performance of estimation models without considering the satellite-derived SCA as a possible predictor. In order to find the suitable data length for model training, which optimize the model performance in terms of goodness of fit, the M-test has been performed on increasing number of inputs for both the input combinations 010101 Fig (4.13) and 111111 Fig (4.14).

It is clear from the Fig (4.13) & Fig (4.14) that the standard error line becomes almost stable after 165 unique data points. So, the data length for model training is considered as 45% (165 weeks) and the testing/ validation as 55% (203 weeks).

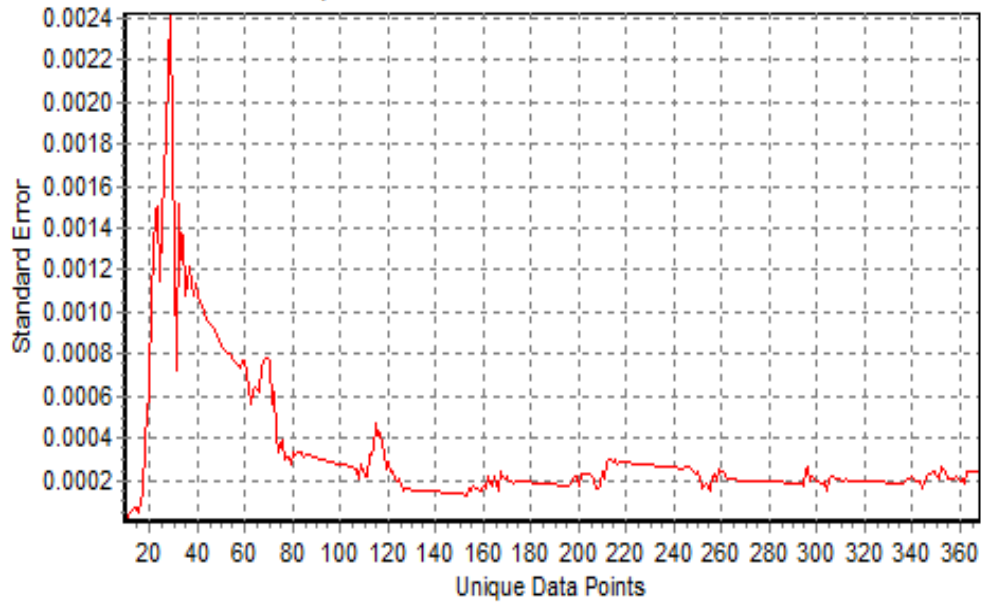


FIGURE 4.13: Stabilizing the Gamma Value with increasing data points for combination 010101

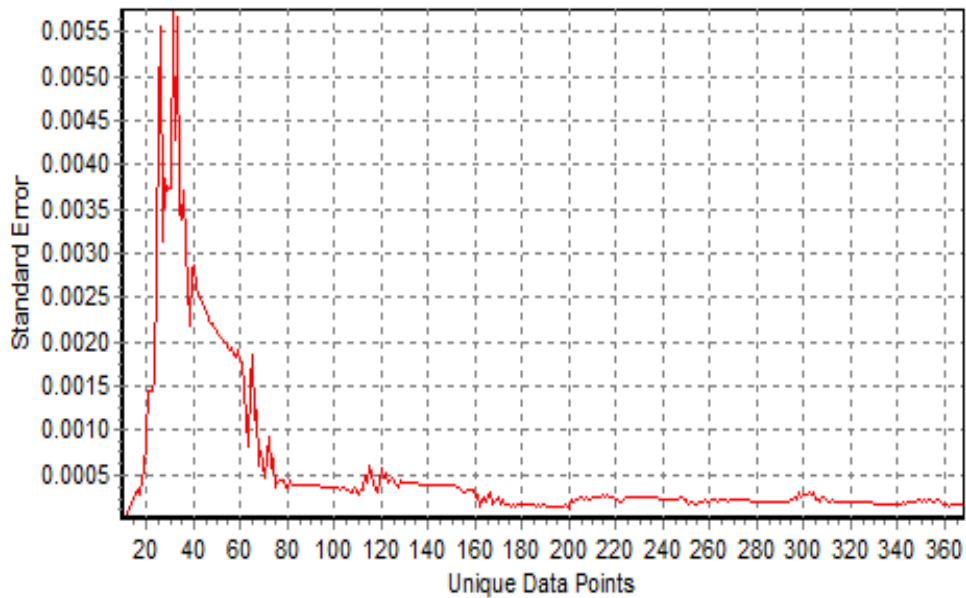


FIGURE 4.14: Stabilizing the Gamma Value with increasing data points for combination 111111

4.3.2 ANN Model Results

The number of hidden layers for all ANN models are fixed as two (2) and a variety of combinations of nodes in these layers are tried (Table 3.4). Since, the change in

number of nodes in each layer doesn't have any significant impact upon the MSE and the correlation coefficient values for the developed models, therefore, the models are developed for all the selected combination of nodes as presented in Table 4.2, tested and evaluated using performance indicators; NSE, RMSE and VAR. Each model with a specific network architecture is trained against the output discharge at BeshamQila for both the combinations, the one (111111) obtained through the Gamma-test and the other (010101) comprised of only gauge-discharges Fig. (4.15).

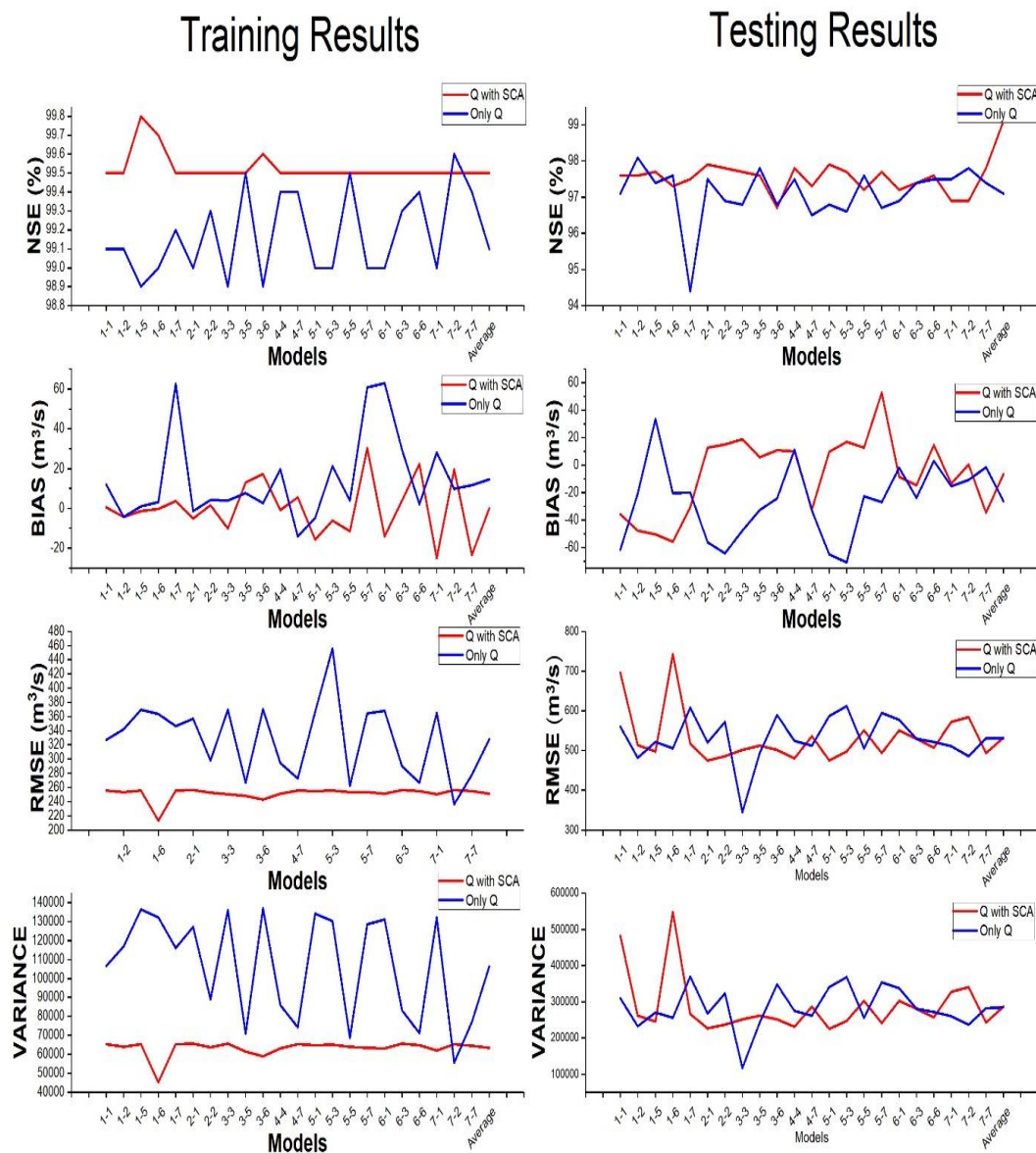


FIGURE 4.15: ANN Modeling Results for both set of input-combinations

4.3.3 Discussion

It is clear from the Fig. 4.15 that the models developed with integrated data-set (111111), performed better with the average values of NSE = 99.5/97.5 (training/testing), BIAS = -0.01/-6.6, RMSE = 251.4/532.3 and VAR = 63218.0/286917.1, as compared to the models developed without SCA in the input variables (010101) with average values of NSE = 99.1/97.1 (training/testing), BIAS = 14.6/-26.1, RMSE = 327.6/531.4 and VAR = 106390.6/284363.4.

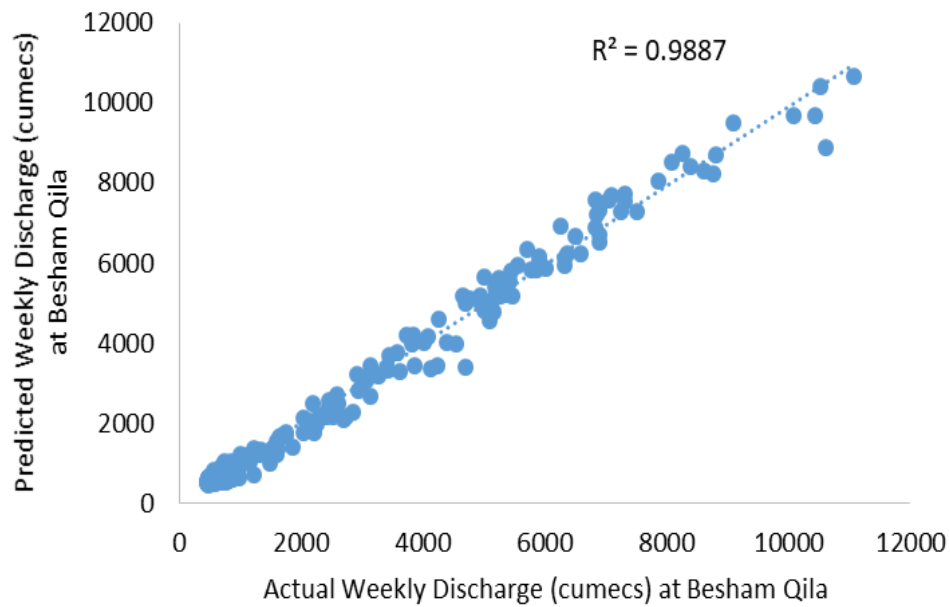
Again, the changing number of nodes in the hidden layers doesn't produce any significant difference in the performance of models with almost similar values of performance indicators, especially for NSE and RMSE in the case of training and testing of models.

However, the moderate difference in values for BIAS and VAR is observed during both the training and testing phases of the developed models. BIAS is a systematic error that represents the difference in values of predicted and actual mean. Positive BIAS means that the mean of predicted discharge is more than the mean of actual or observed discharge.

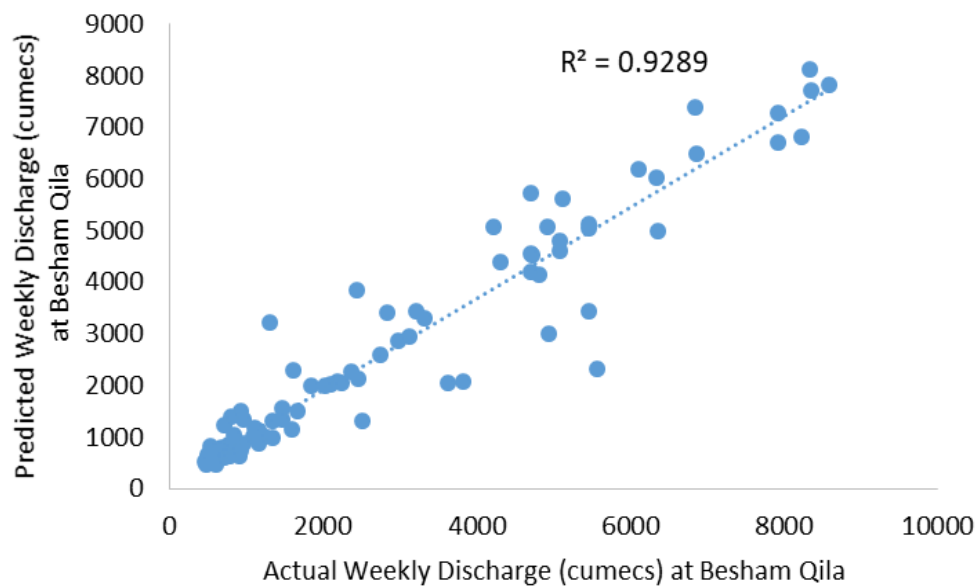
In the training phase of models developed with 010101 (input combination), the value of BIAS in most of the cases is positive, depicting that the trained models are over-predicting; whereas the converse trend of negative BIAS has been observed in most of the cases of testing/validation phase, showing that the models are predicting less than the actual values.

To make the ANN models predict more or less, often depends upon the careful selection of network-architecture as the complex architecture tends to predict more due to over-fitting, while the lighter network tends to predict less due to under-fitting. This is why, this study considered a number of node-combination options while training ANN models through BFGS algorithm to find the best possible option with minimum uncertainty.

For combination 111111, the minimum value of BIAS in the training phase is 0.6 for Model No.1 with a node combination of 1-1. But the same model showed high values of BIAS (-36.0) and RMSE (695.3) in the testing phase. The values of correlation coefficient (R^2) for this node combination is 0.99 and 0.93 as shown in Fig. 4.16(a) & (b) respectively.



(a)

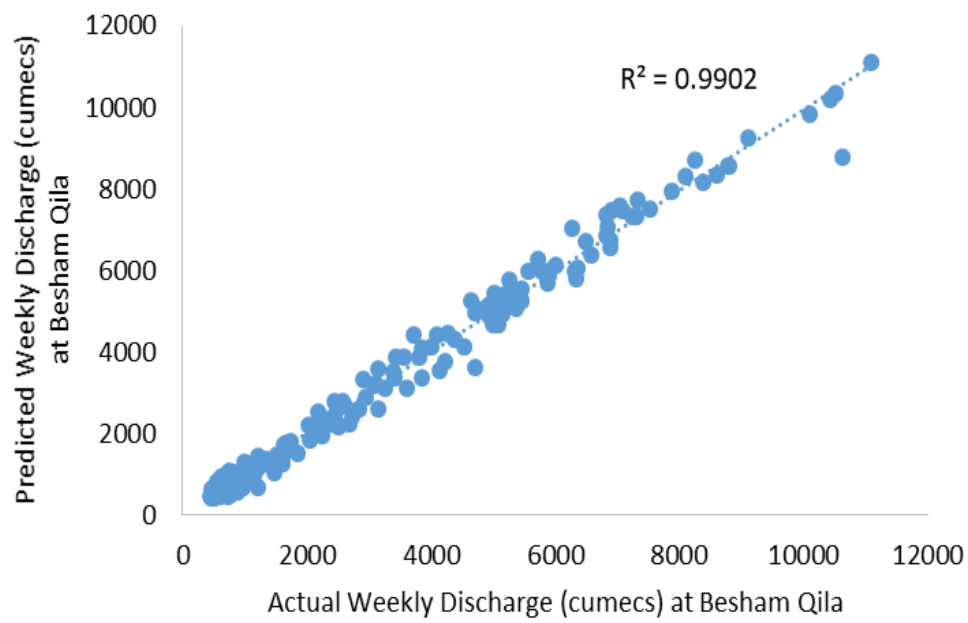


(b)

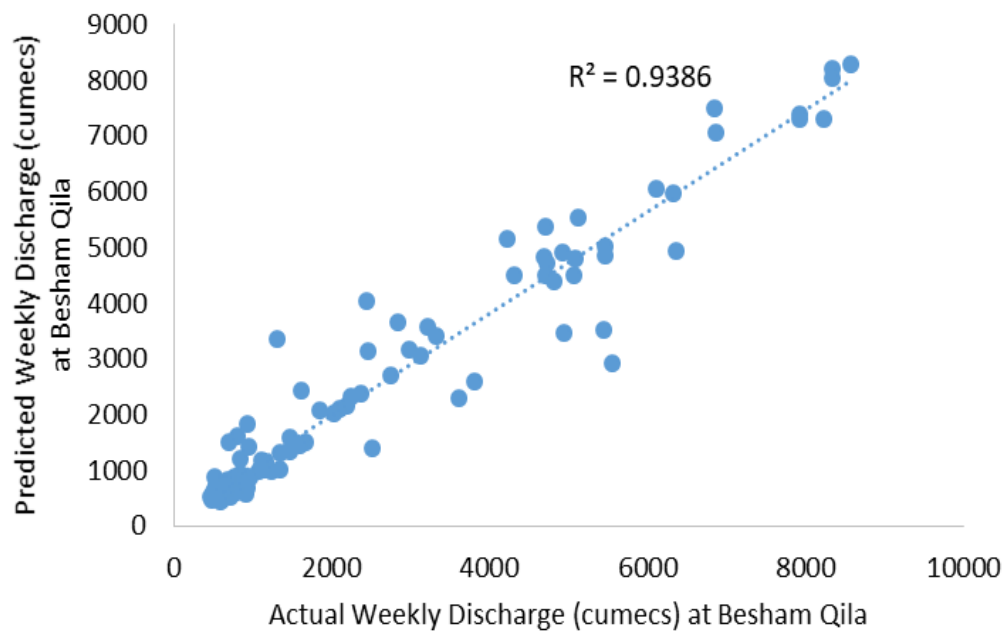
FIGURE 4.16: Model No. 1 (Nodes: 1-1) developed with input combination 111111

(a) Training Phase, (b) Testing Phase

The model No. 7 with a node combination of 2-2 performed well with low values of BIAS Training/Testing = (1.6/15), RMSE (252.4/486.0) and VAR (63709.6/236002.2). The value for R2 in training and testing phases for this node combination are 0.99 and 0.94 respectively for training and testing phases, as shown in Fig 4.17 (a) & (b).



(a)



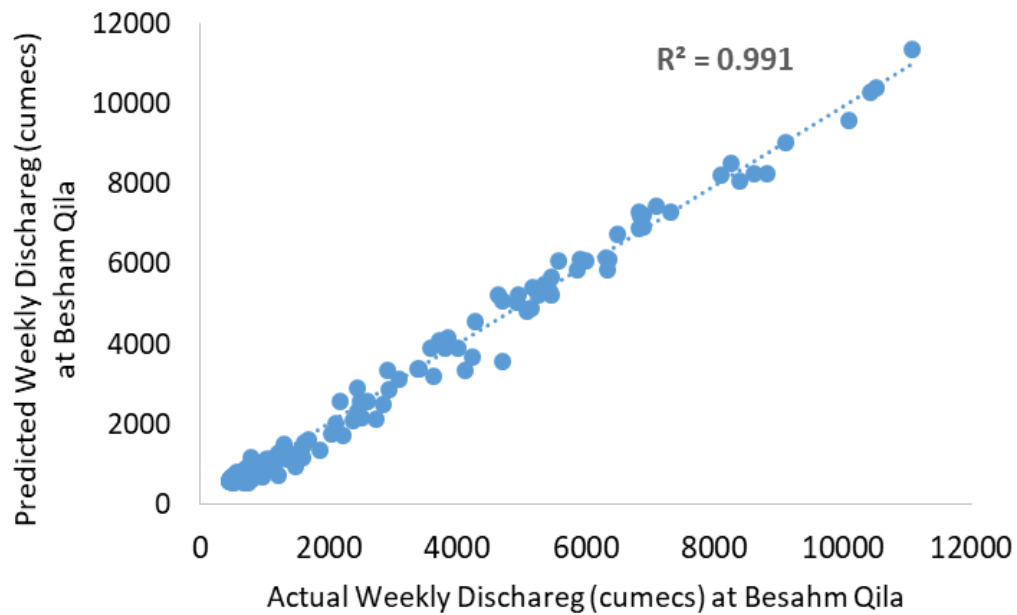
(b)

FIGURE 4.17: Model No. 7 (Nodes: 2-2) developed with input combination 010101

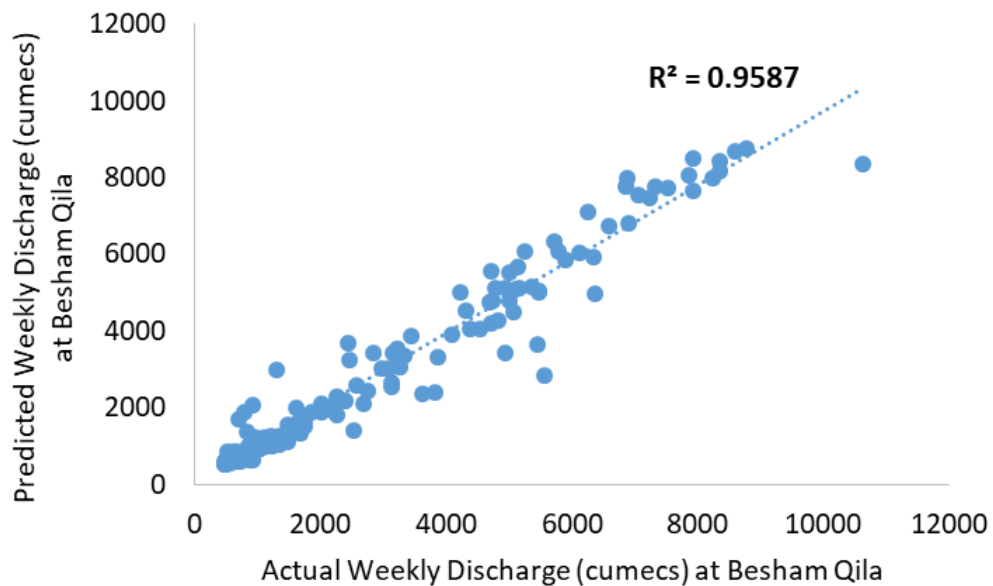
(a) Training Phase, (b) Testing Phase

The model No. 11 with a node combination 4-4 outperformed with low values of BIAS Training/Testing = (-0.9/9.9), RMSE (251.1/480.9) and VAR (63062.9/231196.7), as

compared to other models. The Fig 4.18 (a) & (b) shows the training and testing models developed with a node combination of 4-4. The values for R2 in training and testing phases are 0.99 and 0.96, respectively.



(a)

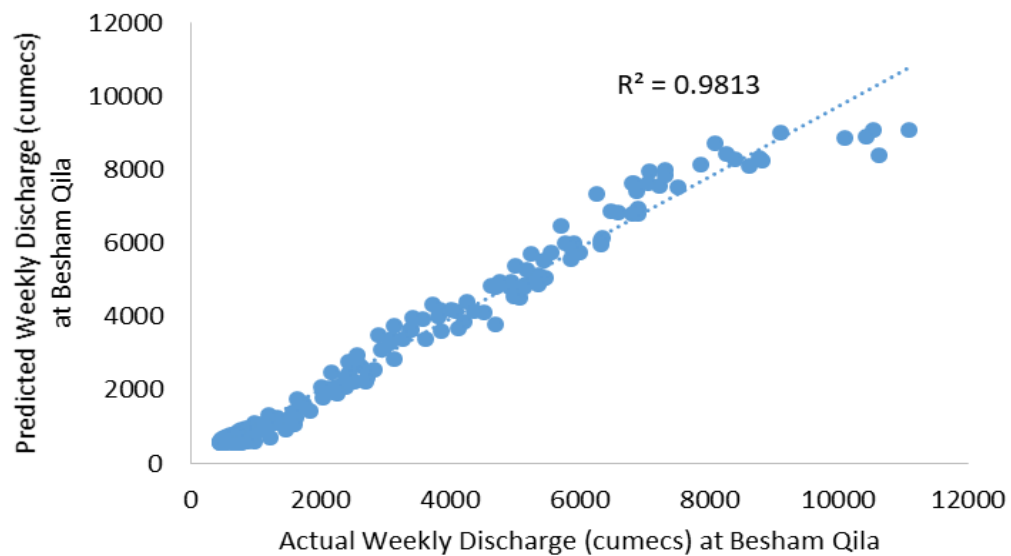


(b)

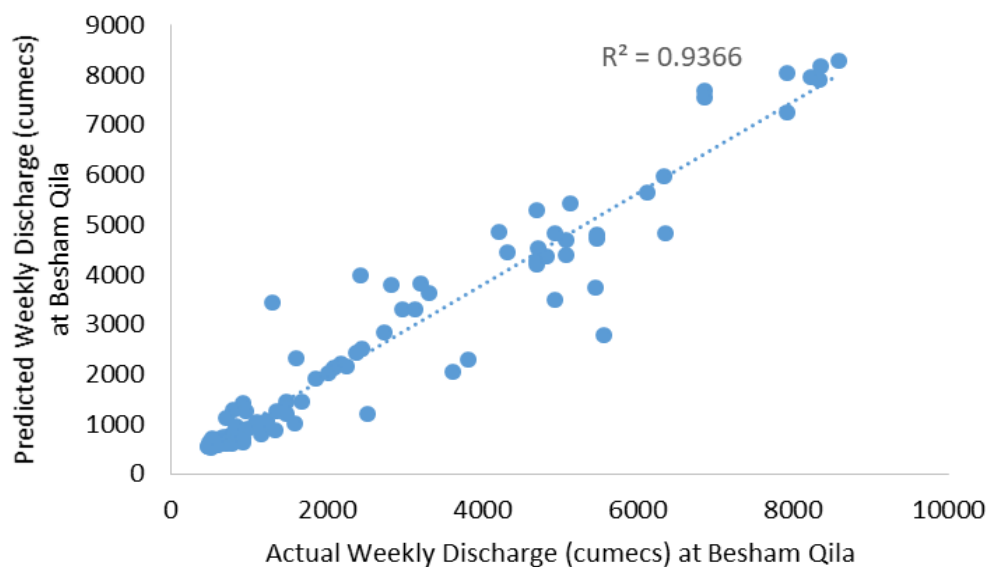
FIGURE 4.18: Model No. 11 (Nodes: 4-4) developed for input combination 111111

(a) Training Phase, (b) Testing Phase

Although, some of the models developed without SCA also performed reasonably well, e.g. the model No. 21 with a node combination of 7-2 and the model No. 19 with a node combination of 6-6 with BIAS (9.7/-10.9) & (2.0/3.2), RMSE (236.0/ 485.9) & (266.9/520.9) and VAR (55593.5/235948.3) & (71221.6/271285.9). The training and testing models for these models are presented in Fig. 4.19 (a) & (b) and Fig. 4.20 (a) & (b), respectively.



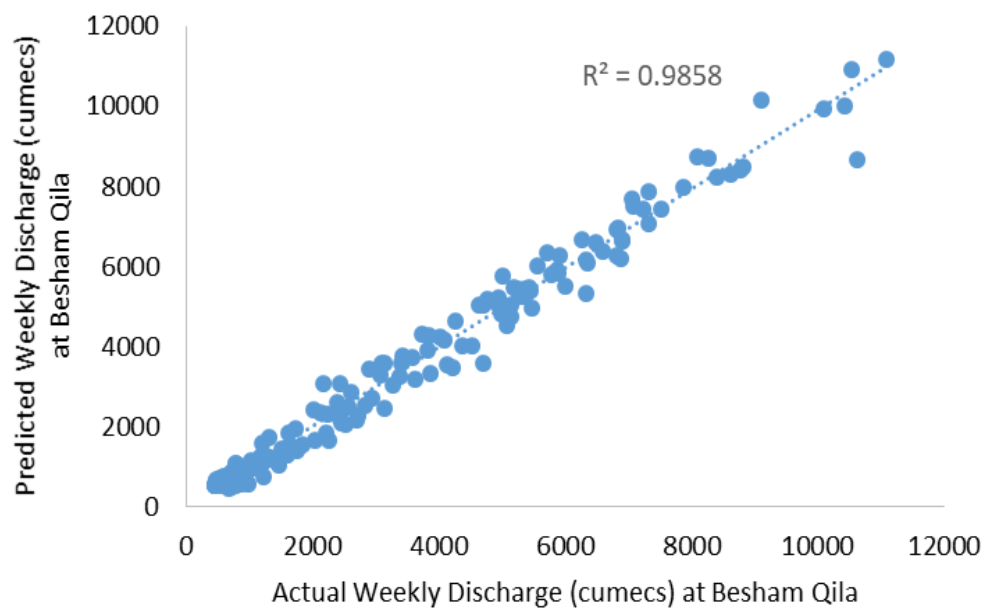
(a)



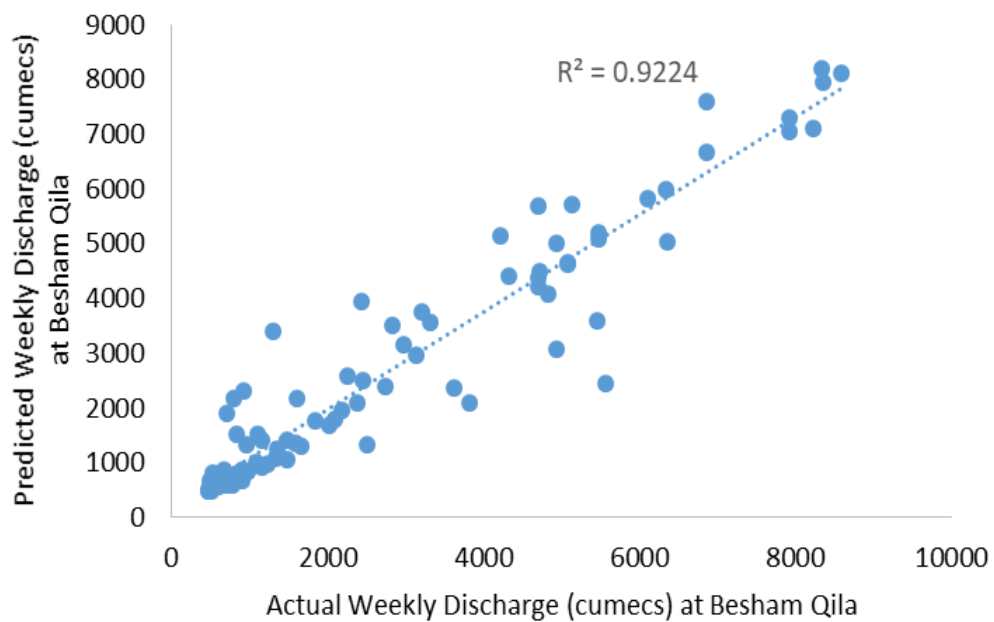
(b)

FIGURE 4.19: Model No. 21 (Nodes: 7-2) developed for input combination 010101

(a) Training Phase, (b) Testing Phase



(a)



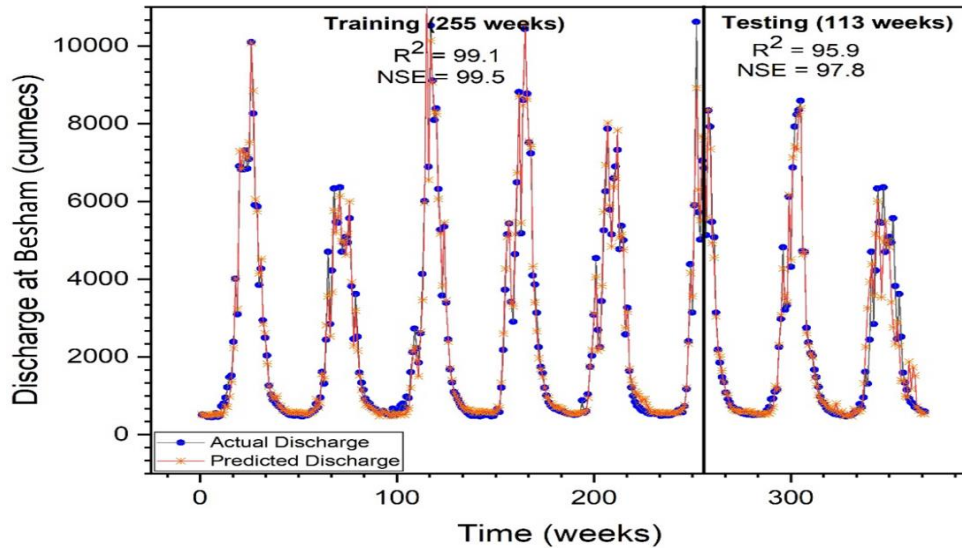
(b)

FIGURE 4.20: Model No. 19 (Nodes: 6-6) developed for input combination 010101

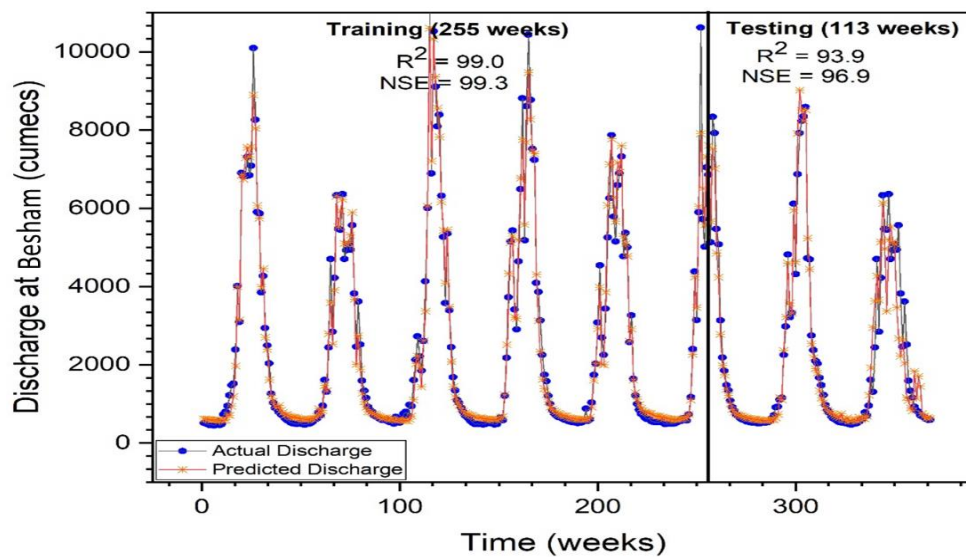
(a) Training Phase, (b) Testing Phase

However, overall trend of results suggest that the models developed with integrated data has the better tendency to perform well with significant high values of NSE. The initial

data-fusion through Gamma-test has also provided the same that the inputs with a mask 111111 (SCA of Gilgit, Q at Gilgit, SCA of Bunji, Q at Bunji, SCA at Astore, Q at Astore) will produce better models with least MSE. The time series plots for actual and predicted stream flows at Besham Qila are presented in Fig. 4.21(a) and Fig. 4.21 (b) respectively for combination 111111 and 010101.



(a)



(b)

FIGURE 4.21: Time series plot for Besham Qila developed with combinationS.
(a) 111111 , (b) 010101

The results indicate that the discharge estimation models for Besham Qila, trained via two layer BFGS algorithm, performed better with the multi-source information (on-ground and satellite) as compared to the single source information (on-ground).

The results clearly showed the dependency of stream-flow of Upper Indus Basin on the upland condition of snow cover area (SCA) and evidenced the importance of satellite derived SCA for a complex terrain of UIB.

4.3.4 Summary

The study is carried out to improve the real time streamflow estimation for a complex terrain of UIB where on ground observations are limited. Since, the most part of the watershed derives their flow from snow melt, so the satellite-derived SCA of the region could be used as a crucial input variable. In this paper, a case study of UIB is considered to improve the streamflow estimation models at BeshamQila through a fused data set, comprising of on ground flow observations and satellite derived SCA of three (03) sub-basins of UIB (Astore, Gilgit and Bunji).

The fusion process is carried out with the help of a novel mathematical tool, Gamma test, which provided the best combination 111111 (SCA of Gilgit, Q at Gilgit, SCA of Bunji, Q at Bunji, SCA at Astore, Q at Astore), with least value of MSE (0.000536).

The feed forward ANN models are trained via two layer BFGS algorithm with a variety of node combinations. The data length for training is optimized with the help of M-test in Win-Gamma environment and the least value of MSE, as determined by Gamma test, is utilized as an early stopping criteria to avoid over-fitting in ANN models. In this case, the best data length for training and testing of models comes out as 45% (165 weeks) and 55% (203 weeks), respectively.

The streamflow estimation models are also developed for input combination 010101, which only contains on ground flow observations. The both type of models are compared on the basis of NSE, BIAS, RMSE and VAR. The results indicate that the models developed with integrated data-set (combination: 111111) performed better with significant high values of NSE and low values of other statistical errors; including BIAS, RMSE and VAR.

4.4 ANN Models Developed Through Data Fusion

This section contains the results targeting the third and final objective of the research work that is carried out to improve ANN based streamflow estimation models by adopting different data fusion options. The section contains results for input combination selection, ANN model development, results comparison, a comprehensive discussion and a summary of the research work.

4.4.1 Gamma Test Results

The GT is performed and MSE value (or Gamma Statistics) is calculated for all the combinations which are; 1. Selected manually (data-fusion) based upon the type/nature of data and 2. Selected through the feature selection methods. The detail of data fusion options tried with respective Gamma value, V_{ratio} and data length for training is presented in Table 4.2.

TABLE 4.2: Gamma & V_{ratio} values along with optimized data length for different Data Fusion options

No.	Inputs	Combination / Mask	Inputs	Data (Normalized)		
				Gamma Values	V_{ratio}	%Data Training
1	P	11111111111100000000000000	12	0.054	0.626	76%
2	SR	00000000000011110000000000	4	0.04	0.459	71%
3	Q	00000000000000000000111111	6	0.004	0.047	63%
4	P + Q	11111111111100000001111111	18	0.002	0.024	38%
5	P + SR	11111111111111110000000000	16	0.023	0.271	71%
6	SCA	00000000000000000111000000	3	0.085	0.976	60%
7	SCA + Q*	00000000000000000111101100	6	0.072	0.827	76%
8	SCA + Q	00000000000000000111111111	9	0.004	0.041	57%
9	P+ S+Q	11111111111111110001111111	22	0.004	0.045	71%

No.	Inputs	Combination / Mask	Data (Normalized)			
			Inputs	Gamma Values	V_{ratio}	%Data Training
10	P+SCA+Q	11111111111100001111111111	21	0.002	0.022	65%
11	ALL	11111111111111111111111111	25	0.002	0.027	60%
Data Fusion through feature selection techniques						
12	Full Embedding	1101111101011011000000000	12	0.014	0.161	54%
13	Genetic Algorithm	1011010000000001111110111	13	2.2×10^{-6}	2.4×10^{-5}	54%
14	Hill Climbing	1111101111111101101111111	22	1.9×10^{-5}	2.2×10^{-4}	54%
15	Sequential Embedding	0010000010111001110111111	14	5.2×10^{-4}	5.9×10^{-3}	71%

The data length for training is optimized using an optimizing function called M-Test, which has already explained under section 3.3.3.

In WinGamma environment, the M-test does not provide any numerical cut-off value to decide the length of data that should be used for model training, rather it provides a graphical relationship between gamma value and increasing number of data points/observations. This graphical representation is then used to find that length of data at which the change in gamma value with respect to increasing number of data points, becomes minimum.

The few M-test outputs in form of graphs are presented in Fig. 4.22 for combination no. 2 (Only SR), Fig. 4.23 for combination no. 8 (SCA+Q), Fig. 4.24 for combination no.10 (P+SCA+Q), Fig. 4.25 for combination no. 12 (FE), Fig. 4.26 for combination no. 13(GA) and Fig. 4.27 for combination no. 15 (SE).

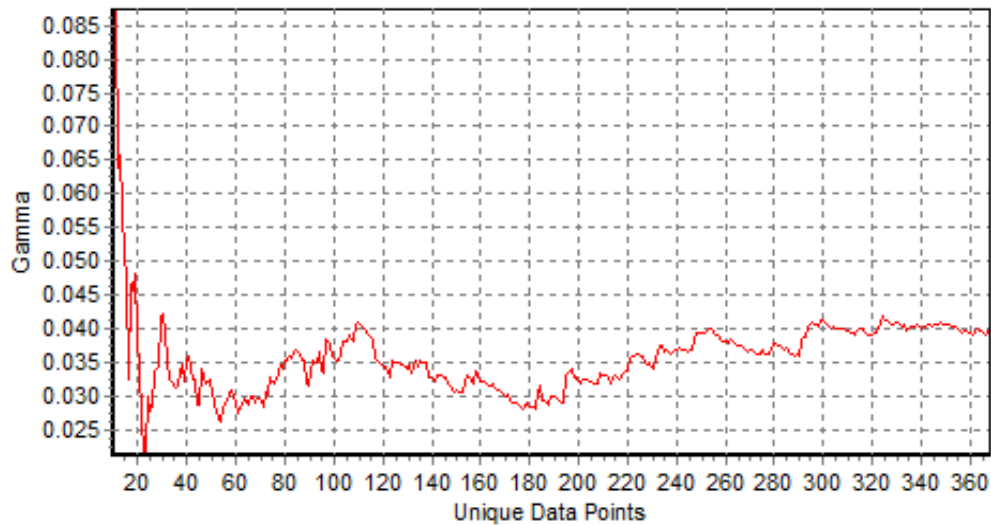


FIGURE 4.22: Stabilizing the Gamma Value with increasing data points for combination no. 2

It is clear from the Fig. 4.22 that the change in gamma value is not significant after 260 points and almost negligible after 300 points, so the data length for training should be considered in between 260 to 300 points. Therefore, the length for training is selected as 260 for training models, which contains inputs comprising of past data condition of solar radiation.

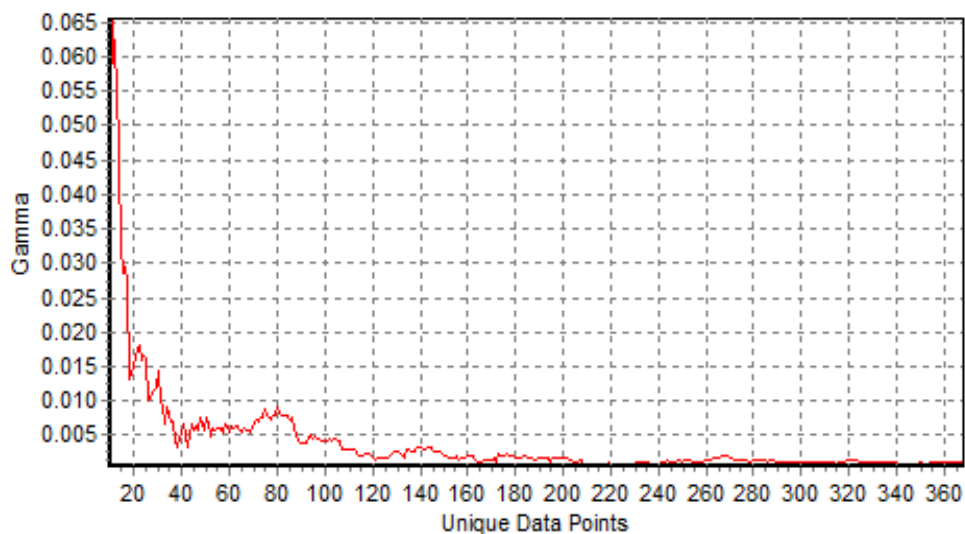


FIGURE 4.23: Stabilizing the Gamma Value with increasing data points for combination no. 8

For inputs containing snow cover area and discharges, the value of gamma error becomes stable at around 200, 210 points, after which the change in graph (Fig. 4.23) is negligible.

So, for this purpose the data length for training is considered as 210 points, which constitutes the 57% of the whole data length.

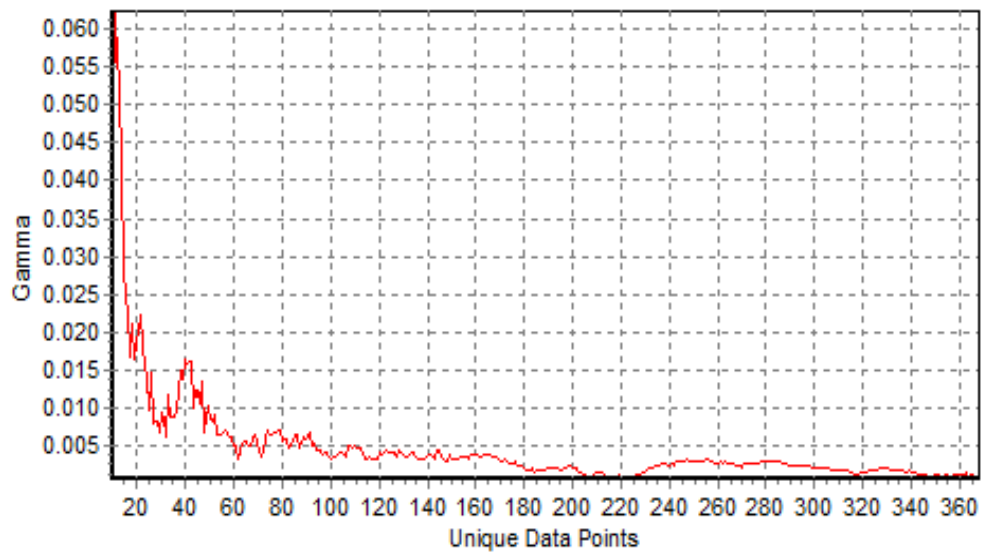


FIGURE 4.24: Stabilizing the Gamma Value with increasing data points for combination no. 10

Fig 4.24 represents the M-test conducted for increasing number of points for combination of inputs containing precipitation, snow cover area and discharge. It could be observed that the gamma value becomes stable after 240 points with no or little change in the gamma value with increasing number of data points. Therefore, the training data length for this combination is taken as 240 points.

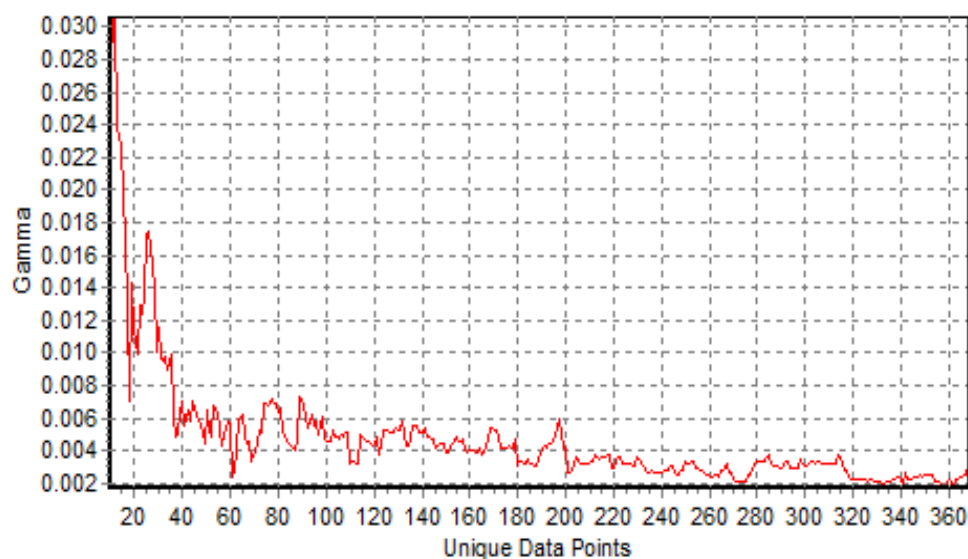


FIGURE 4.25: Stabilizing the Gamma Value with increasing data points for combination no. 12

It is clear from Fig. 4.25 that the graph between gamma values vs. unique data points becomes almost stable after 200 points and there is no sharp variation observed in the gamma value for further increase in data points. So, the length of data is selected as 200 out of total 368 weeks, which is utilized for training the models by using the input combination finalized through full embedding.

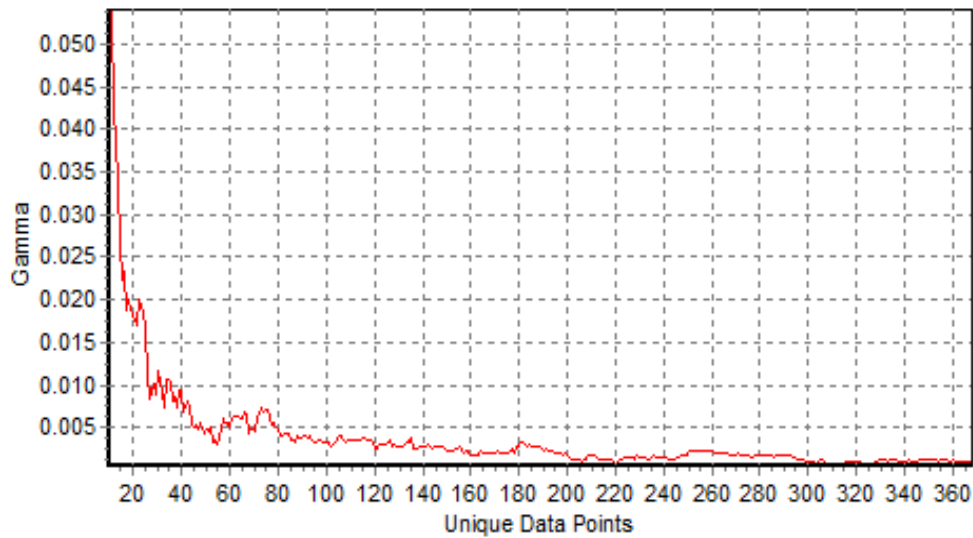


FIGURE 4.26: Stabilizing the Gamma Value with increasing data points for combination no. 13

Fig. 4.26 represents the gamma value variation with respect to increasing data points, when the M-test is conducted on the combination of inputs determined through genetic algorithm. The graph in this figure indicates a stable gamma value after 200 data points, which is The gamma value trend with increasing number of data points for input combination determined through sequential embedding is shown in Fig. 4.27. It is clear from the Fig. 4.27 that the change in gamma value after 260 points is not significant. Therefore, the optimum value of training data length is taken as 260 for this combination of inputs. The M-test results for other combination of inputs, which are not discussed here, are presented in the form of graphs in (Annex-4A).

It is clear from the results that the gamma value for almost all the combinations is close to zero. However, the minimum values are observed for the combinations made through feature selection methods, e.g. 2.2×10^{-6} , 1.9×10^{-5} and 5.2×10^{-4} for GA, SE and HC, respectively. Although, the gamma value close to zero is an indication that the noise among the data is less but alone this value should not be used as a criterion to screen the inputs, because the gamma test bears the assumption that the noise or non-smoothness

in data is only due to the statistical noise. Whereas, this is not true for all cases, e.g. when the outcome predicted is of a probabilistic nature.

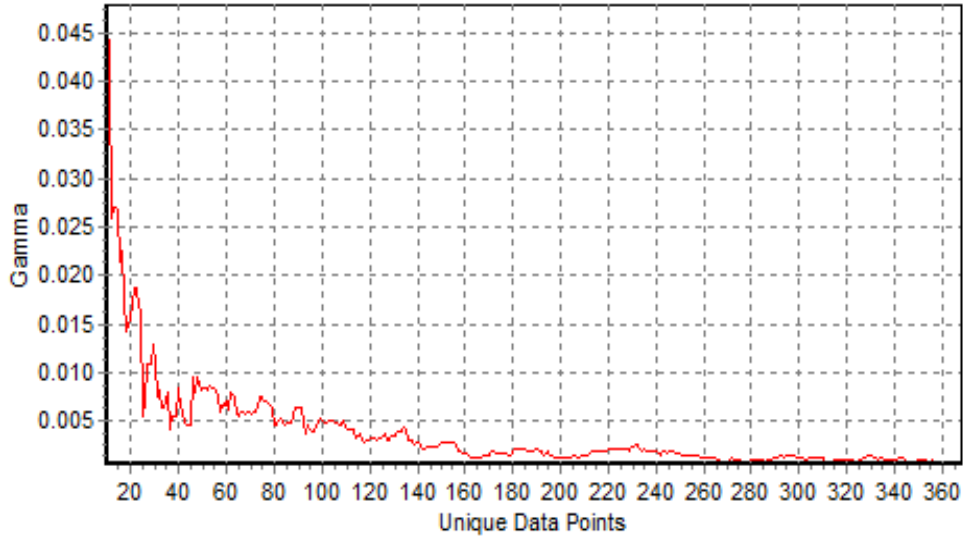


FIGURE 4.27: Stabilizing the Gamma Value with increasing data points for combination no. 15

To overcome this problem, V_{ratio} is also calculated in addition to gamma value. This value is basically a scale invariant noise which is used to standardize the Gamma value. It is the measure of how well a predictor could be modelled by a smooth function. The value of V_{ratio} close to zero means that the gamma value is the true reflection of the MSE present in the data, whereas the value closer to 1 means that the data is more like of a probabilistic nature. The results showed that the value of V_{ratio} is very less (close to 0) when we use; only discharge, precipitation + discharge, snow cover area + discharge, precipitation + solar radiation + discharge; precipitation + snow cover area + discharge; and all inputs together. The value of V_{ratio} is near to 1 for the input combinations; precipitation, solar radiation, precipitation + solar radiation, snow cover area + respective discharge. It shows that the gamma value for these set of input combinations is not reliable and could not be taken as the targeted MSE for the model development process, confidently.

4.4.2 ANN Model Results & Discussion

The targeted MSE calculated through gamma test is used to train ANN models via two layer BFGS algorithm. The models are trained using the optimum data length

determined through M-test. The models are tested on the remaining set of data length which is not utilized in the model training process. The performance of ANN based models for a variety of combinations is evaluated on the basis of a set of performance indicators as explained under section 3.3.6. The results of these models are displayed in the form of Box plots as presented in Figs. (4.28), (4.29), (4.30) and (4.31). (The detail results are presented in Annex-4B) The spread of performance indicators at a given combination shows the variation in the value of that particular indicator with respect to the different architecture of ANN models. i.e. The value of R^2 is calculated for a combination of inputs that contains only precipitation data (P), which is used to develop models with different node combinations in hidden layers.

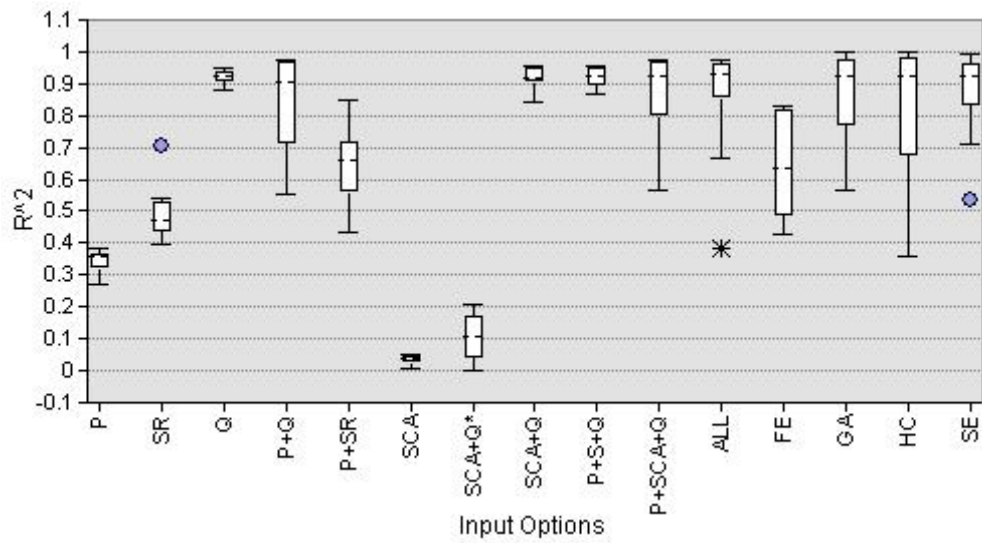


FIGURE 4.28: Variation of R^2 for different input combinations

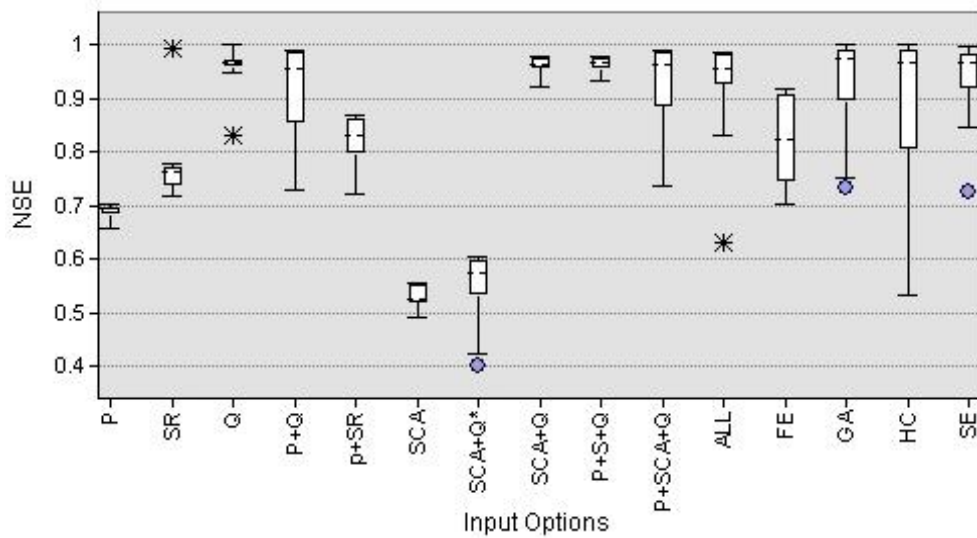


FIGURE 4.29: Variation of NSE for different input combinations

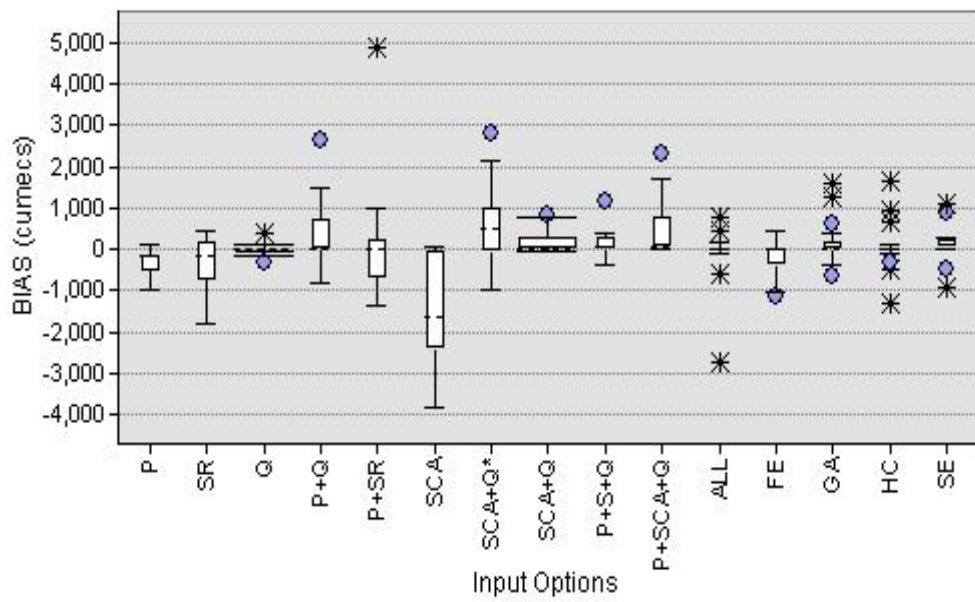


FIGURE 4.30: Variation of BIAS for different input combinations

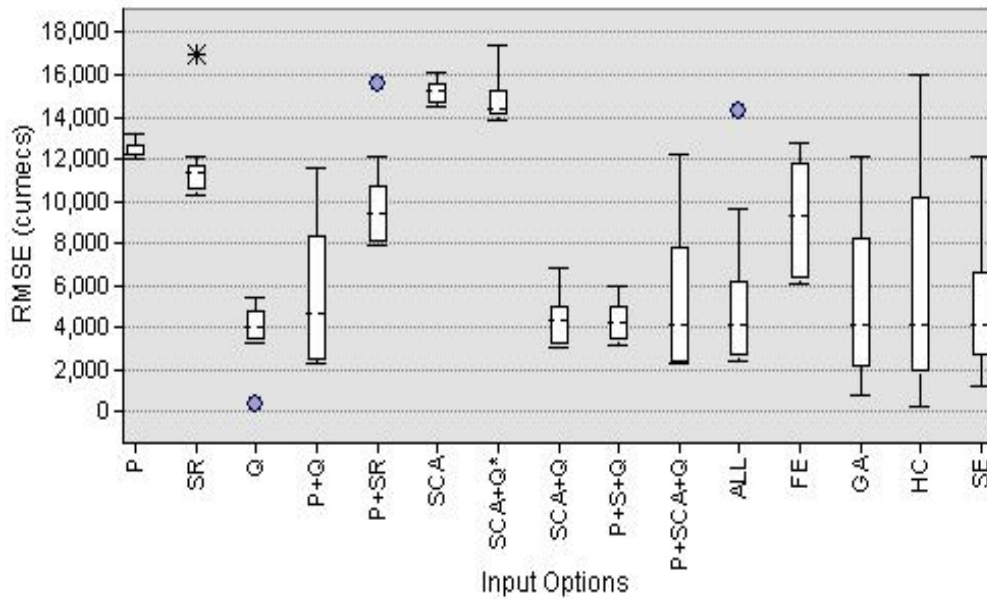


FIGURE 4.31: Variation of RMSE for different input combinations

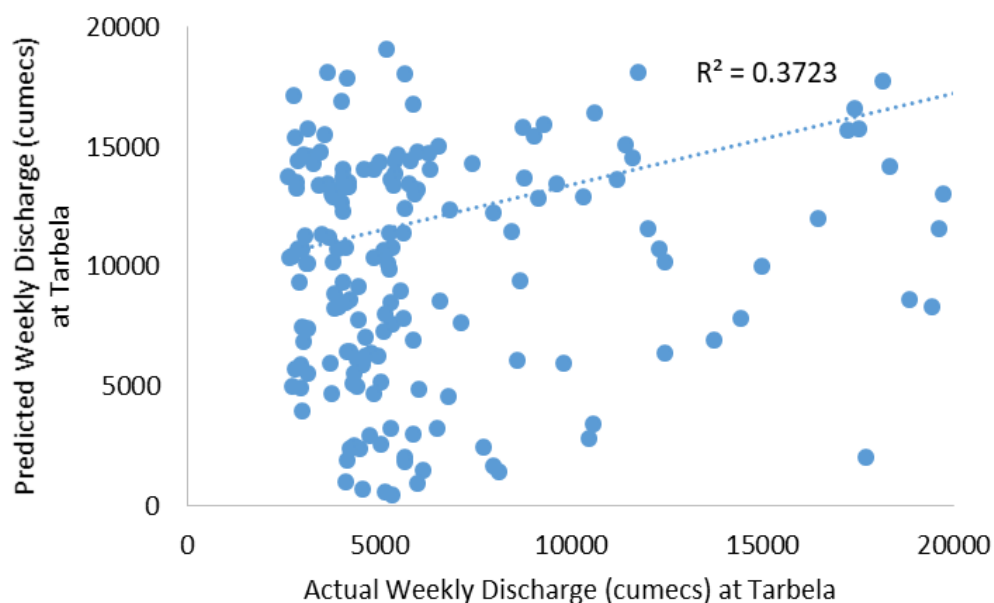
It is clear from Fig. 4.28 that the spread of R^2 using P, SR, SCA and SCA+Q* as inputs for different architectures of ANN models ranges from 0 to 0.5, which is not acceptable in terms of model efficiency. The spread of box-plot while using combinations Q, SCA, SCA+Q and P+S+Q is relatively small (Fig. 4.28 and Fig. 4.29) as compared to the other combination of inputs, which shows that these combination showed a little

variation with respect to changing ANN model structures in the values of R^2 and NSE as compared to the other input options.

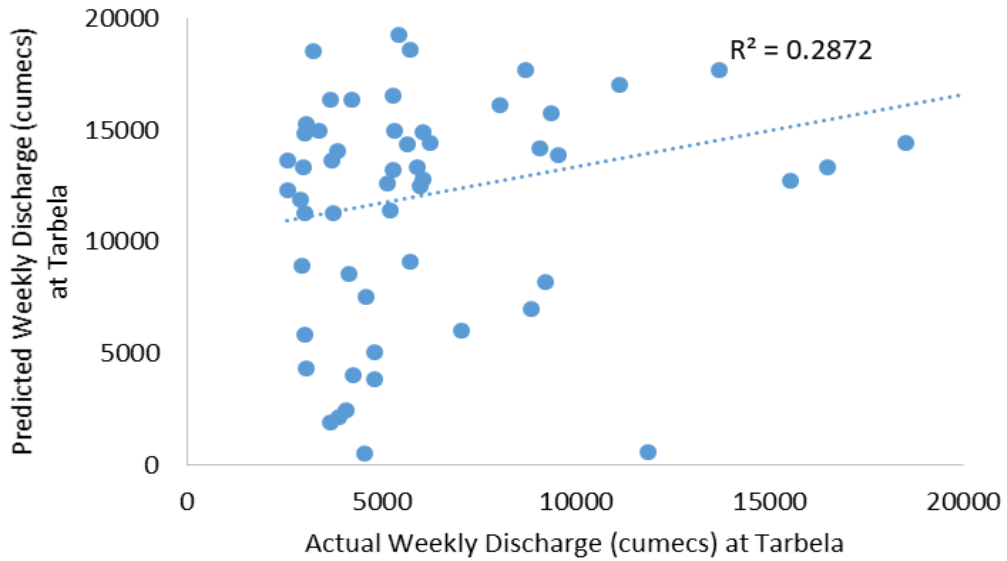
Similarly, most of the values for these indices lie within the range of first and third quartile. However, the combinations made through feature selection methods showed more variation for changing model structures with relatively large box plot and more spread specifically for first quartile. The similar trend of changing values has been observed in case of BIAS (Fig. 4.30) and RMSE (4.31) values for these combinations.

These show that the input combinations determined through feature selection techniques have more flexibility to change their output by changing the model structure or technique. Therefore, a careful selection of model structure is mandatory, otherwise the combinations made through feature section may perform even worse than the combinations made through any other ordinary method/s.

The low values of R^2 for these combinations means that the output (discharge at Tarbela) could not be modelled accurately using these set of input combinations. The week correlation between actual and predicted values can be observed in model, as presented in Figs. 4.32 (a) and 4.32 (b), developed using only P as input with node combination 1-3.



(a)

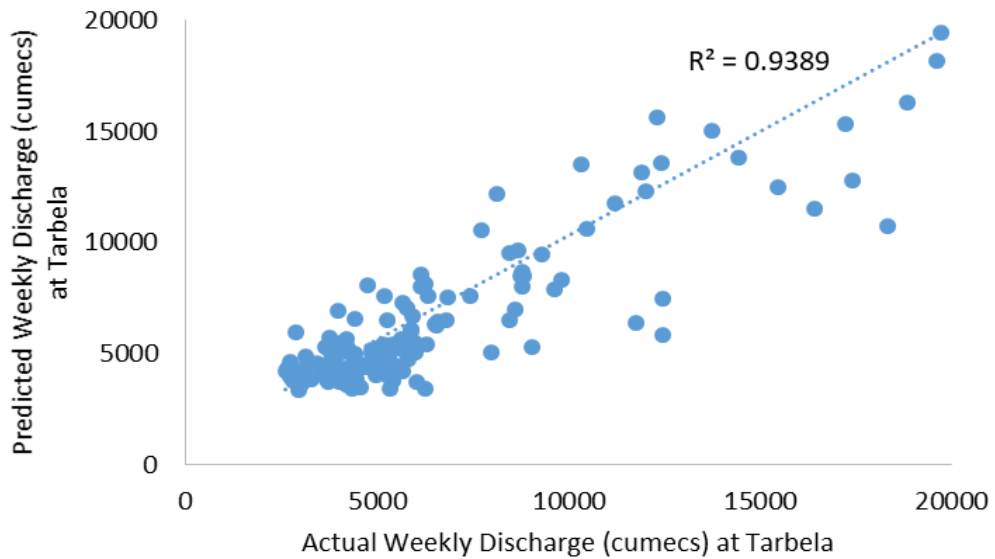


(b)

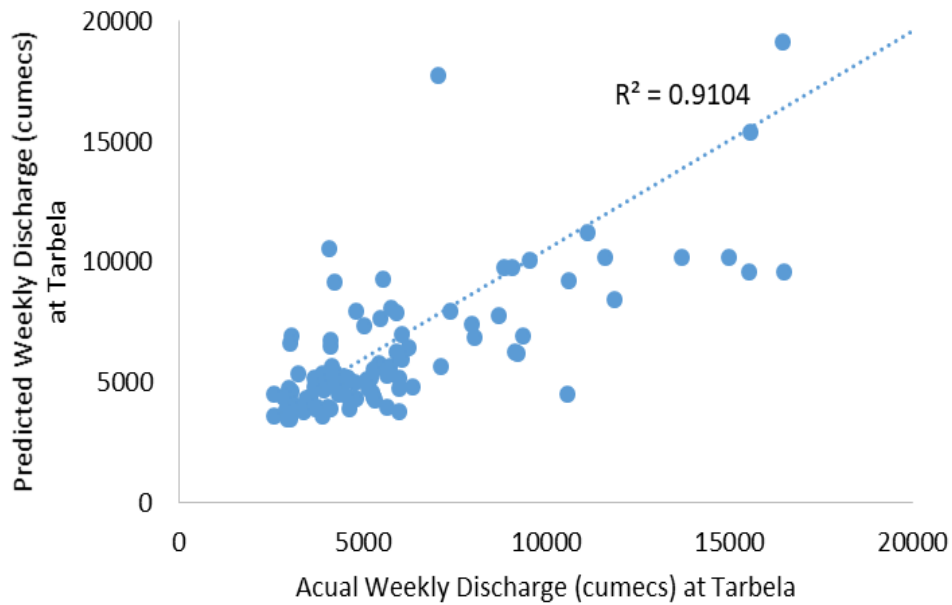
FIGURE 4.32: Model developed using only P (node combination 1-3).

(a) Training Model, (b) Testing Model

The one interesting output of this result is the better values of R^2 for SCA+Q as compared to the SCA+Q*. It means that the SCA of three sub-catchments (Gilgit, Astore and Bunji) along with their respective discharges (Q*) could only be used for the effective modeling of streamflow at downstream (Tarbela), when the discharges of other stations are also included in the inputs set. It is also noted that the values of R^2 are good for input combinations made through feature selection methods, especially for GA and SE. The model developed with combination determined through GA shows a good correlation between actual and modelled values (Fig. 4.33 (a) & 4.33 (b)).



(a)

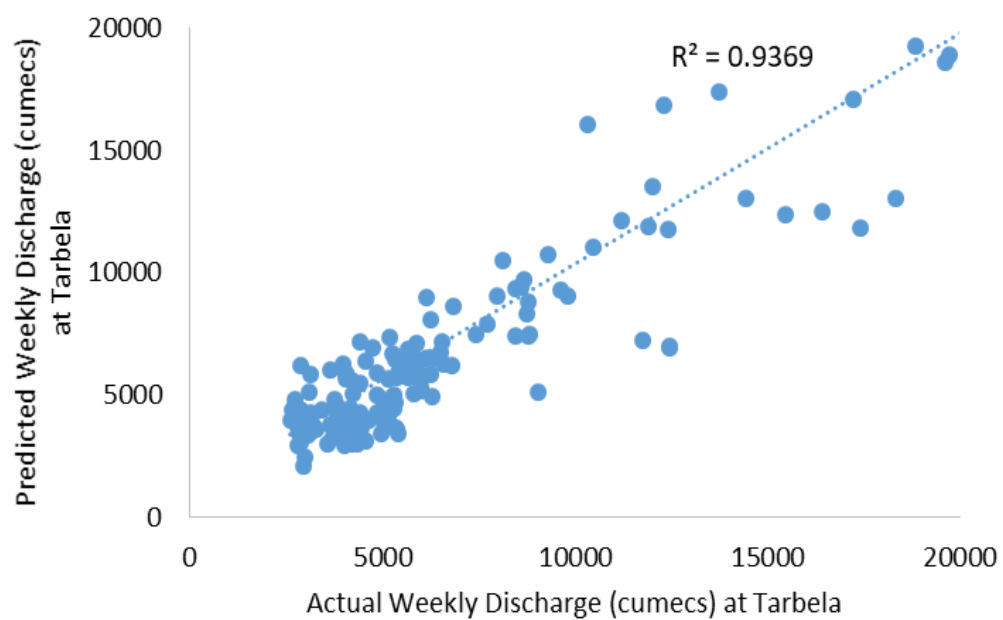


(b)

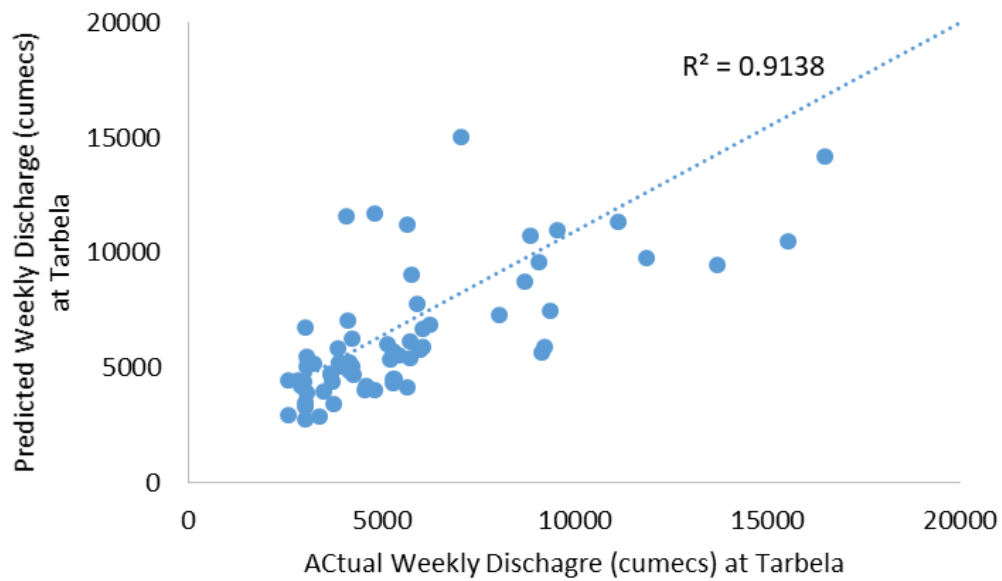
FIGURE 4.33: Model developed using combination determined through GA (with node combination 1-1).

(a) Training Model, (b) Testing Model

Similarly, the model developed using combination determined through SE also performed well with significant high values of R^2 in both training and testing phases as shown in Fig. 4.34 (a) and (b).



(a)

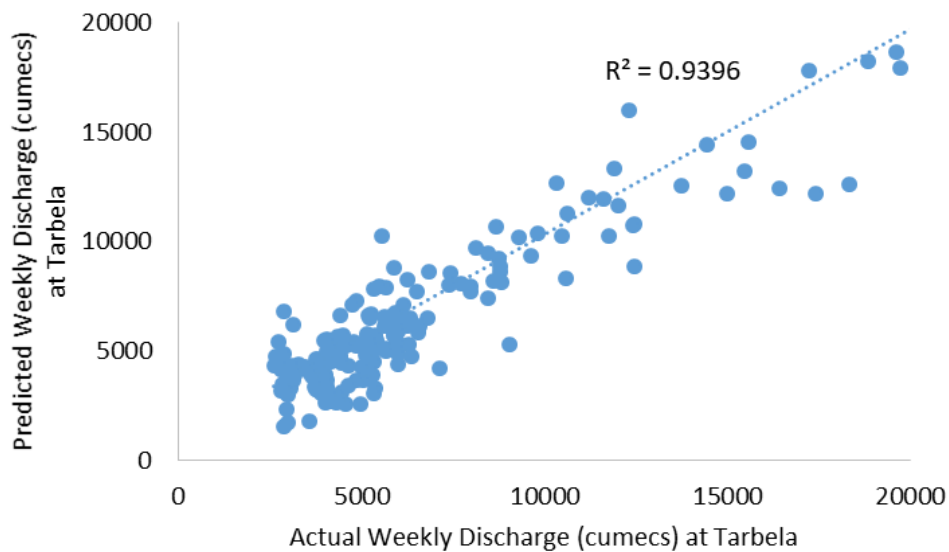


(b)

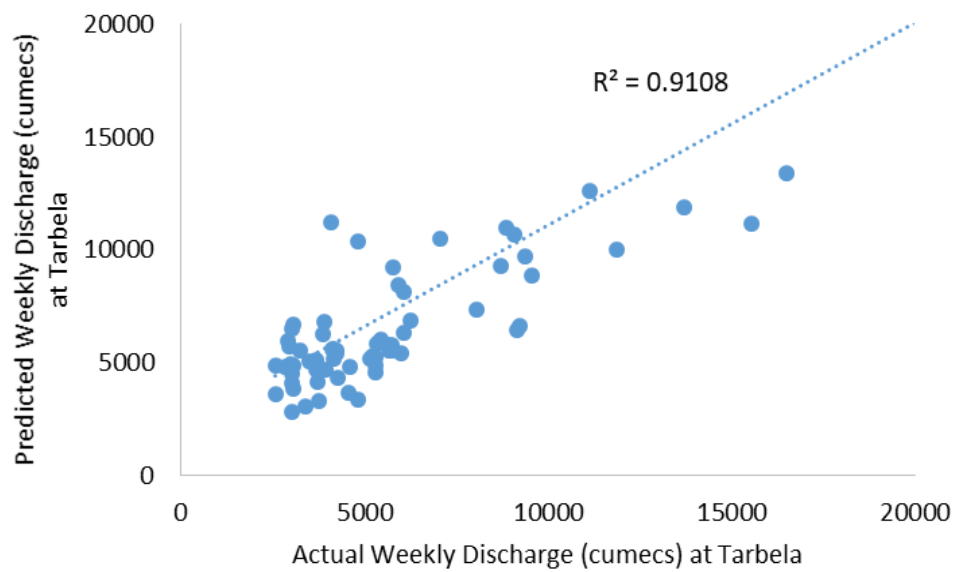
FIGURE 4.34: Model developed using combination determined through SE (with node combination 1-1).

(a) Training Model, (b) Testing Model

The models developed with the number of climate variables more than 3, i.e. in case of P+S+Q, P+SCA+Q and ALL, performed better as compared to the models developed with less number of input variables. The model with node combination 1-1, developed using P+S+Q is shown in Fig. 4.35 (a) and (b). The high value of R^2 in both phases is the proof that the models developed using combination containing these input variables are correlating well with the output.



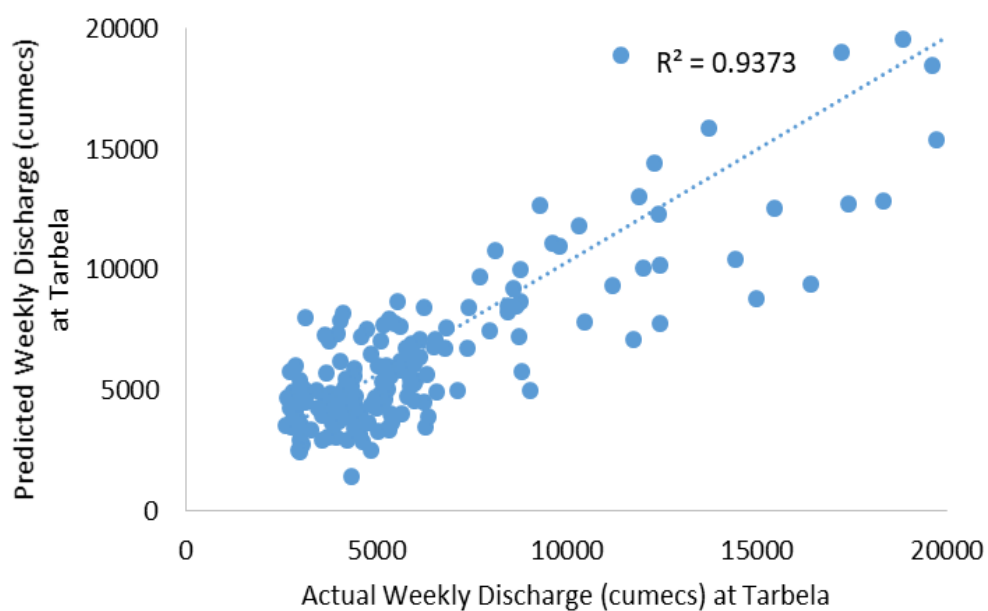
(a)



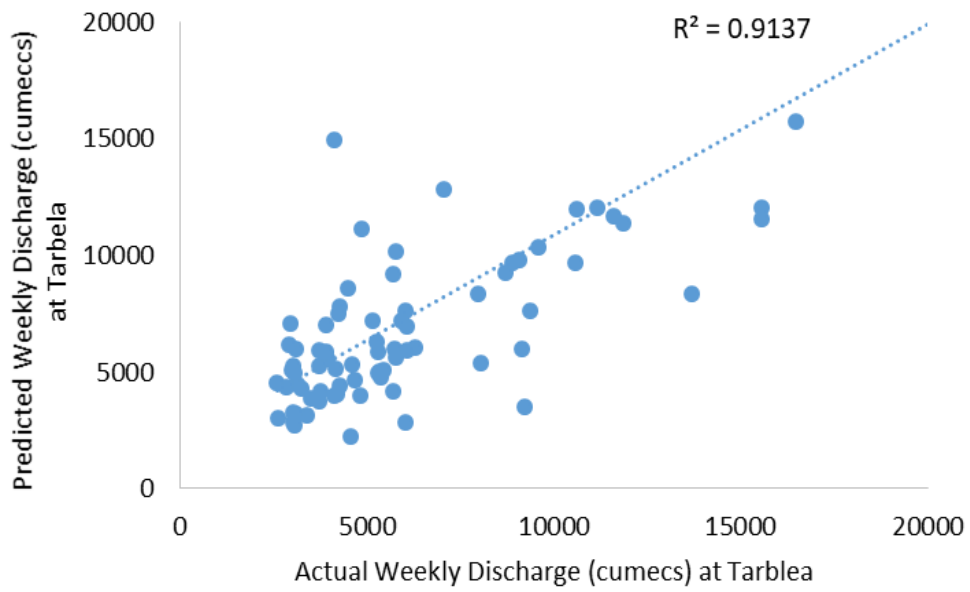
(b)

FIGURE 4.35: Model developed using P+S+Q (with node combination 1-1).
(a) Training Model, (b) Testing Model

The models developed using P+SCA+Q and ALL are presented in Fig. 4.36 (a) & (b) and Fig. 4.37 (a) & (b), respectively. The better correlation in these models is the depiction that more input variables provide a better picture of hydrological process in a catchment.



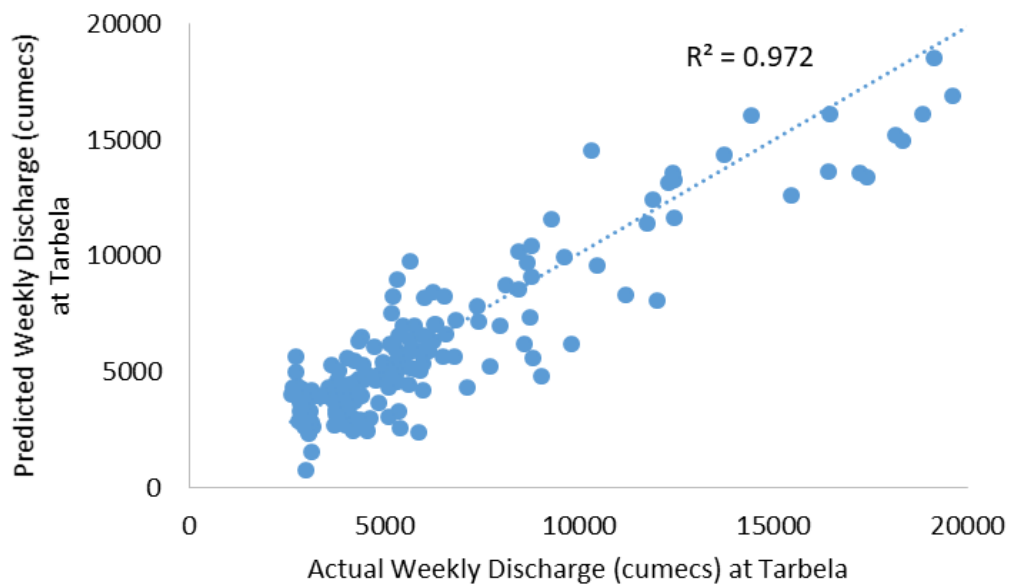
(a)



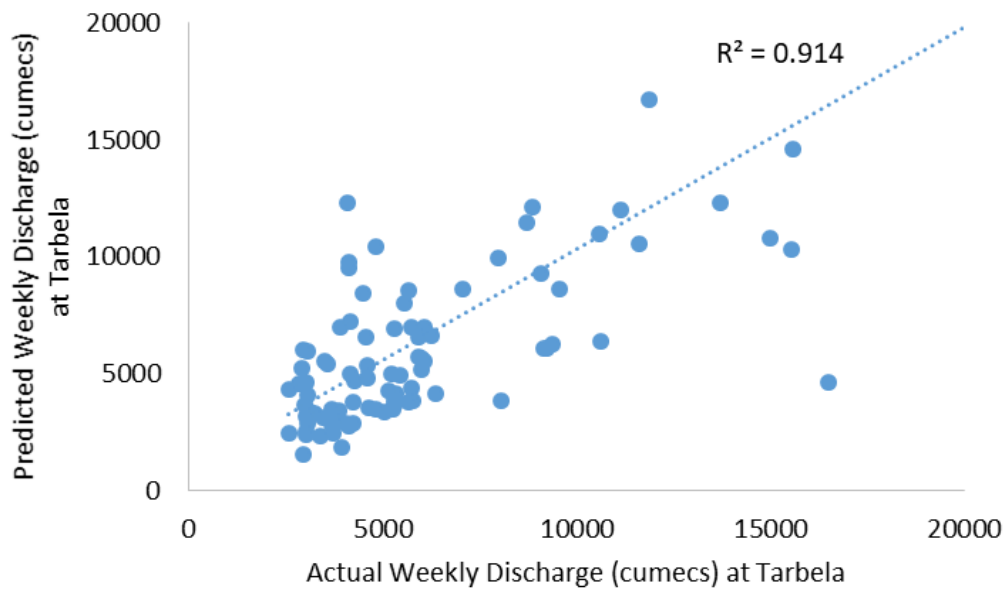
(b)

FIGURE 4.36: Model developed using P+SCA+Q (with node combination 1-1).

(a) Training Model, (b) Testing Model



(a)



(b)

FIGURE 4.37: Model developed using ALL input variables (with node combination 5-5).

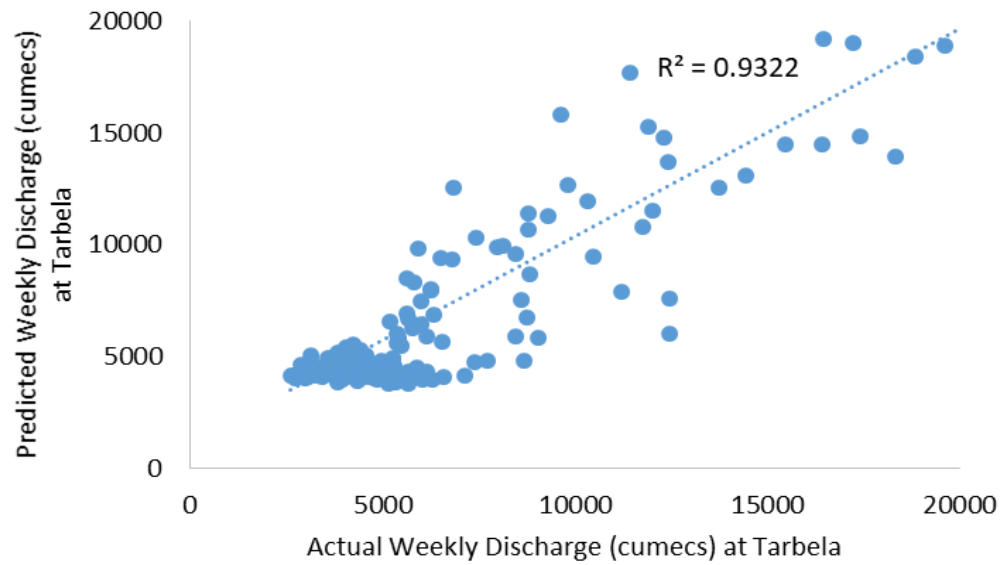
(a) Training Model, (b) Testing Model

A similar trend has been observed for NSE (Fig. 4.29) for developed models, except a slight high values for all input combinations as compared to R^2 .

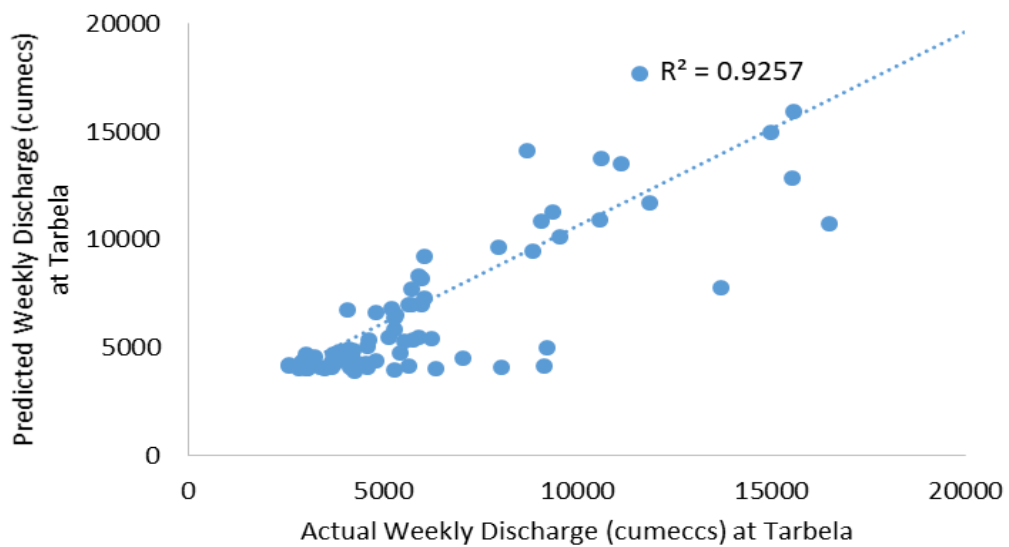
The same trend for R^2 and NSE is due to the reason that both the performance indicators measure that how well an output is correlated with the given input/s.

The spread of BIAS for developed models (Fig. 4.21), shows less values (near 0) for the input combinations made through feature selection methods, ALL, P+S+Q, SCA+Q and Q. it means that the difference in average values of modelled and observed discharges is less for models developed using these set of input combinations. A similar trend has been observed in the RMSE values for all the developed models (Fig. 4.31).

It is also noted from the Fig. (4.28), (4.29), (4.30) and (4.31) that the models developed with the input combination containing only discharges (Q) of different stations also performed well with high values of R^2 and NSE, and less values of BIAS and RMSE. A model with node combination, developed with Q is presented in Fig. 4.38.



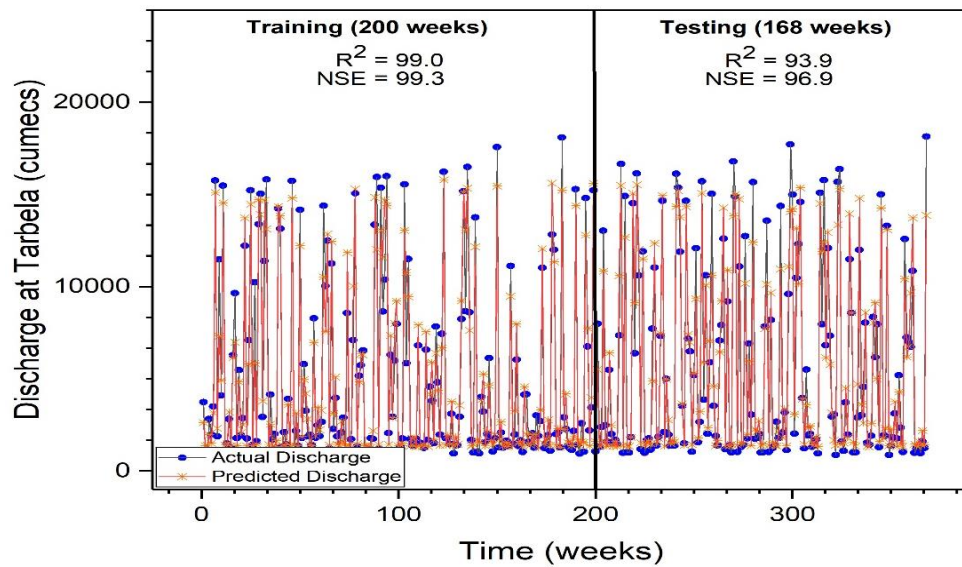
(a)



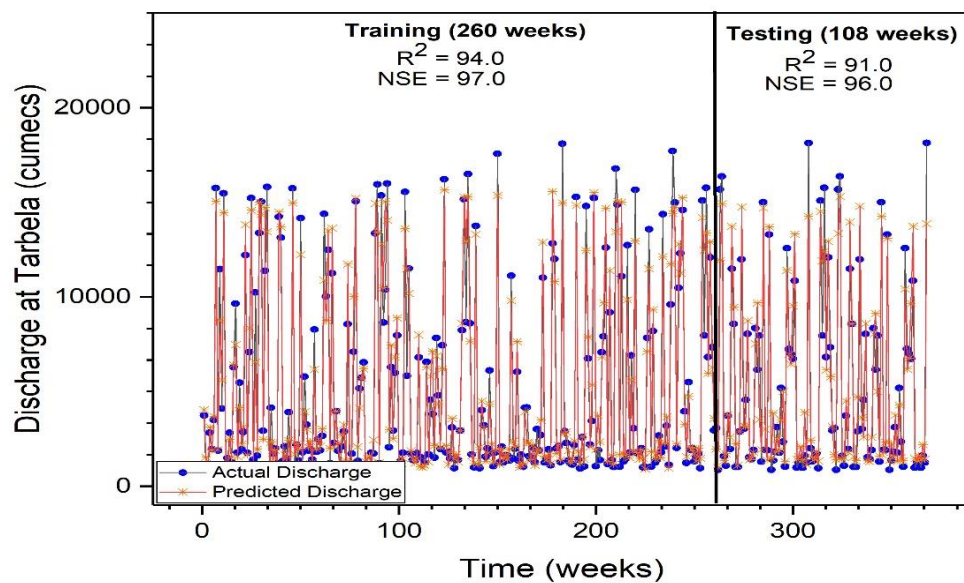
(b)

FIGURE 4.38: Model developed using Q (with node combination 2-2).
(a) Training Model, (b) Testing Model

The time series plots for are presented for the models developed using input variables that contain only discharge (Q) of upstream stations to model the stream flow at Tarbela in and the input variables determined through Sequential Embedding (SE) in Fig. 4.39 (a) and Fig. 4.39(b), respectively.



(a)



(b)

FIGURE 4.39: (a) Time-series plot for models developed using combination of input variables which contains only Q.

(b) Time-series plot for models developed using combination of input variables determined through SE

4.4.3 Summary

The study aims at the development of ANN based streamflow estimation models through a variety of data fusion options. The inputs to the models are basically the antecedent

data condition of UIB catchment. Four types of climate variables are considered including Precipitation (P), Discharge (Q), Solar Radiation (SR) and Snow Cover Area (SCA). A variety of input combinations are made and the impact of each combination is evaluated for our desired output (which is discharge at Tarbela) through a model development process.

The combination of inputs are made on the basis of type/nature of data or through advanced feature selection methods. The feature selection methods utilized in this study are Genetic Algorithm (GA), Hill Climbing (HC), Sequential Embedding (SE) and Full Embedding (FE). Gamma test is performed on all the combinations to assess the value of MSE, prior to model development process. This MSE or gamma value is considered as the targeted MSE for the model training. The data length for training is optimized through a mathematical function, M-test.

The ANN models are trained for all combinations via 2-layered BFGS algorithm. For fifteen (15) combination of inputs, ANN models are developed for nine (9) different architectures that sums to a total of 135 models. Testing phase of models involve checking the performance of developed models for unseen data through a set of performance indicators. The performance indicators used in this study are; Coefficient of determination (R^2), Nash Sutcliffe Efficiency (NSE), Root Mean Square Error (RMSE) and BIAS.

The results showed that; in general, the models developed with combinations containing more than 2 climate variables performed well and in particular, the models developed through input combinations made through feature selection methods outperformed. The results indicated that the upland catchment discharge (Q) is the only input variable that could be used to model the streamflow in the region with reasonable accuracy. However, the results showed that other variables in combination with Q enhances the accuracy of these models.

4.5 Overall Discussion

Provision of better input state to the hydrological models can significantly improve the performance efficiency of models because the hydrological data in raw form may contains undulations, errors and inconsistency problems. Although ANN dont have any explicit requirement of data normality and are considered capable of performing well

for nonlinear data but the improvement in their performance is noted through changing the shape and scale of data. However, the preprocessing options should be applied carefully as a mathematical transformation of physical process may create uncertainties as reported by [3], which identified the decline in efficiency of models due to the use of Wavelet Analysis (WA) and Empirical Mode Decomposition (EMD). These complex transformation are affected by the boundary conditions. Similarly, [204] reported deficiencies in mode mixing for EMD. On the other hand, despite of the many advantages of wavelet based methods, [205] identified adaptability issues in this method. In this case, the simple transformations have less complexities and can improve the hydrological data by reducing skewness and making data normal.

Similarly, the Gamma test proved its significance in estimating the variance of noise on an output before the start of the modeling process. This helped out the appropriate selection of inputs which are capable of modeling the output with more accuracy and less mean square error.

As compared to the conventional approaches used for the generalization, the Gamma test is superior in the context that the noise present in the data is already known and could be used to access the model performance, prior to the model building. It also reduces the need of separate validation data-set which is usually required for conventional early stopping methods to overcome the over-fitting in models. Previously the gamma test have been utilized by [108], [173], [180], to determine the best input combinations for the sediment load estimation models but the approach is limited as the gamma value is calculated for random combinations instead of checking all possible combinations, which is performed in this study with the help of advance feature selection methods.

In this study, BFGS is used to train ANN models which is essentially a gradient based function and these functions often stuck at local minima and may create over-fitting in models. But in this case we already know the value of minima (gamma value) through Gamma test. All the models are trained to achieve this value, and from the results, it is quite clear that all the achieved MSE values for Training are almost equal to the MSE values which are targeted. In Win Gamma, the training data is periodically shuffled to avoid repetitive cycles that may result in the algorithm getting stuck at local minima. Each vector is fed into the network and the error calculated between the expected output and the actual output and the weights in the network are adjusted accordingly. The

algorithm tests to see if the stopping criteria (Gamma Value) has been reached after each iteration.

The data-driven hydrological models are entirely dependent upon the input output data which are essentially the observations of climate variables. In case of modeling the catchment response, the selection of climate variables are very important as a catchment may observe contrasting regimes. Like in this case, the UIB observe contrasting hydrometeorospheric regimes. The most part of the flow is derived by the melting of snow and glaciers, so the snow cover area is considered as an important factor in capturing the catchment response in addition to other climate variables. Due to the complexity of the UIBs terrain and limited availability of the meteorological network, the satellite derived snow covered area is utilized. The dependency of stream flow of the UIB catchment on the satellite derived SCA is proved in the previous chapters. Similarly the Solar radiation is also considered as an important variable in estimating the catchment response, the input combinations containing solar radiation with other primary variables performed well as compared to the input combinations containing primary variables (P and Q), only. The comparison of model results developed with and without considering the solar radiation as an input variable is presented in Annex-4C.

Previously, [206] integrated ANN based decomposition models to predict stream flow in UIB with maximum NSE of 85% in validation phase. Similarly [207] developed Soil Water and Assessment Tool (SWAT) to capture the flow of UIB with reasonable accuracy (Maximum $R^2 = 85\%$). [46] developed models for UIB with $NSE > 90\%$ but the interval of estimation is high (monthly) for which the nonlinearity in the data is usually less as compared to the present research which developed models for weekly flow estimation models. The results of present study showed improvement in the UIB flow estimation models through coupled use of data preprocessing and data fusion with each of R^2 and $NSE > 95\%$.

Chapter 5

Conclusions & Recommendations

5.1 General

The results in chapter 4 revealed that the ANN based hydrological models has shown significant improvement for a complex catchment of UIB through data preprocessing, using satellite derived SCA, and applying different data fusion techniques. The current chapter presents the point wise conclusions made on the basis of all these results, adopted methodologies and literature background, which are discussed in detail in the previous chapters. This chapter also explains broader significance and implications of the research work along with the precincts of the adopted techniques and future recommendations.

5.2 Conclusions

1. The data-set transformed through Box-Cox creates a better input state, by reducing the noise through scaling of the data in a range that is more proportionate to the transfer function in the output layer of an ANN model. Thus, creating better learning maps with improved ANN training capacity and improved Model generalization.
2. Although, there is no specific requirement of data normality to be used for ANN, However the results indicate that the performance of models have significantly increased by using the data that has been transformed towards normality.

3. The Box-Cox Transformation provides an opportunity to change the data in a desired shape/ format by changing the power factor, which is more acceptable for model development process
4. The selection of inputs through the Gamma-Test acts as a part of preprocessing process and provides an opportunity to clean the data through excluding noisy inputs by providing an accurate estimate of variance of inputs on desired output.
5. The natural undulations in hydrological data needs to be preprocessed before applying hydrological models to improve the model's performance as confirmed by this research-work. The results showed the dependency of stream flow of Upper Indus Basin on the upland condition of snow cover area (SCA). It gives an evidence for the importance of satellite-derived SCA for a complex terrain of UIB.
6. This study concluded that the combined data-set comprising of SCA and gauge observations represents the watershed response better than the data-set only consisting of on-ground observations.
7. The present work also demonstrates the importance of conjunctive use of the Gamma Test and ANN to enhance the ability of ANN to perform well in the development of runoff models for mountainous catchments.
8. The higher value for Gamma value and V_{ratio} for (P), (SR), (SCA) and (SCA+Q*) is a clear indication that the gamma value is not a true depiction of statistical noise in the data. The same is reflected in the results.
9. Despite of the contrasting regimes of UIB, the discharge of upland catchment is well modelled to predict Q at Tarbela. Therefore, the antecedent flow condition of upland catchment could be used confidently for the stream flow estimation at Tarbela, as compared to any other climate variable.
10. The study finds that the use of catchment information containing only single climate variable (like P, SR or SCA) is unable to capture the response of the UIB at Tarbela, despite of choosing ANN based non-linear modeling option
11. The ANN models developed using multiple climate variables also performed well (Like; P+SCA+Q, P+SR+Q and P+SR+SCA+Q)

12. It is concluded that the use of multiple type/ sources of variables is beneficial to capture the response of complex catchment of UIB as compared to the single type/ source of information.
13. Selecting the less noisy observation through feature selection methods is found more advantageous as compared to integrating the all available information.
14. It is concluded that the feature selection techniques like GA, Hill Climbing and Sequential Embedding could be used successfully for the data fusion of hydrological time series with the correct measure of statistical noise (Gamma Value) in hydrological data.

5.3 Research Significance and Implications

Data fusion in hydrological forecasting provides an opportunity for the researchers to improve real time hydrological forecasting by incorporating different types, nature and sources of data. It was expected that the amalgamation of data provides a better picture of UIB catchment which not only have different behavioral phases but also observes contrasting regimes. The integrated data-set containing all variables together and/or in combination performed better as expected earlier that the multi-type and multi-source data could provide a better catchment response as compared to one type or single-source data.

This study evidenced that the fused data-set comprising of SCA and gauge observations represents the watershed response better than the data-set only consisting of on-ground observations. The present work also demonstrated the significance of conjunctive use of the Gamma Test and Artificial Neural Networking (ANN) approaches and the ability of ANN to perform well in the development of run-off models for mountainous catchments. Moreover, it is suggested that the uncertainty in the hydrological estimation models could be reduced by knowing more about the watershed. So, the multi-source/type information like climate and meteorological observations (e.g. temperature, solar radiations and rainfall, etc.), more satellite observations (e.g. gridded precipitation and LST) and multisensors data (different satellite products and airborne data) could be used and fused with the traditional observations to improve the real time flow forecasts.

The feature selection methods work on the principal that how well a given input or a set of inputs correlates with the given output. Therefore, these techniques select only those inputs among the all candidate inputs that corresponds well with the presence of less noise while modeling our desired output. Hence, it is suggested that besides the selection of climate variables on the basis of their type, nature or source, the noise present in the data plays a crucial role in input selection criteria. It is also established that the gamma test provides a good estimate of variance of noise on an output through gamma value and V_{ratio} . The models developed through the combination of inputs with V_{ratio} closer to 1, didn't perform well with less values of R2, NSE and high values of RMSE and BIAS. This clearly showed the importance of Gamma test in preliminary selection of input combination (data fusion).

The Box-Cox transformation provides an opportunity to transform the hydrological data through a family of power transformation. This family of power transformation contains other well-known transformations under its umbrella like square root transformation, cube root transformation, inverse transformation and log transformation. However, compared to these transformations, the results indicate that the data transformed through the power factor = 0.005, provides the best approximation to the normal distribution. Therefore, it is suggested to use Box-cox transformation as a preprocessing option as it offers an optimal solution for the researchers to transform their data in a shape that is more acceptable for model calibration process.

5.4 Precincts of Techniques used in Study

The study is carried out with the help of techniques, which are applied with some precincts/boundaries and are mentioned below:

1. The power factor (λ) for Box-Cox transformation could take infinite values and similarly results in infinite number of transformations. The current research uses only eleven (11) different values of λ through hit & trial and found 0.01 and 0005 as the best values. Further, the data transformed through 0.005 is only used for model development process.
2. The Gamma test is used to calculate the noise present among the data, prior to model development. This noise (gamma value) only describe the "statistical noise"

present among the data. The maximum input variables used for this research work are twenty five (25). For this, the possible combinations of inputs are $2^{25} - 1$ that require an extensive computational effort. Therefore, the combination are finalized either (on the basis of this gamma value) through a set of feature selection techniques or manual selection based upon the type/nature of data.

3. The nodes in hidden layers of ANN architecture have been selected through hit & trial and evaluated on the basis of set of performance indicators. The limited set of nodes are tried with two fixed hidden layers and ANN models are trained using only BFGS Algorithm. The results obtained through this specific set of conditions are generalized for overall ANN based streamflow estimation models.

5.5 Recommendations

1. The use of data preprocessing is recommended for hydrological model development, especially for streamflow estimation models. Besides Box-Cox transformation, the other simple transformations like Moving Average, Log Normalization, Chi square, etc. may also be tried and compared for a variety of nonlinear modeling options.
2. It is recommended to carry out the selection of climate variables for hydrological forecasting through advanced feature selection methods.
3. It is recommended to use satellite derived SCA as one of the possible input variable for complex and mountainous catchments where most part of the flow is derived by melting of snow and glaciers.
4. It is recommended to combine new observations like in-situ sensors data, airborne data and remotely sensed data with the traditional on ground observations to create a “more informed data state” for hydrological models.
5. For future possibilities, data-fusion could be used along with model-fusion to improve the hydrological forecasting.

Bibliography

- [1] R. Remesan, A. Ahmadi, M. A. Shamim, and D. Han, “Effect of data time interval on real-time flood forecasting,” no. 2004, pp. 396-408, 2010.
- [2] Wang, *Stochasticity, Nonlinearity and Forecasting of Streamflow Processes*. IOS Press, Amsterdam, 2006.
- [3] X. Zhang, Y. Peng, C. Zhang, and B. Wang, “Are hybrid models integrated with data preprocessing techniques suitable for monthly streamflow forecasting? Some experiment evidences,” *J. Hydrol.*, 2015.
- [4] C. Juan, W. Genxu, M. Tianxu, and S. Xiangyang, “ANN Model-Based Simulation of the Runoff Variation in Response to Climate Change on the Qinghai-Tibet Plateau, China,” *Adv. Meteorol.*, vol. 2017, pp. 1-13, 2017.
- [5] M. Hassan et al., “Predicting streamflows to a multipurpose reservoir using artificial neural networks and regression techniques,” *Earth Sci. Informatics*, vol. 8, no. 2, pp. 337-352, Jun. 2015.
- [6] T. Peng, J. Zhou, C. Zhang, and W. Fu, “Streamflow Forecasting Using Empirical Wavelet Transform and Artificial Neural Networks,” *Water*, vol. 9, no. 6, p. 406, Jun. 2017.
- [7] T. A. Awchi, “River Discharges Forecasting In Northern Iraq Using Different ANN Techniques,” *Water Resour. Manag.*, vol. 28, no. 3, pp. 801-814, Feb. 2014.
- [8] G. of Pakistan, “Agriculture Statistics: Pakistan Bureau of Statistics,” 2021. [Online]. Available: <https://www.pbs.gov.pk/content/agriculture-statistics>.
- [9] Imran Mukhtar, “With water scarce, Pakistan helps farmers grow more with less,” *Reuters*, 2020.

- [10] D. R. Archer, N. Forsythe, H. J. Fowler, and S. M. Shah, "Sustainability of water resources management in the Indus Basin under changing climatic and socio economic conditions," *Hydrol. Earth Syst. Sci.*, vol. 14, no. 8, pp. 1669-1680, Aug. 2010.
- [11] World Bank, "Pakistan Country Water Resources Assistance Strategy Water Economy: Running Dry," 2005.
- [12] WAPDA, *Handbook on water statistics of Pakistan*. 2010.
- [13] S. M.H., "Apportionment of Indus Water Accord," 2001.
- [14] A. Nadeem, "Indus Basin Water Management Challenges and Strategies," *Lahore Journal Econ.*, no. 15, pp. 1-25, 2010.
- [15] S. B. Cheema, M. Afzaal, T. Koike, and M. Rasmy, "Improvement of Applicability of Snow Hydrological Model by Introducing Snow Correction Factor in the Gilgit Basin," *Pakistan J. Meteorol.*, vol. 12, no. 24, pp. 95-106, 2016.
- [16] L. See and R. J. Abrahart, "Multi-model data fusion for hydrological forecasting," vol. 27, pp. 987-994, 2001.
- [17] M. B. Ercan and J. L. Goodall, "Estimating Watershed-Scale Precipitation by Combining Gauge- and Radar-Derived Observations," *J. Hydrol. Eng.*, vol. 18, no. 8, pp. 983-994, Aug. 2013.
- [18] L. Sun et al., "Investigating water use over the Choptank River Watershed using a multi-satellite data fusion approach," *Water Resour. Res.*, vol. 53, no. 7, 2017.
- [19] K. Nagarajan, C. Krekeler, S. Member, K. C. Slatton, S. Member, and W. D. Graham, "A Scalable Approach to Fusing Spatiotemporal Data to Estimate Streamflow via a Bayesian Network," vol. 48, no. 10, pp. 3720-3732, 2010.
- [20] R. J. Abrahart and L. See, "Multi-model data fusion for river flow forecasting: an evaluation of six alternative methods based on two contrasting catchments," *Hydrol. Earth Syst. Sci.*, vol. 6, no. 4, pp. 655-670, Aug. 2002.
- [21] H. J. Fowler and D. R. Archer, "Hydro-climatological variability in the Upper Indus Basin and implications for water resources," *IAHS Publ.*, no. April, pp. 131-138, 2005.

- [22] S. ul Hasson, J. Bhner, and V. et al. Lucarini, "Prevailing climatic trends and runoff response from Hindukush-Karakoram-Himalaya, upper Indus Basin," *Earth Syst. Dyn.*, vol. 8, no. 2, pp. 337-355, 2017.
- [23] W. W. Immerzeel, P. Droogers, S. M. De Jong, and M. F. P. et al. Bierkens, "Remote Sensing of Environment Large-scale monitoring of snow cover and runoff simulation in Himalayan river basins using remote sensing," *Remote Sens. Environ.*, vol. 113, no. 1, pp. 40-49, 2009.
- [24] M. Arfan, J. Lund, D. Hassan, M. Saleem, and A. Ahmad, "Assessment of Spatial and Temporal Flow Variability of the Indus River," *Resources*, vol. 8, no. 2, p. 103, May 2019.
- [25] H. Bilal, S. Chamhuri, M. Bin Mokhtar, and K. D. Kanniah, "Recent snow cover variation in the Upper Indus Basin of Gilgit Baltistan, Hindukush Karakoram Himalaya," *J. Mt. Sci.*, vol. 16, no. 2, pp. 296-308, 2019..
- [26] A. K. Prasad, K.-H. S. Yang, H. M. El-Askary, and M. Kafatos, "Melting of major Glaciers in the western Himalayas: evidence of climatic changes from long term MSU derived tropospheric temperature trend (1979-2008)," *Ann. Geophys.*, vol. 27, no. 12, pp. 4505-4519, Dec. 2009.
- [27] X. Liu and B. Chen, "Climatic Warming in the Tibetan Plateau during recent decades," *Int. J. Climatol.*, vol. 1742, pp. 1729-1742, 2000.
- [28] W. Wang, Y. Xiang, Y. Gao, A. Lu, and T. Yao, "Rapid expansion of glacial lakes caused by climate and glacier retreat in the Central Himalayas," *Hydrol. Process.*, vol. 29, no. 6, pp. 859-874, Mar. 2015.
- [29] B. Mukhopadhyay and A. Dutta, "A Stream Water Availability Model of Upper Indus Basin Based on a Topologic Model and Global Climatic Datasets," *Water Resour. Manag.*, vol. 24, no. 15, pp. 4403-4443, Dec. 2010.
- [30] A. A. Tahir, J. F. Adamowski, P. Chevallier, A. U. Haq, and S. Terzago, "Comparative assessment of spatiotemporal snow cover changes and hydrological behavior of the Gilgit, Astore and Hunza River basins (Hindukush-Karakoram-Himalaya region, Pakistan)," *Meteorol. Atmos. Phys.*, vol. 128, no. 6, pp. 793-811, Dec. 2016.

- [31] D. R. Archer and H. J. Fowler, "Spatial and temporal variations in precipitation in the Upper Indus Basin , global teleconnections and hydrological implications," *Hydrol. Earth Syst. Sci.*, vol. 8, no. 1, pp. 47-61, 2004.
- [32] S. Hasson, V. Lucarini, M. R. Khan, M. Petitta, T. Bolch, and G. Gioli, "Early 21st century snow cover state over the western river basins of the Indus River system," *Hydrol. Earth Syst. Sci.*, vol. 18, no. 10, pp. 4077-4100, Oct. 2014.
- [33] F. A. De Scally, "Relative importance of snow accumulation and monsoon rainfall data for estimating annual runoff, jhelum basin, pakistan," *Hydrol. Sci. J.*, vol. 39, no. 3, pp. 199-216, 1994.
- [34] N. Forsythe, C. G. Kilsby, H. J. Fowler, and D. R. Archer, "Assessment of Runoff Sensitivity in the Upper Indus Basin to Interannual Climate Variability and Potential Change Using MODIS Satellite Data Products," *Mt. Res. Dev.*, vol. 32, no. 1, p. 16, Feb. 2012.
- [35] K. Hewitt, "Glacier change, concentration, and elevation effects in the Karakoram Himalaya, upper indus basin," *Mt. Res. Dev.*, vol. 31, no. 3, pp. 188-200, 2011.
- [36] D. Archer, "Contrasting hydrological regimes in the upper Indus Basin," *J. Hydrol.*, vol. 274, no. 1-4, pp. 198-210, Apr. 2003.
- [37] D. R. Archer and H. J. Fowler, "Using meteorological data to forecast seasonal runoff on the River Jhelum, Pakistan," *J. Hydrol.*, vol. 361, no. 1-2, pp. 10-23, 2008.
- [38] N. Forsythe, H. J. Fowler, C. G. Kilsby, and D. R. Archer, "Opportunities from Remote Sensing for Supporting Water Resources Management in Village/Valley Scale Catchments in the Upper Indus Basin," *Water Resour. Manag.*, vol. 26, no. 4, pp. 845-871, Mar. 2012.
- [39] W. W. Immerzeel, N. Wanders, A. F. Lutz, J. M. Shea, and M. F. P. Bierkens, "Reconciling high-altitude precipitation in the upper Indus basin with glacier mass balances and runoff," *Hydrol. Earth Syst. Sci.*, vol. 19, no. 11, pp. 4673-4687, 2015.
- [40] A. F. Lutz, W. W. Immerzeel, P. D. A. Kraaijenbrink, A. B. Shrestha, and M. F. P. Bierkens, "Climate Change Impacts on the Upper Indus Hydrology: Sources, Shifts and Extremes," *PLoS One*, vol. 11, no. 11, pp. 1-33, Nov. 2016.

- [41] N. Forsythe et al., "Variability and Potential Change Using MODIS Satellite Data Products Assessment of Runoff Sensitivity in the Upper Indus Basin to Interannual Climate Variability and Potential Change Using MODIS Satellite Data Products," *BioOne*, vol. 32, no. 1, pp. 16-29, 2011.
- [42] A. Khan, B. S. Naz, and L. C. Bowling, "Separating snow, clean and debris covered ice in the Upper Indus Basin, Hindukush-Karakoram-Himalayas, using Landsat images between 1998 and 2002," *J. Hydrol.*, vol. 521, pp. 46-64, 2015.
- [43] M. P. Rao et al., "Six Centuries of Upper Indus Basin Streamflow Variability and Its Climatic Drivers," *Water Resour. Res.*, vol. 54, no. 8, pp. 5687-5701, Aug. 2018.
- [44] G. J. Young and K. Hewitt, "Hydrology research in the upper Indus basin , Karakoram Himalaya , Pakistan," *IAHS Publ.*, no. 190, pp. 139-152, 1990.
- [45] M. Yaseen, H. A. Bhatti, T. Rientjes, G. Nabi, and M. Latif, "Temporal and Spatial Variations in Summer Flows of Upper Indus Basin, Pakistan," in *72th Annual Session of Pakistan Engineering Congress*, 2013, no. 747, pp. 315-334.
- [46] H. Hayat, T. A. Akbar, A. A. Tahir, Q. K. Hassan, A. Dewan, and M. Irshad, "Simulating Current and Future River-Flows in the Snowmelt-Runoff Model and RCP Scenarios," *Water*, vol. 11, no. 4, pp. 1-19, 2019.
- [47] S. P. Charles et al., "Seasonal streamflow forecasting in the upper Indus Basin of Pakistan: an assessment of methods," *Hydrol. Earth Syst. Sci.*, vol. 22, no. 6, pp. 3533-3549, Jun. 2018.
- [48] D. Bocchiola, G. Diolaiuti, A. Soncini, C. Mihalcea, C. D. Agata, and C. Mayer, "Prediction of future hydrological regimes in poorly gauged high altitude basins: the case study of the upper Indus , Pakistan," *Hydrol. Earth Syst. Sci. Discuss.*, vol. 8, pp. 3743-3791, 2011.
- [49] K. Garee, X. Chen, A. Bao, Y. Wang, and F. Meng, "Hydrological Modeling of the Upper Indus Basin: A Case Study from a High-Altitude Glacierized Catchment Hunza," *Water*, vol. 9, no. 1, p. 17, Jan. 2017.
- [50] M. T. Afzal, M. Arslan, S. Zafar, and M. M. Waqar, "Satellite Derived Snow Cover Status and Trends in the Indus Basin," *J. Sp. Technol.*, vol. 4, no. 1, pp. 26-31, 2014.

- [51] F. Bashir and G. Rasul, "Estimation of Water Discharge from Gilgit Basin using Remote Sensing , GIS and Runoff Modeling Introduction:," *Pakistan J. Meteorol.*, vol. 6, no. 12, pp. 97-113, 2003.
- [52] M. Tayyab, J. Zhou, R. Adnan, and X. Zeng, "Application of Artificial Intelligence Method Coupled with Discrete Wavelet Transform Method," *Procedia Comput. Sci.*, vol. 107, pp. 212-217, 2017.
- [53] J. Liu, S. Kang, K. Hewitt, L. Hu, and L. Xianyu, "Large observational bias on discharge in the Indus River since 1970s," *Sci. Rep.*, vol. 8, no. 1, p. 17291, Dec. 2018.
- [54] Q. Xia, X. Gao, W. Chu, and S. Sorooshian, "Estimation of daily cloudfree, snow-covered areas from MODIS based on variational interpolation," *Water Resour. Res.*, vol. 48, no. 9, pp. 1-9, Sep. 2012.
- [55] R. Remesan, M. A. Shamim, and D. Han, "Model data selection using gamma test for daily solar radiation estimation," *Hydrol. Process.*, vol. 22, no. 21, pp. 4301-4309, Oct. 2008.
- [56] S. Galelli and A. Castelletti, "Tree-based iterative input variable selection for hydrological modeling," *Water Resour. Res.*, vol. 49, no. 7, pp. 4295-4310, 2013.
- [57] V. Moya Quiroga, A. Mano, Y. Asaoka, S. Kure, K. Udo, and J. Mendoza, "Snow glacier melt estimation in tropical Andean glaciers using artificial neural networks," *Hydrol. Earth Syst. Sci.*, vol. 17, no. 4, pp. 1265-1280, Apr. 2013.
- [58] S. Anwar, B. Yu, and G. Nabi, "Application of weather generator for environmental parameters estimation for upper indus basin," *Soil Environ.*, vol. 31, no. 1, pp. 11-20, 2012.
- [59] S. Singh, S. Jain, and A. Brdossy, "Training of Artificial Neural Networks Using Information-Rich Data," *Hydrology*, vol. 1, no. 1, pp. 40-62, Jul. 2014.
- [60] L. Zhang et al., "Comparison of SWAT and DLBRM for Hydrological Modeling of a Mountainous Watershed in Arid Northwest China," *J. Hydrol. Eng.*, vol. 21, no. 5, pp. 04016007-1-11, 2016.

- [61] E. Wang, H. Zheng, F. Chiew, Q. Shao, and J. Luo, "Monthly and seasonal streamflow forecasts using rainfall-runoff modeling and POAMA predictions," in *19th International Congress on Modelling and Simulation*, 2011, pp. 12-16.
- [62] K. L. Brubaker, R. T. Pinker, and E. et al. Deviatova, "Evaluation and Comparison of MODIS and IMS Snow-Cover Estimates for the Continental United States Using Station Data," *J. Hydrometeorol.*, vol. 6, pp. 1002-1017, 2005.
- [63] I. Snmez, A. E. Tekeli, and E. Erdi, "Snow cover trend analysis using Interactive Multisensor Snow and Ice Mapping System data over Turkey," *Int. J. Climatol.*, vol. 34, no. 7, pp. 2349-2361, Jun. 2014.
- [64] P. Rastner et al., "On the Automated Mapping of Snow Cover on Glaciers and Calculation of Snow Line Altitudes from Multi-Temporal Landsat Data," *Remote Sens.*, vol. 11, no. 12, p. 1410, Jun. 2019.
- [65] A. Sakai, T. Nuimura, K. Fujita, S. Takenaka, H. Nagai, and D. Lamsal, "Climate regime of Asian glaciers revealed by GAMDAM glacier inventory," *Cryosph.*, vol. 9, no. 3, pp. 865-880, May 2015.
- [66] C. Andermann, S. Bonnet, and R. Gloaguen, "Evaluation of precipitation data sets along the Himalayan front," *Geochemistry, Geophys. Geosystems*, vol. 12, no. 7, pp. 1-16, Jul. 2011.
- [67] U. Naeem, H.-U.-R. Mughal, A. R. Ghumman, and M. A. Shamim, "Ranking Sensitive Calibrating Parameters of UBC Watershed Model Ranking Sensitive Calibrating Parameters of UBC Watershed Model," *KSCE J. Civ. Eng.*, vol. 19, pp. 1538-1547, 2015.
- [68] J.-F. Exbrayat, N. R. Viney, J. Seibert, H.-G. Frede, and L. Breuer, "Multi-model data fusion as a tool for PUB: example in a Swedish mesoscale catchment," *Adv. Geosci.*, vol. 29, pp. 43-50, Feb. 2011.
- [69] S. Wi, Y. C. E. Yang, S. Steinschneider, A. Khalil, and C. M. Brown, "Calibration approaches for distributed hydrologic models in poorly gaged basins: implication for streamflow projections under climate change," *Hydrol. Earth Syst. Sci.*, vol. 19, no. 2, pp. 857-876, Feb. 2015.

- [70] T. Razavi and P. Coulibaly, "Improving streamflow estimation in ungauged basins using a multi-modelling approach," *Hydrol. Sci. J.*, vol. 61, no. 15, pp. 2668-2679, Nov. 2016.
- [71] C. Augusto, G. Santos, and G. Barbosa, "Daily streamflow forecasting using a wavelet transform and artificial neural network hybrid models," *Hydrol. Sci. J.*, vol. 59, no. 2, pp. 312-324, 2014.
- [72] R. S. Govindaraju, "By the ASCE Task Committee on Application of Artificial Neural Networks in Hydrology 1," *J. Hydrol. Eng.*, vol. 5, no. 2, pp. 115-123, 2000.
- [73] R. S. Govindaraju, "Artificial Neural Networks in Hydrology. II: Hydrologic Applications," *J. Hydrol. Eng.*, vol. 5, no. 2, pp. 124-137, 2000.
- [74] O. Kisi and H. Sanikhani, "Prediction of long-term monthly precipitation using several," *Int. J. Climatol.*, vol. 35, no. 14, pp. 4139-4150, 2015.
- [75] G. Uysal and A. Arda, "Streamflow forecasting using different neural network models with satellite data for a snow dominated region in turkey," *Procedia Eng.*, vol. 154, pp. 1185-1192, 2016.
- [76] G. B. Humphrey, M. S. Gibbs, G. C. Dandy, and H. R. Maier, "A hybrid approach to monthly streamflow forecasting: Integrating hydrological model outputs into a Bayesian artificial neural network," *J. Hydrol.*, vol. 540, pp. 623-640, Sep. 2016.
- [77] X. Zhu and X. Wu, "Class Noise vs . Attribute Noise: A Quantitative Study of Their Impacts," *Artif. Intell. Rev.*, vol. 22, no. 1, pp. 177-210, 2004.
- [78] F. Ahmed, M. Hassan, and H. N. Hashmi, "Developing nonlinear models for sediment load estimation in an irrigation canal," *Acta Geophys.*, vol. 66, no. 6, pp. 1485-1494, Dec. 2018.
- [79] G. Uysal and A. . orman, "Monthly streamflow estimation using wavelet-artificial neural network model: A case study on amldere dam basin, Turkey," *Procedia Comput. Sci.*, vol. 120, pp. 237-244, 2017.
- [80] A. P. Piotrowski and J. J. Napiorkowski, "A comparison of methods to avoid overfitting in neural networks training in the case of catchment runoff modelling," *J. Hydrol.*, vol. 476, pp. 97-111, 2013.

-
- [81] A. N. Tikhnov, "Solution of incorrectly formulated problems and the regularization method," *Sov. Math Dokl*, 1963.
- [82] K. J. Lang, A. H. Waibel, and G. E. Hinton, "A time-delay neural network architecture for isolated word recognition," *Neural Networks*, vol. 3, no. 1, pp. 23-43, 1990.
- [83] Y. Le Cun, J. S. Denker, and S. A. Solla, "Optimal Brain Damage," *Adv. Neural Inf. Process. Syst.*, pp. 598-605, 1990.
- [84] O. Giustolisi, "Sparse solution in training artificial neural networks," *Neurocomputing*, vol. 56, no. 1-4, pp. 285-304, 2004.
- [85] M. Khan, N. Muhammad, and A. El-Shafie, "Wavelet-ANN versus ANN-Based Model for Hydrometeorological Drought Forecasting," *Water*, vol. 10, no. 8, p. 998, Jul. 2018.
- [86] A. Famili, W. Shen, R. Weber, and E. Simoudis, "Data Preprocessing and Intelligent Data Analysis," *Intell. Data Anal.*, vol. 1, no. 1-4, pp. 3-23, 1997.
- [87] N. M. Nawi, W. H. Atomi, and M. Z. et al Rehman, "The optimization of Neural Networks through data preprocessing techniques," *Procedia Technol.*, vol. 11, no. Iceei, pp. 32-39, 2013.
- [88] B. Cannas, A. Fanni, G. Sias, S. Tronci, and M. K. Zedda, "River flow forecasting using neural networks and wavelet analysis," *Geophys. Res. Abstr.*, vol. 7, 2005.
- [89] R. J. Abrahart et al., "Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting," *Prog. Phys. Geogr. Earth Environ.*, vol. 36, no. 4, pp. 480-513, Aug. 2012.
- [90] N. Mohd, W. Hasen, and M. Z. Rehman, "The Effect of Data Pre-Processing on Optimized Training of Artificial Neural Networks," *Procedia Technol.*, vol. 8, pp. 33-40, 2013.
- [91] K. Kuniar and M. Zajc, "Some methods of pre-processing input data for neural networks," *Comput. Assist. Methods Eng. Sci.*, vol. 22, no. 2, pp. 141-151, 2015.
- [92] E. K. Haimoudi and L. Cherrat, "Practical Application of the Data Preprocessing Method for Kohonen Neural Networks in Pattern Recognition Tasks," in *The*

- Sixth International Conference on Advances in Information Mining and Management*, 2016, no. 1, pp. 38-44.
- [93] H. Ba, S. Guo, Y. Wang, X. Hong, Y. Zhong, and Z. Liu, "Improving ANN model performance in runoff forecasting by adding soil moisture input and using data preprocessing techniques," *Hydrol. Res.*, vol. 49, no. 3, pp. 744-760, Jun. 2018.
- [94] X. Zheng, M. Wang, and J. Ordieres-Mer, "Comparison of Data Preprocessing Approaches for Applying Deep Learning to Human Activity Recognition in the Context of Industry 4.0," *Sensors*, vol. 18, no. 7, p. 2146, Jul. 2018.
- [95] H. A. Afan et al., "Input attributes optimization using the feasibility of genetic nature inspired algorithm: Application of river flow forecasting," *Sci. Rep.*, vol. 10, no. 1, pp. 1-15, 2020.
- [96] L. Diop et al., "The influence of climatic inputs on stream-flow pattern forecasting: case study of Upper Senegal River," *Environ. Earth Sci.*, vol. 77, no. 5, 2018.
- [97] C. L. Wu, K. W. Chau, and C. Fan, "Prediction of rainfall time series using modular artificial neural networks coupled with data-preprocessing techniques," *J. Hydrol.*, vol. 389, no. 1-2, pp. 146-167, Jul. 2010.
- [98] T. Zhou, F. Wang, and Z. Yang, "Comparative Analysis of ANN and SVM Models Combined with Wavelet Preprocess for Groundwater Depth Prediction," *Water*, vol. 9, no. 10, p. 781, Oct. 2017.
- [99] T. Partal and . Kii, "Wavelet and neuro-fuzzy conjunction model for precipitation forecasting," *J. Hydrol.*, vol. 342, no. 1-2, pp. 199-212, Aug. 2007.
- [100] V. Nourani, A. Hosseini Baghanam, J. Adamowski, and O. Kisi, "Applications of hybrid wavelet-Artificial Intelligence models in hydrology: A review," *J. Hydrol.*, vol. 514, pp. 358-377, Jun. 2014.
- [101] G. Napolitano, F. Serinaldi, and L. See, "Impact of EMD decomposition and random initialisation of weights in ANN hindcasting of daily stream flow series: An empirical examination," *J. Hydrol.*, vol. 406, no. 3-4, pp. 199-214, 2011.
- [102] L. Karthikeyan and D. N. Kumar, "Predictability of nonstationary time series using wavelet and EMD based ARMA models," *J. Hydrol. Elsevier*, vol. 502, pp. 103-119, 2013.

- [103] P. Addison, *The illustrated wavelet transform handbook*. London: Institute of Physics Publishing, 2002.
- [104] T. Xiong, Y. Bao, and Z. Hu, "Neurocomputing Does restraining end effect matter in EMD-based modeling framework for time series prediction? Some experimental evidences," *Neurocomputing*, vol. 123, pp. 174-184, 2014.
- [105] G. Brown and J. L. Wyatt, "The Use of the Ambiguity Decomposition in Neural Network Ensemble Learning Methods.," in *Proceedings of the Twentieth International Conference on Machine Learning*, 2003, pp. 67-74.
- [106] J. W. Osborne, "Improving your data transformations: Applying the Box-Cox transformation," *Pract. Assessment, Res. Eval.*, vol. 15, no. 12, p. 12, 2010.
- [107] R. Dirk, "Frequency analysis of rainfall data," in *College on Soil Physics 30th Anniversary (1983-2013)*, The Abdus Salam International Centre for Theoretical Physica, 2013, pp. 244-10.
- [108] M. A. Shamim, M. Hassan, S. Ahmad, and M. Zeeshan, "A Comparison of Artificial Neural Networks (ANN) and Local Linear Regression (LLR) Techniques for Predicting Monthly Reservoir Levels," *KSCE J. Civ. Eng.*, vol. 20, pp. 971-977, 2016.
- [109] M. Hassan et al., "Development of sediment load estimation models by using artificial neural networking techniques," *Environ. Monit. Assess.*, vol. 187, no. 11, p. 686, Nov. 2015.
- [110] G. E. P. Box and D. R. Cox, "An Analysis of Transformations," *J. R. Stat. Soc.*, vol. 26, no. 2, pp. 211-252, 1964.
- [111] K. Beven, *Rainfall Runoff Modelling: The Primer*, Second. Wiley Blackwell, 2011.
- [112] H. L et al., "The streamflow estimation using the Xinanjiang rainfall runoff model and dual state-parameter estimation method," *J. Hydrol.*, vol. 480, pp. 102-114, 2013.
- [113] D. Shrestha and D. Solomatine, "Assessing uncertainty in rainfall-runoff models: Application of data-driven models," in *Flood Risk Management: Research and Practice*, no. July, Taylor & Francis, 2009, pp. 1563-1573.

- [114] P. Darbandsari and P. Coulibaly, “Journal of Hydrology: Regional Studies Inter-comparison of lumped hydrological models in data-scarce watersheds using different precipitation forcing data sets: Case study of Northern Ontario , Canada,” *J. Hydrol. Reg. Stud.*, vol. 31, no. August, p. 100730, 2020.
- [115] L. Breuer et al., “Assessing the impact of land use change on hydrology by ensemble modeling (LUCHEM). I: Model intercomparison with current land use,” *Adv. Water Resour.*, vol. 32, no. 2, pp. 129-146, 2009.
- [116] F. Garavaglia et al., “Impact of model structure on flow simulation and hydrological realism: From a lumped to a semi-distributed approach,” *Hydrol. Earth Syst. Sci.*, vol. 21, no. 8, pp. 3937-3952, 2017.
- [117] G. Seiller, F. Anctil, and C. Perrin, “Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions,” *Hydrol. Earth Syst. Sci.*, vol. 16, no. 4, pp. 1171-1189, 2012.
- [118] F. H. S. Chiew, M. J. Stewardson, and T. A. McMahon, “Comparison of six rainfall-runoff modelling approaches,” *J. Hydrol.*, vol. 147, no. 1-4, pp. 1-36, 1993.
- [119] T. Y. Gan, Y. Gusev, S. J. Burges, and O. Nasonova, “Performance comparison of a complex physics-based land surface model and a conceptual, lumped-parameter hydrological model at the basin-scale,” *IAHS Publ.*, no. 307, pp. 196-207, 2006.
- [120] T. Das, A. Brdossy, E. Zehe, and Y. He, “Comparison of conceptual model performance using different representations of spatial variability,” *J. Hydrol.*, vol. 356, no. 1-2, pp. 106-118, 2008.
- [121] A. H. Linde, J. C. J. H. Aerts, R. T. W. L. Hurkmans, and M. Eberle, “Comparing model performance of two rainfall-runoff models in the Rhine basin using different atmospheric forcing data sets,” *Hydrol. Earth Syst. Sci. Discuss.*, vol. 4, no. 6, pp. 4325-4360, 2007.
- [122] P. Shi et al., “Evaluating the SWAT Model for Hydrological Modeling in the Xixian Watershed and a Comparison with the XAJ Model,” *Water Resour. Manag.*, vol. 25, no. 10, pp. 2595-2612, 2011..

- [123] A. H. A. Suliman, M. Jajarmizadeh, S. Harun, and I. Z. Mat Darus, "Comparison of Semi-Distributed, GIS-Based Hydrological Models for the Prediction of Streamflow in a Large Catchment," *Water Resour. Manag.*, vol. 29, no. 9, pp. 3095-3110, 2015.
- [124] T. Vansteenkiste et al., "Intercomparison of five lumped and distributed models for catchment runoff and extreme flow simulation," *J. Hydrol.*, vol. 511, pp. 335-349, 2014.
- [125] J. Koch et al., "Inter-comparison of three distributed hydrological models with respect to seasonal variability of soil moisture patterns at a small forested catchment," *J. Hydrol.*, vol. 533, pp. 234-249, 2016.
- [126] G. Tegegne, D. K. Park, Y. Kim, and Y. O. Kim, "Selecting hydrologic modelling approaches for water resource assessment in the yongdam watershed," *J. Hydrol. New Zeal.*, vol. 56, no. 2, pp. 155-164, 2017.
- [127] M. Azmi, S. Araghinejad, and M. Kholghi, "Multi Model Data Fusion for Hydrological Forecasting using K-Nearest Neighbour Method," *Iran. J. Sci. Technol.*, vol. 34, pp. 81-92, 2010.
- [128] A. Behrangi, B. Khakbaz, T. Chun, A. Aghakouchak, and K. Hsu, "Hydrologic evaluation of satellite precipitation products over a mid-size basin," *J. Hydrol.*, vol. 397, no. 3-4, pp. 225-237, 2011.
- [129] A. I. J. M. Van Dijk, "Model-data fusion: using observations to understand and reduce uncertainty in hydrological models," in *International Congress on Modeling & Simulation*, 2011, no. December, pp. 12-16.
- [130] B. V. Dasarathy and S. Member, "Sensor Fusion Potential Exploitation Innovative Architectures and Illustrative Applications," *Proc. IEEE*, vol. 85, no. 1, 1997.
- [131] B. V. Dasarathy, "Information Fusion what , where , why , when , and how?," *Inf. Fusion*, vol. 2, pp. 75-76, 2001.
- [132] L. Alparone, S. Baronti, A. Garzelli, and F. Nencini, "A Global Quality Measurement of Pan-Sharpned Multispectral Imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 1, no. 4, pp. 313-317, 2004.

- [133] V. Tsagaris, V. Anastassopoulos, and G. A. Lampropoulos, "Fusion of Hyperspectral Data Using Segmented PCT for Color Representation and Classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 10, pp. 2365-2375, 2005.
- [134] K. C. Slatton, S. Cheung, and H. Jhee, "Reduced-Complexity Fusion of Multiscale Topography and Bathymetry Data Over the Florida Coast," *IEEE Geosci. Remote Sens. Lett.*, vol. 2, no. 4, pp. 389-393, 2005.
- [135] Y. H. Kaheil, M. K. Gill, M. Mckee, L. A. Bastidas, and E. Rosero, "Downscaling and Assimilation of Surface Soil Moisture Using Ground Truth Measurements," *IEEE Geosci. Remote Sens. Lett.*, vol. 46, no. 5, pp. 1375-1384, 2008.
- [136] L. Zhou et al., "A study on availability of ground observations and its impacts on bias correction of satellite precipitation products and hydrologic simulation efficiency," *J. Hydrol.*, vol. 610, p. 127595, Jul. 2022.
- [137] D. Guo, H. Wang, X. Zhang, and G. Liu, "Evaluation and analysis of grid precipitation fusion products in Jinsha river basin based on China meteorological assimilation datasets for the SWAT model," *Water (Switzerland)*, vol. 11, no. 2, 2019.
- [138] J. Pang et al., "Hydrological evaluation of open-access precipitation data using SWAT at multiple temporal and spatial scales," *Hydrol. Earth Syst. Sci.*, vol. 24, no. 7, pp. 3603-3626, 2020.
- [139] P. Mishra, V. V. Zaphu, N. Monica, A. Bhadra, and A. Bandyopadhyay, "Accuracy Assessment of MODIS Fractional Snow Cover Product for Eastern Himalayan Catchment," *J. Indian Soc. Remote Sens.*, vol. 44, no. 6, pp. 977-985, 2016.
- [140] G. A. Riggs, D. K. Hall, and M. O. Romn, "Overview of NASAs MODIS and VIIRS Snow-Cover Earth System Data Records," *Earth Syst. Sci. Data Discuss.*, no. April, pp. 1-30, 2017.
- [141] K. Rittger, K. J. Bormann, E. H. Bair, J. Dozier, and T. H. Painter, "Evaluation of VIIRS and MODIS Snow Cover Fraction in High-Mountain Asia Using Landsat 8 OLI," *Front. Remote Sens.*, vol. 2, May 2021.
- [142] Y. Bai et al., "Reservoir Inflow Forecast Using a Clustered Random Deep Fusion Approach in the Three Gorges Reservoir , China," vol. 23, no. 10, pp. 1-15, 2018.

- [143] A. Roy, A. Royer, and R. Turcotte, "Improvement of springtime streamflow simulations in a boreal environment by incorporating snow-covered area derived from remote sensing data," *J. Hydrol.*, vol. 390, no. 1-2, pp. 35-44, 2010.
- [144] P. J. Block, F. Assis, S. Filho, L. Sun, and H. Kwon, "A STREAMFLOW FORECASTING FRAMEWORK USING MULTIPLE CLIMATE AND HYDROLOGICAL MODELS 1," vol. 45, no. 4, pp. 828-843, 2009.
- [145] S. A. Quadri and O. Sidek, "Development of heterogeneous multisensor data fusion system to improve evaluation of concrete structures," *Int. J. Image Data Fusion*, vol. 5, no. 2, pp. 97-108, Apr. 2014.
- [146] Z. Wang et al., "Data fusion in data scarce areas using a back-propagation artificial neural network model: a case study of the South China Sea," *Front. Earth Sci.*, vol. 12, no. 2, pp. 280-298, Jun. 2018.
- [147] Q. Liu, K. Brigham, and N. S. V Rao, "Estimation and Fusion for Tracking Over Long-Haul Links Using Artificial Neural Networks," *IEEE Trans. Signal Inf. Process. over Networks*, vol. 3, no. 4, pp. 760-770, Dec. 2017.
- [148] E. Garca Plaza, P. Nez Lpez, and E. Beamud Gonzlez, "Multi-Sensor Data Fusion for Real-Time Surface Quality Control in Automated Machining Systems," *Sensors*, vol. 18, no. 12, p. 4381, Dec. 2018.
- [149] J. Liu, Y. Hu, Y. Wang, B. Wu, J. Fan, and Z. Hu, "An integrated multi-sensor fusion-based deep feature learning approach for rotating machinery diagnosis," *Meas. Sci. Technol.*, vol. 29, no. 5, p. 055103, 2018.
- [150] C. Shu and D. H. Burn, "Artificial neural network ensembles and their application in pooled flood frequency analysis," *Water Resour. Res.*, vol. 40, no. 9, pp. 1-10, 2004.
- [151] S. Srinivasulu and A. Jain, "A comparative analysis of training methods for artificial neural network rainfall-runoff models," *Appl. Soft Comput.*, vol. 6, no. 3, pp. 295-306, Mar. 2006.
- [152] F. Gkbulak et al., "Comparison of Rainfall-Runoff Relationship Modeling using Different Methods in a Forested Watershed," *Water Resour. Manag.*, vol. 29, no. 12, pp. 4229-4239, Sep. 2015.

- [153] B. Mishra, S. Hub, N. K. Tripathi, and M. S. Babel, "An artificial neural network-based snow cover predictive modeling in the higher Himalayas," *J. Mt. Sci.*, vol. 11, no. 4, pp. 825-837, 2014.
- [154] G. Lee, D. Kim, H. Kwon, and E. Choi, "Estimation of Maximum Daily Fresh Snow Accumulation Using an Artificial Neural Network Model," *Adv. Meteorol.*, p. 11, 2019.
- [155] P. D. Broxton, X. Zeng, D. Sulla-Menashe, and P. A. Troch, "A Global Land Cover Climatology Using MODIS Data," *J. Appl. Meteorol. Climatol.*, vol. 53, no. 6, pp. 1593-1605, Jun. 2014.
- [156] M. Ata, A. E. Tekeli, S. Dnmez, and H. Fouli, "Use of interactive multisensor snow and ice mapping system snow cover maps (IMS) and artificial neural networks for simulating river discharges in Eastern Turkey," *Arab. J. Geosci.*, vol. 9, no. 2, p. 150, Feb. 2016.
- [157] A. G. Yilmaz, M. A. Imteaz, and G. Jenkins, "Catchment flow estimation using Artificial Neural Networks in the mountainous Euphrates Basin," *J. Hydrol.*, vol. 410, no. 1-2, pp. 134-140, 2011.
- [158] F. Wang, G. Huang, G. Cheng, and Y. Li, "Advances in Water Resources Multi-level factorial analysis for ensemble data-driven hydrological prediction," *Adv. Water Resour.*, vol. 153, no. June 2020, p. 103948, 2021.
- [159] J. M. Quilty, A. E. Sikorska-senoner, and D. Hah, "A stochastic conceptual-data-driven approach for improved hydrological simulations," *Environ. Model. Softw.*, vol. 149, no. December 2021, p. 105326, 2022.
- [160] R. Duan, G. Huang, Y. Li, X. Zhou, J. Ren, and C. Tian, "Stepwise clustering future meteorological drought projection and multi-level factorial analysis under climate change: A case study of the Pearl River Basin, China," *Environ. Res.*, vol. 196, no. October, p. 110368, 2021.
- [161] DAWN, "Geography: The Rivers of Pakistan," 2009.
- [162] World Bank, "Climate change 2007: Synthesis report, Contribution of working groups i, ii and iii to the fourth assessment report of the Intergovernmental Panel on Climate Change," 2007.

- [163] K. Jamal, S. Ahmad, X. Li, M. Rizwan, H. Li, and J. Feng, "Climate change and runoff contribution by hydrological zones of cryosphere catchment of Indus River , Pakistan," *Hydrol. Earth Syst. Sci. Discuss.*, no. November, pp. 1-31, 2018.
- [164] M. Akhtar, N. Ahmad, and M. J. Booij, "The impact of climate change on the water resources of Hindukush-Karakorum-Himalaya region under different glacier coverage scenarios," *J. Hydrol.*, vol. 355, no. 1-4, pp. 148-163, 2008.
- [165] G. Rasul, Q. Dahe, and Q. Z. Chaudhry, "Global Warmin and Melting Glaciers along Southern Slopes of HKH Ranges," *Pakistan J. Meteorol.*, vol. 5, no. 9, pp. 63-76, 2008.
- [166] S. S. Hussain, M. Mudasser, and M. Munir, "Climate change and variability in mountain regions of Pakistan Implications for water and Agriculture," *Pakistan J. Meteorol.*, vol. 2, no. 4, pp. 75-90, 2005.
- [167] U. A. Naeem, H. N. Hashmi, M. A. Shamim, and N. Ejaz, "Flow Variation in Astore River under Assumed Glaciated Extents Due to Climate Change," *Pakistan J. Eng. Appl. Sci.*, vol. 11, pp. 73-81, 2012.
- [168] F. Malik, "Box-Cox Transformation Approach for Data Normalization: a Study of New Product Development in Manufacturing Sector of Pakistan," *IBT J. Bus. Stud.*, vol. 14, no. 1, pp. 110-119, 2018.
- [169] J. W. Tukey, "On the Comparative Anatomy of Transformations," *Ann. Math. Stat.*, vol. 32, no. 1, pp. 12-40, 1957.
- [170] B. F. . Manly, "Exponential Data Transformation," vol. 25, no. 1, pp. 37-42, 1976.
- [171] J. A. John and N. R. Draper, "An Alternative Family of Transformations," *Appl. Stat.*, vol. 29, no. 2, p. 190, 1980.
- [172] P. J. Bickel and K. A. Doksum, "An analysis of transformations revisited," *J. Am. Stat. Assoc.*, vol. 76, no. 374, pp. 296-311, 1981.
- [173] R. Remesan, M. Ali, D. Han, and J. Mathew, "Runoff prediction using an integrated hybrid modelling scheme," *J. Hydrol.*, vol. 372, no. 1-4, pp. 48-60, 2009.
- [174] A. Stefansson, K. N, and J. Antonia J, "A note on the Gamma test," *Neurocomputing Appl.*, vol. 5, pp. 131-133, 1997.

- [175] A. Sharifi, Y. Dinpashoh, and R. Mirabbasi, "Daily runoff prediction using the linear and non-linear models," *Water Sci. Technol.*, vol. 76, no. 4, pp. 793-805, 2017.
- [176] A. B. Dariane, M. A. Gol, and F. Karami, "Forecasting of rainfall using different input selection methods on climate signals for neural network inputs," *J. Hydraul. Structures*, vol. 5, no. 1, pp. 42-59, 2019.
- [177] L. Alzubaidi et al., *Review of deep learning: concepts, CNN architectures, challenges, applications, future directions*, vol. 8, no. 1. Springer International Publishing, 2021.
- [178] P. J. Durrant, "A non-linear data analysis and modeling tool with applications to flood prediction.," University of Wales, 2001.
- [179] A. Elshorbagy, G. Corzo, S. Srinivasulu, and D. P. Solomatine, "Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology - Part 2: Application," *Hydrol. Earth Syst. Sci.*, vol. 14, no. 10, pp. 1943-1961, 2010.
- [180] M. Hassan et al., "Development of sediment load estimation models by using artificial neural networking techniques," *Environ. Monit. Assess.*, vol. 187, no. 11, p. 686, Nov. 2015.
- [181] M. A. Shamim, M. Hassan, S. Ahmad, and M. Zeeshan, "A comparison of Artificial Neural Networks (ANN) and Local Linear Regression (LLR) techniques for predicting monthly reservoir levels," *KSCE J. Civ. Eng.*, vol. 20, no. 2, pp. 971-977, Mar. 2016.
- [182] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press Cambridge, 2016.
- [183] M. Minsky and S. Papert, *Perceptrons*. MIT Press Cambridge, 1969.
- [184] A. Jones, "New tools in non-linear modelling and prediction," *Comput. Manag. Sci.*, vol. 1, no. 2, pp. 109-149, 2004.
- [185] J. Brownlee, *Better Deep Learning: Train Faster, Reduce Overfitting, and Make Better Predictions*. 2018.

- [186] A. J. Litta, S. M. Idicula, and U. C. Mohanty, "Artificial Neural Network Model in Prediction of Meteorological Parameters during Premonsoon Thunderstorms," *Int. Journal Atmos. Sci.*, 2013.
- [187] O. Kisi and M. Cimen, "Engineering Applications of Artificial Intelligence Precipitation forecasting by using wavelet-support vector machine conjunction model," *Eng. Appl. Artif. Intell.*, vol. 25, no. 4, pp. 783-792, 2012.
- [188] A. Khan, B. S. Naz, and L. C. Bowling, "Separating snow, clean and debris covered ice in the Upper Indus Basin, Hindukush-Karakoram-Himalayas, using Landsat images between 1998 and 2002," *J. Hydrol.*, vol. 521, pp. 46-64, 2015.
- [189] A. E. Tekeli, I. Snmez, and E. Erdi, "Snow-covered area determination based on satellite-derived probabilistic snow cover maps," *Arab. J. Geosci.*, vol. 9, no. 3, p. 198, Mar. 2016.
- [190] A. G. Klein and A. C. Barnett, "Validation of daily MODIS snow cover maps of the Upper Rio Grande River Basin for the 2000-2001 snow year," *Remote Sens. Environ.*, vol. 86, no. 2, pp. 162-176, 2003.
- [191] Z. Pu, L. Xu, and V. V. Salomonson, "MODIS/Terra observed seasonal variations of snow cover over the Tibetan Plateau," *Geophys. Res. Lett.*, vol. 34, no. 6, pp. 1-6, 2007.
- [192] D. K. Hall and G. A. Riggs, "Accuracy assessment of the MODIS snow products," vol. 1547, pp. 1534-1547, 2007.
- [193] X. Huang, T. Liang, X. Zhang, and Z. Guo, "Validation of MODIS snow cover products using landsat and ground measurements during the 2001-2005 snow seasons over northern Xinjiang, China," *Int. J. Remote Sens.*, vol. 32, no. 1, pp. 133-152, 2011.
- [194] A. Gafurov, D. Kriegel, S. Vorogushyn, and B. Merz, "Evaluation of remotely sensed snow cover product in Central Asia," *Hydrol. Res.*, vol. 44, no. 3, pp. 506-522, 2013.
- [195] E. Snieder, R. Shakir, and U. T. Khan, "Comparison of four input variable selection methods for artificial neural network based flood forecasting models," in *Proceedings, Annual Conference - Canadian Society for Civil Engineering*, 2019, vol. 2019-June, pp. 1-10.

- [196] R. May, G. Dandy, and H. Maier, "Review of Input Variable Selection Methods for Artificial Neural Networks," in *Artificial Neural Networks - Methodological Advances and Biomedical Applications*, no. January 2016, InTech, 2011.
- [197] N. Vo, H. Shi, and J. Szajman, "Sensitivity analysis and optimisation to input variables using winGamma and ANN: A case study in automated residential property valuation," *Int. J. Adv. Appl. Sci.*, vol. 2, no. 12, pp. 19-24, 2015.
- [198] T. Tirelli and D. Pessani, "Importance of feature selection in decision-tree and artificial-neural-network ecological applications. Alburnus alburnus alborella: A practical example," *Ecol. Inform.*, vol. 6, no. 5, pp. 309-315, 2011.
- [199] S. E. Kemp, "Gamma test analysis tools for non-linear time series by," University of Glamorgan, Wales UK, 2006.
- [200] S. Katoch, S. S. Chauhan, and V. Kumar, "A review on genetic algorithm: past, present, and future," *Multimed. Tools Appl.*, vol. 80, no. 5, pp. 8091-8126, Feb. 2021.
- [201] S. Goswami, S. Chakraborty, P. Guha, A. Tarafdar, and A. Kedia, "Filter-Based Feature Selection Methods Using Hill Climbing Approach," 2019, pp. 213-234.
- [202] F. G. Lobo and C. F. Lima, "Adaptive population sizing schemes in genetic algorithms," *Stud. Comput. Intell.*, vol. 54, no. 2007, pp. 185-204, 2007.
- [203] R. M. Sakia, "The Box-Cox transformation technique: a review," *Stat.*, pp. 169-178, 1992.
- [204] L.-H. Tang, Y.-L. Bai, J. Yang, and Y.-N. Lu, "A hybrid prediction method based on empirical mode decomposition and multiple model fusion for chaotic time series," *Chaos, Solitons & Fractals*, vol. 141, p. 110366, Dec. 2020.
- [205] W. Wang, "Evaluating The Performance of Several Data Preprocessing Methods Based On GRU in Forecasting Monthly Runoff Time Series," *Research Sq.*, 2021.
- [206] M. Tayyab, I. Ahmad, N. Sun, J. Zhou, and X. Dong, "Application of Integrated Artificial Neural Networks Based on Decomposition Methods to Predict Streamflow at Upper Indus Basin, Pakistan," *Atmosphere (Basel)*, vol. 9, no. 12, p. 494, Dec. 2018.

-
- [207] M. I. Shah, A. Khan, T. A. Akbar, Q. K. Hassan, A. J. Khan, and A. Dewan, “Predicting hydrologic responses to climate changes in highly glacierized and mountainous region Upper Indus Basin:,” *R. Soc. Open Sci.*, vol. 7, no. 8, 2020.

Annex-4A

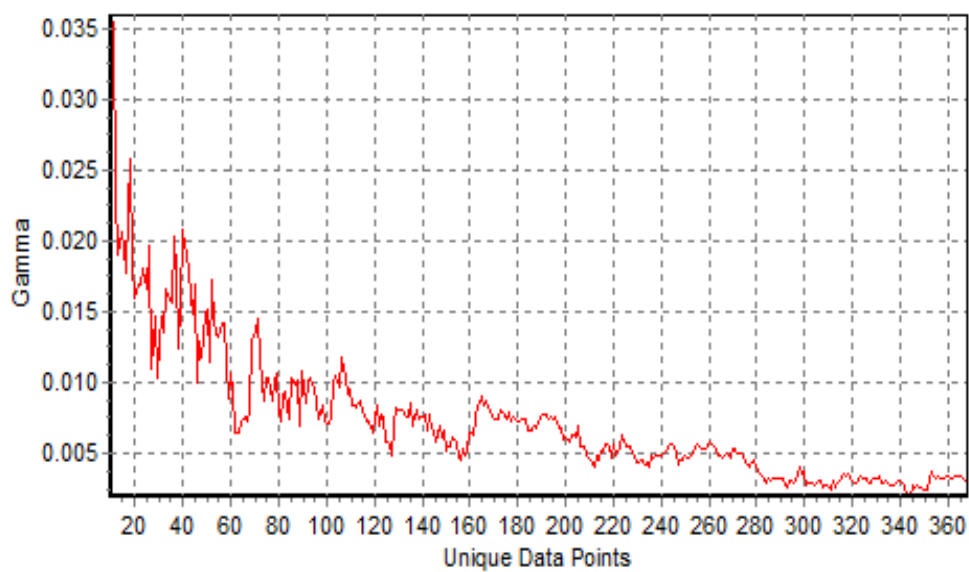


FIGURE 1: M-Test Result for combination 01 (Only P)

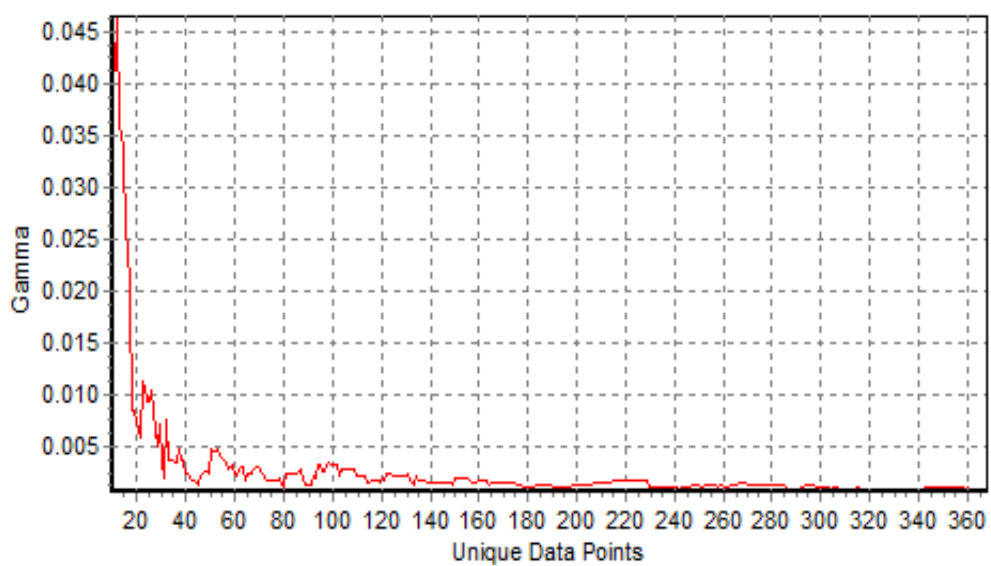


FIGURE 2: M-Test result for combination 03 (Only Q)

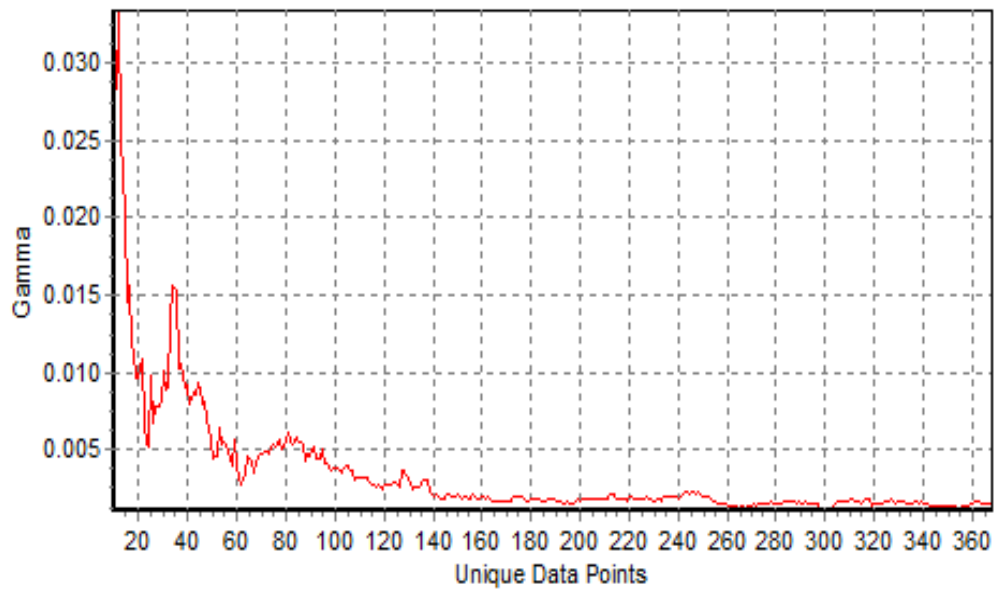


FIGURE 3: M-Test result for combination 04 (P+Q)

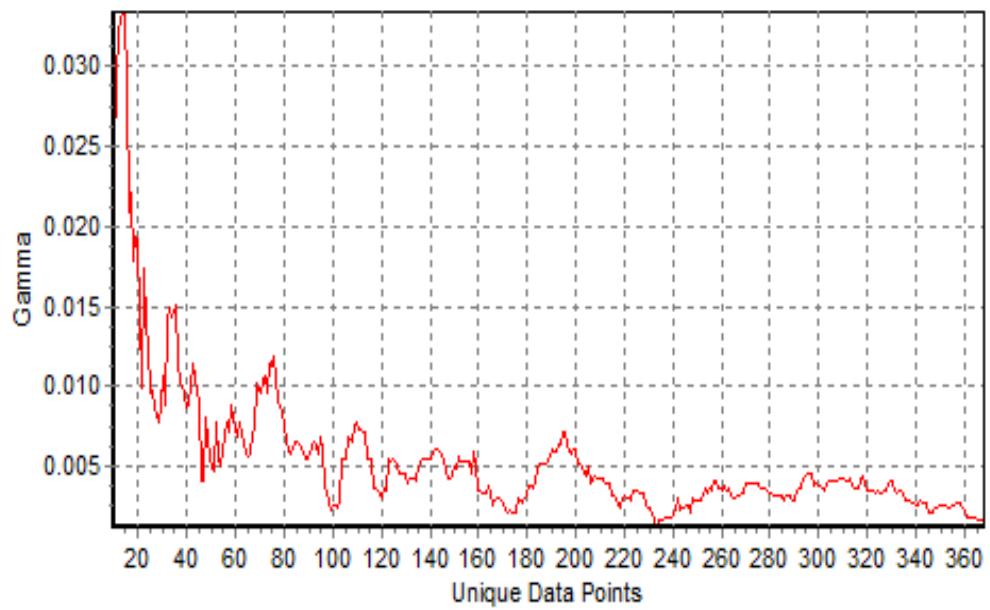


FIGURE 4: M-Test result for combination 05 (P+SR)

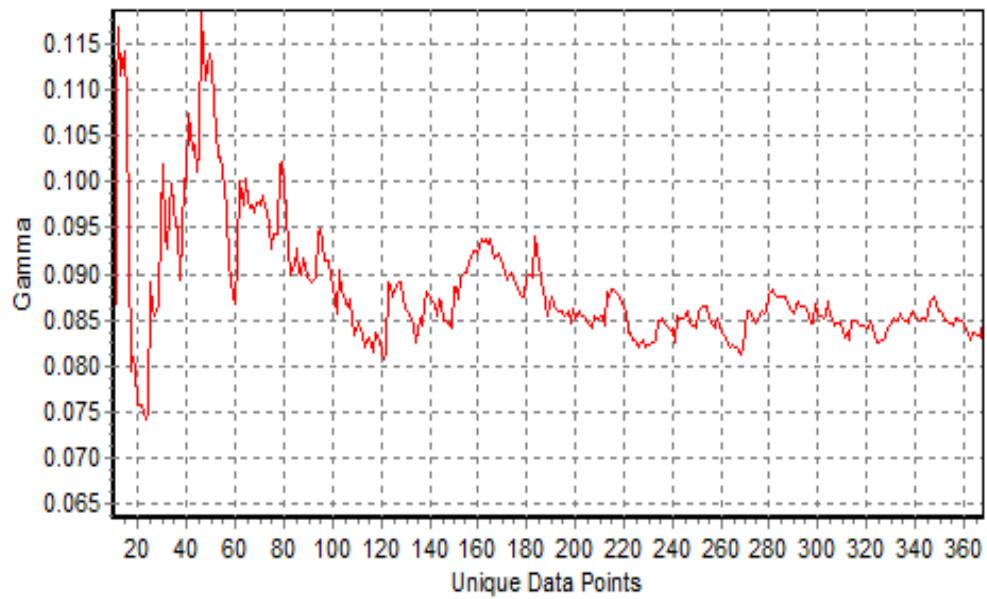


FIGURE 5: M-Test result for combination 06 (Only SCA)

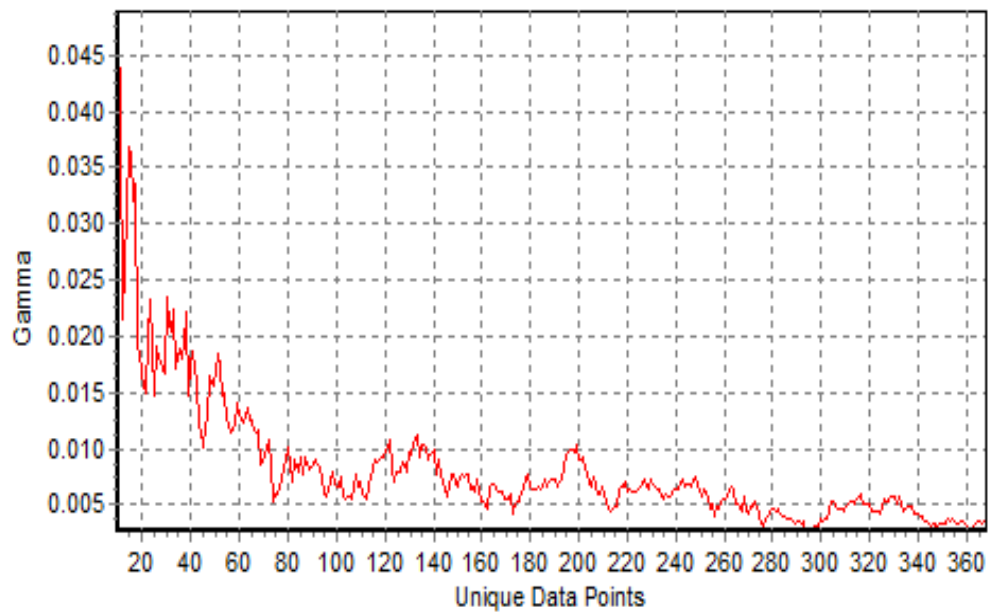


FIGURE 6: M-Test result for combination 07 (SCA+Q*)

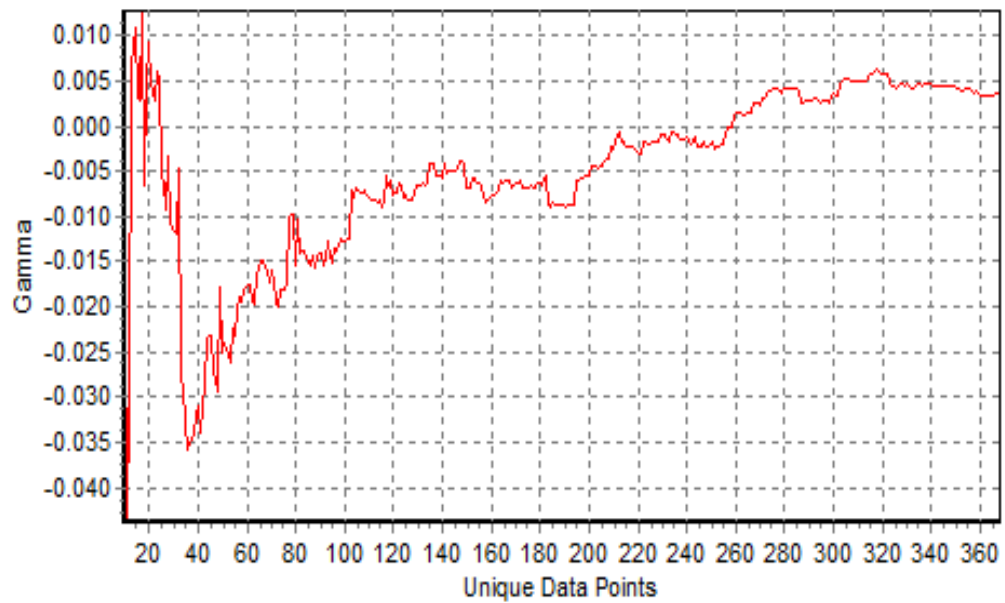


FIGURE 7: M-Test result for combination 09 (P+S+Q)

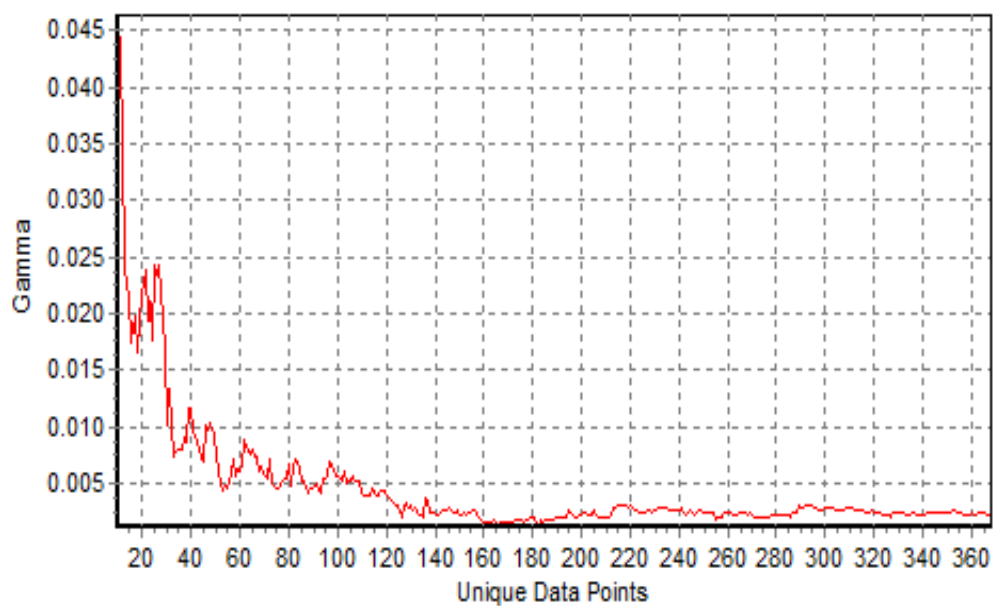


FIGURE 8: M-Test result for combination 11 (ALL)

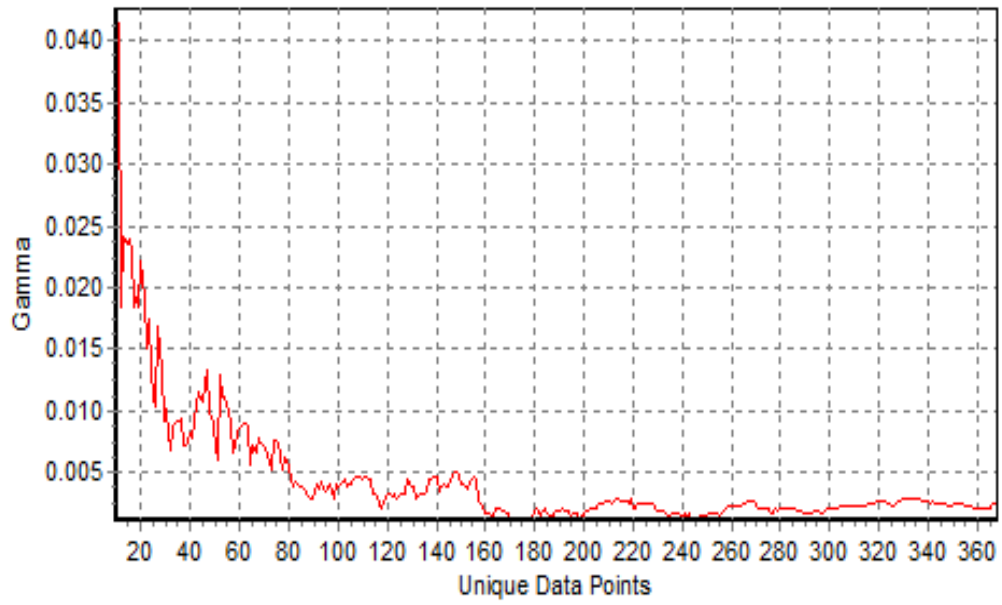


FIGURE 9: M-Test result for combination 14 (HC)

Annex-4B

- Table 1: R^2 Values for different architectures of ANN using multiple combination options
- Table 2: NSE Values for different architectures of ANN using multiple combination options
- Table 3: RMSE Values for different architectures of ANN using multiple combination options
- Table 4: BIAS Values for different architectures of ANN using multiple combination options

TABLE 1: R^2 Values for different architectures of ANN using multiple combination options

Models	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1-1 (Train)	0.37	0.44	0.92	0.93	0.66	0.04	0.12	0.93	0.94	0.94	0.94	0.66	0.94	0.94	0.94
1-1 (Test)	0.31	0.4	0.91	0.88	0.66	0.05	0.05	0.91	0.91	0.91	0.89	0.45	0.91	0.91	0.91
1-3 (Train)	0.37	0.46	0.93	0.93	0.72	0.04	0.1	0.93	0.95	0.94	0.96	0.68	0.95	0.95	0.94
1-3 (Test)	0.29	0.46	0.9	0.88	0.43	0.05	0.00	0.91	0.91	0.91	0.89	0.47	0.88	0.73	0.91
1-5 (Train)	0.37	0.42	0.93	0.94	0.67	0.01	0.17	0.94	0.95	0.95	0.96	0.79	0.96	0.96	0.95
1-5 (Test)	0.27	0.43	0.88	0.55	0.43	0.03	0.06	0.84	0.87	0.89	0.38	0.53	0.56	0.78	0.91
2-2 (Train)	0.38	0.53	0.93	0.94	0.73	0.01	0.18	0.94	0.95	0.96	0.97	0.76	0.95	0.97	0.96
2-2 (Test)	0.29	0.44	0.93	0.66	0.56	0.02	0.00	0.86	0.9	0.74	0.85	0.47	0.59	0.74	0.83
3-1 (Train)	0.38	0.53	0.94	0.97	0.73	0.05	0.18	0.95	0.95	0.98	0.97	0.82	0.98	0.98	0.97
3-1 (Test)	0.34	0.41	0.9	0.56	0.57	0.05	0.15	0.84	0.88	0.78	0.67	0.43	0.83	0.52	0.83
3-3 (Train)	0.37	0.53	0.95	0.98	0.85	0.04	0.18	0.95	0.95	0.98	0.97	0.83	0.99	0.99	0.98
3-3 (Test)	0.32	0.46	0.89	0.73	0.58	0.01	0.01	0.88	0.9	0.57	0.77	0.48	0.62	0.44	0.8
4-4 (Train)	0.37	0.53	0.95	0.98	0.73	0.04	0.17	0.96	0.95	0.98	0.97	0.83	0.99	1.00	0.99
4-4 (Test)	0.33	0.52	0.91	0.81	0.59	0.02	0.08	0.91	0.89	0.8	0.86	0.49	0.79	0.71	0.71
5-1 (Train)	0.37	0.54	0.95	0.97	0.72	0.03	0.18	0.95	0.95	0.98	0.97	0.83	0.99	1.00	0.98
5-1 (Test)	0.34	0.48	0.89	0.66	0.51	0.05	0.05	0.90	0.90	0.74	0.83	0.52	0.85	0.36	0.82
5-5 (Train)	0.37	0.54	0.95	0.98	0.73	0.03	0.20	0.96	0.96	0.98	0.97	0.83	1.00	1.00	0.99
5-5 (Test)	0.35	0.7	0.91	0.78	0.54	0.05	0.02	0.91	0.87	0.82	0.91	0.62	0.69	0.54	0.54

TABLE 2: NSE Values for different architectures of ANN using multiple combination options

Models	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1-1 (Train)	0.70	0.73	0.83	0.97	0.84	0.52	0.57	0.96	0.97	0.76	0.93	0.83	0.97	0.97	0.97
1-1 (Test)	0.68	0.99	1.00	0.94	0.82	0.56	0.55	0.96	0.96	0.96	0.95	0.74	0.96	0.96	0.96
1-3 (Train)	0.70	0.74	0.96	0.97	0.87	0.52	0.57	0.96	0.97	0.97	0.98	0.84	0.98	0.98	0.97
1-3 (Test)	0.66	0.75	0.95	0.94	0.72	0.56	0.51	0.96	0.96	0.96	0.95	0.75	0.94	0.87	0.96
1-5 (Train)	0.70	0.72	0.97	0.97	0.84	0.51	0.60	0.97	0.98	0.97	0.98	0.90	0.98	0.98	0.97
1-5 (Test)	0.66	0.74	0.95	0.76	0.73	0.54	0.55	0.92	0.93	0.95	0.63	0.77	0.73	0.89	0.96
2-2 (Train)	0.70	0.77	0.97	0.97	0.87	0.49	0.60	0.97	0.98	0.98	0.99	0.88	0.98	0.99	0.98
2-2 (Test)	0.67	0.74	0.97	0.80	0.79	0.53	0.40	0.94	0.95	0.85	0.93	0.74	0.75	0.87	0.92
3-1 (Train)	0.70	0.77	0.97	0.99	0.87	0.53	0.60	0.98	0.98	0.99	0.99	0.91	0.99	0.99	0.98
3-1 (Test)	0.69	0.72	0.96	0.73	0.8	0.56	0.60	0.93	0.95	0.89	0.83	0.70	0.89	0.73	0.91
3-3 (Train)	0.70	0.77	0.98	0.99	0.86	0.52	0.60	0.98	0.98	0.99	0.99	0.91	0.99	1.00	0.99
3-3 (Test)	0.68	0.75	0.95	0.86	0.81	0.54	0.45	0.95	0.96	0.74	0.88	0.73	1.00	0.56	0.90
4-4 (Train)	0.70	0.77	0.98	0.99	0.87	0.52	0.60	0.98	0.98	0.99	0.99	0.91	1.00	1.00	1.00
4-4 (Test)	0.69	0.78	0.96	0.90	0.81	0.55	0.57	0.96	0.95	0.89	0.94	0.75	0.89	0.83	0.85
5-1 (Train)	0.70	0.77	0.98	0.99	0.87	0.51	0.60	0.98	0.98	0.99	0.99	0.91	1.00	1.00	0.99
5-1 (Test)	0.69	0.76	0.95	0.83	0.77	0.56	0.54	0.96	0.95	0.87	0.92	0.76	0.93	0.53	0.92
5-5 (Train)	0.70	0.78	0.98	0.99	0.87	0.51	0.60	0.98	0.98	0.99	0.99	0.92	1.00	1.00	1.00
5-5 (Test)	0.70	0.77	0.96	0.88	0.79	0.56	0.42	0.96	0.94	0.91	0.96	0.82	0.82	0.67	0.72

TABLE 3: RMSE Values for different architectures of ANN using multiple combination options

Models	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1-1 (Train)	12077.4	11361.7	4109.6	3964.1	8781.4	14562.8	14349.6	3896.6	3717.0	3717.1	3505.1	8656.9	3648.1	3669.9	3710.1
1-1 (Test)	12737.5	12168.9	387.8	5298.9	8862.9	15609.3	15052.4	4762.4	4692.6	4675.8	5281.3	11984.7	4728.7	4666.0	4603.2
1-3 (Train)	12101.4	11161.4	3970.0	3872.1	7960.1	14566.1	14499.3	3885.6	3530.2	3577.8	3051.2	8295.4	3257.4	3242.8	3592.1
1-3 (Test)	13035.1	11561.0	5043.1	5390.3	12157.6	15612.6	15770.3	4766.3	4686.6	4649.0	5473.6	11785.1	5573.8	8592.7	4778.7
1-5 (Train)	12113.7	11505.6	3927.8	3624.0	8665.1	14769.6	13921.1	3438.9	3254.4	3393.9	3098.9	6764.4	2861.8	2872.5	3519.5
1-5 (Test)	13170.8	11864.4	5451.1	11000.7	12027.5	15856.9	15104.6	6784.4	5933.5	5250.1	14247.2	11244.8	12151.1	7625.0	4798.9
2-2 (Train)	12029.3	10379.7	3888.1	3628.5	7930.1	14968.4	13861.1	3563.5	3253.9	2834.1	2521.0	7268.5	3222.1	2390.2	3089.9
2-2 (Test)	12931.7	11720.2	4311.0	9878.7	10469.2	16133.4	17396.9	5979.4	4932.7	9126.1	6202.7	12052.9	11646.7	8332.3	6687.0
3-1 (Train)	12016.5	10378.0	3700.8	2372.0	7931.7	14479.8	13860.6	3112.1	3258.2	2276.4	2515.2	6274.0	2239.4	1926.3	2798.8
3-1 (Test)	12543.2	12160.2	4888.2	11565.7	10266.8	15621.7	14145.0	6515.0	5319.7	7883.0	9607.3	12760.3	7637.6	12093.5	6838.5
3-3 (Train)	12097.1	10376.6	3328.1	2355.5	7965.3	14536.4	13842.6	3109.0	3217.9	2265.7	2494.5	6113.1	1756.2	1418.0	2340.4
3-3 (Test)	12769.8	11457.4	5249.5	8231.6	10186.7	15832.3	16642.9	5545.9	4899.8	12210.6	8163.1	12077.7	11057.6	15483.4	7417.0
4-4 (Train)	12112.4	10331.0	3326.0	2363.9	7909.8	14592.7	13920.9	3101.6	3228.6	2256.9	2470.8	6122.8	1151.9	511.3	1220.4
4-4 (Test)	12585.4	10923.6	4845.2	6992.0	10048.3	15804.0	14707.9	4801.9	5195.1	7929.8	5868.2	11805.8	7820.7	9733.2	9077.2
5-1 (Train)	12103.1	10307.5	3329.7	2373.8	7939.8	14641.8	13897.4	3108.4	3255.3	2269.0	2512.8	6145.4	1133.5	504.5	2157.3
5-1 (Test)	12502.3	11310.0	5175.7	9298.7	11156.9	15633.0	15273.4	5014.7	5016.4	8547.2	6745.8	11378.1	6150.9	16041.3	6677.1
5-5 (Train)	12106.7	10256.0	3328.6	2352.2	7873.8	14649.8	13917.1	3092.4	3196.7	2269.9	2488.1	6064.7	771.5	261.7	1182.9
5-5 (Test)	12395.6	16986.0	4773.4	7565.3	10664.3	15654.9	17053.1	4870.3	5695.1	6950.5	4670.1	9955.2	9989.5	13525.3	12148.7

TABLE 4: RBIAS Values for different architectures of ANN using multiple combination options

Models	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1-1 (Train)	-25.5	0	1.3	2.6	2.9	-6.6	-60.1	-0.6	-1.7	-0.4	0.8	14.3	0.1	0.5	100.2
1-1 (Test)	-693.5	-950.6	-109.3	599.9	1009.4	-2435.5	546.9	40.7	382.3	102	-48	-1012.5	-167.7	111.7	277.6
1-3 (Train)	-203.1	2.1	-0.1	8.9	-62.7	-3.8	-2.1	2	1	-7.3	3	2.5	-51.2	0.2	0.3
1-3 (Test)	-959.4	-990.3	-171.5	737.8	-1262.8	-2452.7	-993.5	17.2	336.2	52.5	29.1	-1150.3	-364.4	-513.8	268.2
1-5 (Train)	83.9	352.7	-5.5	-1.1	1.2	21.4	377.9	-0.1	98.7	-6.7	0.8	24.1	0	1	0.2
1-5 (Test)	-266.7	-637.8	-322.1	-833.9	-1342	-2723	1053.7	843.2	1149.2	161.2	-2758	-580.7	-649.6	-353.5	249.9
2-2 (Train)	-163.4	227.7	-10.8	-157.8	176.2	-1213.7	-929.8	0.1	321.1	0.6	-15.9	-1.3	1	2.2	0.1
2-2 (Test)	-133	-666	-117.8	1463.3	-546.7	-3831.9	1332.4	435.6	328.3	1681.9	175	-303.5	1613.7	-1285.9	1112.7
3-1 (Train)	-468.5	-697.1	-1.3	29.5	457.4	48.5	689.2	8	-15	26.9	-9	16.1	-0.8	-10.1	0.5
3-1 (Test)	-538.4	-1791.8	101.2	2621.7	134.2	-2428.7	465.4	748.1	30	978.5	441.5	444.3	1242.2	-118.5	44.2
3-3 (Train)	99.5	-110.8	3.4	-48.6	-69.8	71.8	1023.9	1	89.4	49.1	-3.5	144.1	0.9	1.3	-0.4
3-3 (Test)	-119.3	58.9	359.1	-171.9	-1183.7	-2122.7	2781	97.6	189.2	2279.8	794.4	-431.5	600.8	1656.8	246.6
4-4 (Train)	-103.1	-218.3	-14.6	67.8	679.9	41	-399.2	23.2	-51.9	79.1	16	304	-2	-0.4	0.8
4-4 (Test)	-909.2	-1052.2	14	1139.4	-426.2	-2212.2	214.4	294	-361	1702.1	-612.1	-80.6	383.2	957	-939.7
5-1 (Train)	-97.8	447.4	-2.3	-27.1	241.9	59.9	683.2	25.4	-148.1	58.9	-87.3	278.5	-0.8	0.1	-10.3
5-1 (Test)	-546.7	188.3	-320.8	-294.2	-858.6	-2481	931.3	-69.6	182.1	690.3	192.4	13.9	-50.6	669	-496.4
5-5 (Train)	-46.3	-510.2	-21.9	173.7	171.3	-104	-307.7	33.2	-85	-24.9	-25.1	-192.7	-4.2	-0.3	-0.4
5-5 (Test)	-555.4	428.1	66.2	912.9	-677	-2538.3	2128.7	392	-179.1	792.6	-39.6	-421.1	137.8	59.5	900.2

Annex-4C

- Table 5: Results of Models developed without Solar Radiation
- Table 6: Results of models developed by input combinations selected through GA and Gamma Test

TABLE 5: Results of Models developed without Solar Radiation

Node Arrangement	Original Data ($\lambda=1$)										Transformed Data ($\lambda=0.005$)									
	Target MSE= 0.0012976										Target MSE= 4.3006×10^{-7}									
	MASK (All data except SR)										MASK (All data except SR)									
	R^2		NSE		RMSE		VARIANCE		BIAS		R^2		NSE		RMSE		VARIANCE		BIAS	
						$(\times 10^5)$								$(\times 10^5)$						
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
5-5	90.8	92.0	95.6	95.2	674.6	691.0	4.5	4.7	92.5	-72.6	92.2	89.7	96.3	93.5	618.0	805.2	3.8	6.2	-62.0	-170.0
6-3	90.8	94.0	95.6	95.6	672.6	18.6	4.5	-0.1	41.9	-102.8	92.2	75.0	96.2	87.2	620.4	1131.4	3.8	12.8	-48.0	-43.2
4-6	90.8	93.0	95.6	95.5	668.3	671.3	4.5	4.4	-0.1	-124.9	92.4	91.8	96.3	95.0	614.1	704.4	3.7	4.7	-57.8	-144.3
3-3	90.7	90.1	95.6	94.5	675.0	742.1	4.5	5.5	57.6	-75.4	91.1	83.8	95.8	91.6	658.7	914.4	4.3	8.3	-50.6	18.2
2-2	90.8	91.4	95.6	94.6	672.6	732.2	4.5	5.3	64.6	-83.9	89.1	91.5	75.7	95.2	1578.2	691.1	25.0	4.7	1.9	-31.4
1-1	90.7	93.5	95.6	95.5	671.3	669.0	4.5	4.1	2.8	-193.0	89.1	93.6	94.7	96.0	733.0	628.7	5.3	3.8	-70.0	-77.5
6-2	90.8	82.0	95.6	90.5	673.7	979.2	4.5	9.4	81.4	-94.2	92.7	92.0	96.5	94.5	595.2	726.2	3.5	5.1	-23.2	-114.5
8-3	90.7	92.4	95.6	95.2	673.8	686.8	4.5	4.6	68.4	-99.3	91.8	89.8	96.0	94.0	635.6	778.7	4.0	5.7	-69.2	-181.6
3-8	90.8	93.1	95.6	95.8	672.1	649.7	4.5	4.2	55.0	-58.3	91.5	91.5	96.0	96.7	644.8	552.3	4.1	3.0	-53.8	-75.0

TABLE 6: Results of models developed by input combinations selected through GA and Gamma Test

Node Arrangement	Original Data ($\lambda=1$)										Transformed Data ($\lambda=0.005$)									
	Target MSE= 0.0012976										Target MSE= 4.3006×10^{-7}									
	MASK 10110010111111110001										MASK 10101110100110111011									
	R^2	NSE		RMSE		VARIANCE		BIAS			R^2	NSE		RMSE		VARIANCE		BIAS		
					$(\times 10^5)$							$(\times 10^5)$								
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
5-5	97.1	70.2	98.7	85.5	372.2	1192.7	1.4	13.8	1.26	-200.7	89.5	79.1	97.7	89.4	485.6	1035.4	2.4	10.7	-28.1	-12.7
6-3	97.1	67.8	98.7	79.1	372.2	1451.2	1.4	20.9	-1.35	98.5	94.2	78.7	97.2	89	531.9	1050.7	2.8	10.9	-33.9	105.7
4-6	97	18.2	98.6	35.6	382.3	2548	1.5	63.7	0.085	-347.5	94.3	63.1	97.3	79.4	528.8	1442.1	2.8	20.8	-34.6	-18.1
3-3	95.2	63.2	97.7	79.6	483.1	1434.3	2.3	20.4	0.24	-142.1	93.2	87.9	96.7	93.5	578.8	807.7	3.3	6.5	-39.3	-85.7
2-2	88.4	64.2	94.4	80.3	758	1407.2	5.7	19.8	21.4	-45.6	91.2	90.4	95.8	94	658.6	776.7	4.3	5.8	-55.5	-144.9
1-1	91.3	73.1	95.9	87.2	651.6	651.8	4.3	3.9	-0.08	-189.5	89.5	94.1	95	96.1	719.6	630.1	5.1	3.9	-64.9	-103.8
6-2	97.1	54.8	98.7	75	372.2	1584.6	1.4	24.9	4.55	-127.5	94.7	70.6	97.5	83.6	506.5	1284.5	2.6	16.5	-29.1	44.9
8-3	97.2	69.2	98.7	84	371.8	1269.7	1.4	15.9	-3.87	-163.9	96.2	55.4	98.2	71.4	429.2	1696.2	1.8	28	-21.4	278.4
3-8	96.7	56.2	98.4	75.2	402.8	1580.8	1.6	24.9	-0.08	33.6	94.9	80.4	97.6	89.8	498.6	1008.3	2.5	10.1	-29.5	-71.7