LONG RANGE PREDICTION OF INDIAN SUMMER MONSOON RAINFALL USING DATA MINING AND STATISTICAL APPROACHES

Thesis

Submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

by

VATHSALA H

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA,
SURATHKAL, MANGALORE - 575025
MARCH 2018

Dedicated to my beloved parents

Behind Every Young Child Who Believes In Himself Is A Parent Who Believed First.

- Matthew Jacobson

DECLARATION

by the Ph.D. Research Scholar

I hereby declare that the research thesis entitled **LONG RANGE PREDICTION OF INDIAN SUMMER MONSOON RAINFALL USING DATA MINING AND STATISTICAL APPROACHES** which is being submitted to the **National Institute of Technology Karnataka, Surathkal** in partial fulfilment of the requirements for the award of the Degree of **Doctor of Philosophy** in **Computer Science and Engineering** is a *bonafide report of the research work carried out by me.* The material contained in this research thesis has not been submitted to any university or institution for the award of any degree.

(CO10P01, Vathsala H)
Department of Computer Science & Engineering

Place: NITK, Surathkal.
Date:

CERTIFICATE

This is to *certify* that the research thesis entitled **LONG RANGE PREDICTION OF INDIAN SUMMER MONSOON RAINFALL USING DATA MINING AND STATISTICAL APPROACHES** submitted by **Vathsala H**, (Register Number: CO10P01) as the record of the research work carried out by her, is *accepted as the research thesis submission* in partial fulfilment of the requirements for the award of degree of **Doctor of Philosophy**.

Dr. Shashidhar G Koolagudi
Research Guide

Chairman - DRPC

ACKNOWLEDGMENT

Gratitude is the heart's memory.
It is the fairest blossom which springs from the heart.

After an intensive period of six years, writing this note of gratitude is the finishing touch on my dissertation. It has been a period of intense learning for me, not only in the scientific arena, but also on a personal level. I would like to reflect on the people who have supported and helped me so much throughout this period.

I would first like to thank my mentor Dr.Shashidhar G Koolagudi for the guidance, support and courage he provided during the ups and downs of this tenure; being a major force that made me see the completion of my research work. I am ever grateful to him for all his encouragement that brought out the ability in me to accomplish the otherwise tough goal. I sincerely thank him for his patience that has lead me to this achievement.

Words fail me, as I thank my previous research guide, the late Dr.K.C.Shet. for having faith in me and giving me an opportunity to pursue this research under his guidance.

I express my heartfelt thanks to my Research Progress Assessment Committee (RPAC) members Prof.D.V.R.Murthy and Prof.Muralidhar Kulkarni, for their enthusiasm, their valuable suggestions and constant encouragement that consistently helped in improving the research work.

I acknowledge all the help extended by Mr.Pravin, Mrs.Nagarathna, Miss.Keerthi, Mrs.Fathima, Mr.Vishnu, Mrs.Deepa and other research scholars at the computer science and engineering department of NITK.

I sincerely thank all teaching, technical and administrative staff of the Department of Computer Science and Engineering, NITK, for their help during my research work.

I thank Dr.P.Ramanujan (Associate Director, IHLC), CDAC, Bangalore for extending excellent cooperation in this regard.

Dr.S.Janakiraman, Mr.Abhishek Srivastava, Mr.Ramesh Naidu, Mrs.Savitha Gowda, Mr.Mohit Ved, Dr.R.C.Saritha and my other colleges of CDAC, Bangalore, were all instrumental during my learning phase and beyond. I thank them from the bottom of my heart.

A special thanks to my family. Words cannot express my sincere gratitude towards my parents Mr.M.D.Ananda Naik and Mrs.Prema for all of the sacrifices that they have made on my behalf. Their prayers, unconditional support and the ever lasting faith in me is what sustained me thus far. I would also like to thank my brother, Mr.Vikas H and my husband, Mr.Narendra Babu G. Thank you for supporting me for everything, especially for encouraging me throughout this experience.

Finally I thank the almighty, for giving me the chance, the courage and the strength to endure and pursue till the end. You are the one, I always looked up to, for all my joys and trails. My trust in you will always stay firm. Thank you, God.

Place: Surathkal Vathsala H
Date:

ABSTRACT

The Indian subcontinent is mainly dependent on South-West monsoon for her fresh water needs. The variability of South-West monsoon decides the state of the economy of this region. This rainfall in the Indian context is also known as Indian Summer Monsoon. From the literature available, it may be noted that people of India have been aware of the reversal of winds over the Arabian sea from the turn of Christian era as the important information regarding the wind regime was understood by the traders due to the movement of ships for trade across the Indian Ocean. Nevertheless, since ancient times, there has been a great demand for accurate forecasting of Indian Summer Monsoon Rainfall (ISMR).

Predictors are the parameters that have high influence on rainfall patterns. In the past, several predictors have been proposed by hard-core meteorologists for ISMR prediction. However, over the times, the role of the predictors for ISMR prediction is drastically changing due to climate shift and new issues like; pollution, global warming etc. In this regard, a holistic approach is essential to use popularly available computing techniques to study the correlation between various existing predictors and to define a new set for efficient prediction. It is also true that compared to linear prediction techniques the use of recent approaches such as probabilistic and ensemble methods are expected to give improved performance. New approaches use inherent nonlinear relations among the data for better prediction. Using the techniques like fuzzy logic, a new scenario may be defined in prediction, where prediction value can be more precise and quantitative than conventional range prediction. In the present work the ISMR data provided by reputed data acquisition agencies like; National Centers for Environmental Prediction - National Center for Atmospheric Research (NCEP-NCAR), India Meteorological Department (IMD) and Indian Institute of Tropical Meteorology (IITM); over a period of 37 years (1969-2005), have been used to predict the rainfall. Data mining approaches, like correlation analysis and association rule mining, are used to identify highly influential predictors. Sophisticated state-of-the-art classifiers such as Neural Networks, Neuro-Fuzzy systems are used to predict ISMR using highly influential predictors. An approach of fuzzy logic has been applied, to quantitatively predict rainfall in a small geographical area of South India. The proposed method is used to analyse the effect of South-Western and North-Eastern monsoons on the Indian peninsular region. The findings of the present research are observed to be highly efficient, compared to the existing traditional prediction approaches including the ones used by IMD; the official government organization responsible for ISMR prediction.

Keywords : Association rule, Indian Summer Monsoon Rainfall, Statistical techniques, Closed itemsets, Cluster membership.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS AND NOMENCLATURE

AGCM	Atmospheric General Circulation Model
AISMR	All India Summer Monsoon Rainfall
AO/NAO	Arctic Oscillation/North Atlantic Oscillation
AWS	Automatic Weather Stations
BASS	Bellingshausen and Amundsen Sea Sector
BoB	Bay of Bengal
CAM	Community Atmosphere Model
CC	Correlation Coefficient
CCA	Canonical Correlation Analysis
CCs	Correlation Coefficients
CFS	Climate Forecast System
CFSv2	Climate Forecast System Version-2
CI	Central India
CMAP	CPC - Merged Analysis of Precipitation
CPU	Central Processing Unit
DJF	December-January-February
DOE	Department Of Energy

DSLP	Darwin Sea Level Pressure
DSM	Digital Surface and Terrain Models
ECAS	East-Central Arabian Sea
ECHAM	European Centre-Hamburg Model
ECMWF	European Center for Medium-Range Weather Forecasts
EEIO	East Equatorial Indian Ocean
EM	Expectation Maximization
ENSO	El Niño-Southern Oscillation
EOF	Empirical Orthogonal Function
EQUINOO	Equatorial Indian Ocean Oscillation
EQWIN	Equatorial Wind
ESDIM	Environmental Services Data and Information Management
FSUCGCM	The Florida State University Coupled Ocean-Atmosphere general Circulation Model
GA	Genetic Algorithm
GARP	Global Atmospheric Program, Global Experiment (FGGE)
GCM	General Circulation Model
GDP	Gross Domestic Product
GEDEX	Greenhouse Effect Detection Experiment
GISST	Global Sea-Ice and Sea Surface Temperature
HL	Heat Low
hPa	Hectopascal

HRR	Homogeneous Rainfall Regions
HSDSD	Historical Soviet Daily Snow Depth
HTP	Himalayan Tibetan Plateau
IA	Inter Annual
IITM	Indian Institute of Tropical Meteorology
IMD	India Meteorological Department
IMR	Indian Monsoon Rainfall
IOD	Indian Ocean Dipole
IR	Infra red
IS	Inter Seasonal
ISM	Indian Summer Monsoon
ISMR	Indian Summer Monsoon Rainfall
ISRO	Indian Space Research Organization
ITCZ	Tropical Convergence Zone
JJAS	June, July, August, and September
LDA	Linear Discriminant Analysis
LLJ	Low Level Jet
LPS	Low-Pressure System
LRF	Long Range Forecast
LWCRF	Long Wave Cloud Irradiative Forcing
MAM	March-April-May

MetUM	Met Office Unified Model
MF	Membership Function
MH	Mascarene High
MLP	Multi Layer Perceptron
MMEs	Multi-Model Ensembles
MOSDAC	Meteorological and Oceanographic Satellite Data Archival Centre
MR	Multiple Linear Regression
MSLP	Mean Sea Level Pressure
MT	Monsoon Trough
NCAR	National Center for Atmospheric Research
NCEI	National Centers for Environmental Information
NCEP	National Centers for Environmental Prediction
NEI	North East India
NEMR	North East Monsoon Rainfall
NETCDF	Network Common Data Form
NKSMR	North Interior Karnataka Summer Monsoon Rainfall
NN	Neural Networks
NPO	North Pacific Oscillation
NWI	North West India
NWIWP	Northwest India Winter Precipitation
OLR	Out Going Long Wave Radiation Flux

OV	Onset Vortex
PCR	Principal Component Regression
PISMR	Peninsular Indian Summer Monsoon Rainfall
PNA	Pacific/North American
PPR	Projection Pursuit Regression
PRWONAM	Prediction of Regional Weather with Observational Meso-Network and Atmospheric Modeling
QBO	Quasi Biennial oscillation
QRP	Quantitative Rainfall Prediction
RMSE	Root Mean Squared Error
RV	Rainfall Value
SC	Subtractive Clustering
SD	Standard Deviation
SDR	Sub Division Rainfall
SIE	Sea-Ice Extent
SLP	Sea Level Pressure
SMR	Summer Monsoon Rainfall
SO	Southern Oscillation
SOI	Southern Oscillation Index
SPI	South Peninsular India
SPMF	Sequential Pattern Mining Framework
SR	Standardized Rainfall

SST	Sea Surface Temperature
TEJ	Tropical Easterly Jet
TH	Tibetan High
TPSC	Tibetan Plateau Snow Cover
TRMM	Tropical Rainfall Measuring Mission
URL	Uniform Resource Locater
VHRR	Very High Resolution Scanning Radiometer
VIS	Visible
WCI	Western Central Indian
WEIO	West Equatorial Indian Ocean
WMO	World Meteorological Organization

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

The word "Mausim" in Arabic language means season. Mausim is the word from which the word Monsoon is derived. Monsoon represents the seasonal reversal of winds, accompanied by changes in rainfall patterns. The general mechanism of monsoon is the shift of the global wind patterns that result in heavy rains in the tropical and subtropical parts of the world.

Depending on the direction of the winds; there are two types of monsoons, they are called the South-West and North-East monsoon. In the Indian context, the moisture bearing winds that cross the equator enter India from the South-West direction. Hence, the rains brought by these winds are called the South-West monsoon. The physics of the South-West monsoon is shown in Figure 1.1. At the end of this season (September end); due to reversal of the pressure systems, the wind blows back from the land towards the water bodies surrounding India from the North-East direction. Thus, the rains brought about by these winds are called the North-East monsoon. This phenomenon is shown in Figure 1.2.

The main mechanism that sets in the monsoon is the high pressure and the low pressure systems. The mechanism of South-West monsoon proceeds in three stages, namely: onset, progress and withdrawal. Starting from the high pressure at the Mascarene region in the southern Indian Ocean, the moisture laden winds

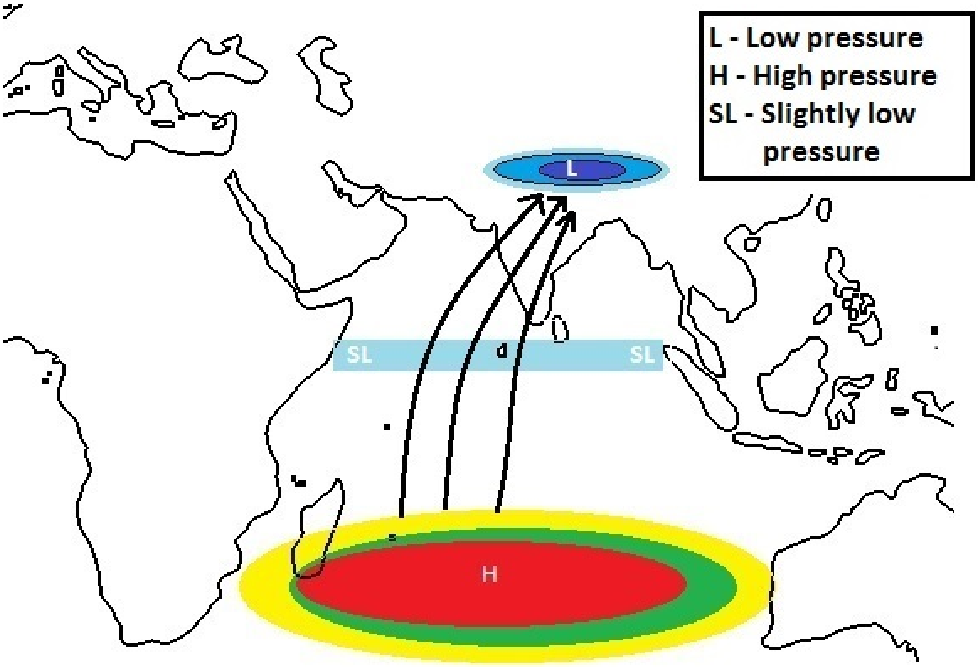


Figure 1.1: Mechanism of South-West monsoon

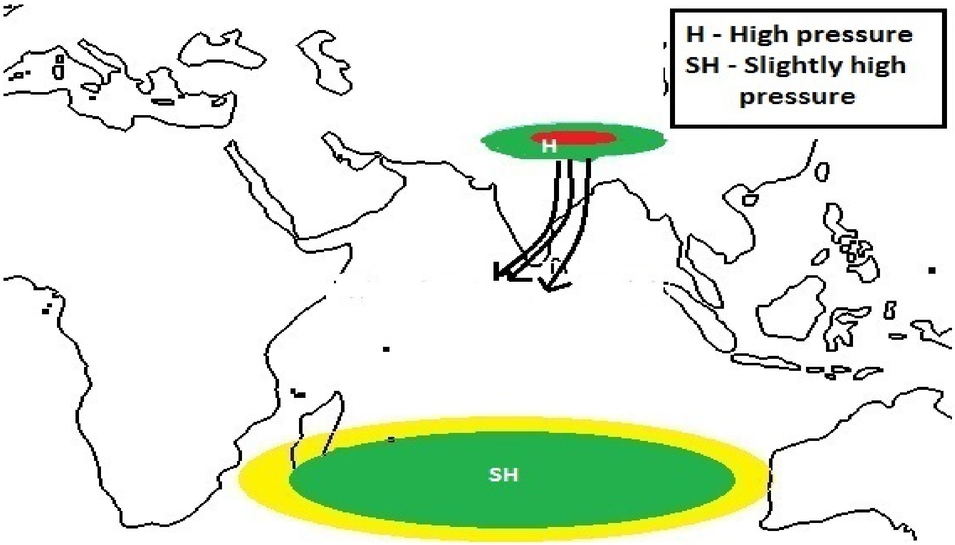


Figure 1.2: Mechanism of North-East monsoon

are directed towards India. Guided by the slightly low pressure at the equator indicating the onset of Indian monsoon, it enters India to reach the low pressure zone at the northern plains. These winds give rains all along the path of their travel. This phase of monsoon is called the progress. Due to the reversal of pressure systems at the end of the South-West monsoon season (June to September) the dry winds from the land start blowing towards the Indian Ocean. A part of these dry winds on their path, pick up moisture from the Bay of Bengal to give rains to Tamilnadu and some parts of Orissa, Karnataka and Kerala. This phase of monsoon is called the withdrawal of Indian monsoon. Since North-East monsoon gives rains to a small part of the country it is not given proper attention as compared to South-West monsoon.

Majority of the rain in India is from the South-West monsoon, also known as Indian summer monsoon. The variability of South-West monsoon decides the state of the economy of this region. Agriculture is a major occupation in India. Hence, the general public of India has been curious about monsoon year after year. From literature it is seen that Indians have been aware of reversal of winds over the Arabian sea during the turn of Christian era (according to Monsoon Monograph by IMD). The important information regarding the wind regime was understood by the traders due to the movement of ships for trade over the Indian Ocean. This indicates that since ancient times, there has been a demand for accurate forecasting of Indian Summer Monsoon Rainfall (ISMR).

"Weather prediction" is the use of technology for finding the state of the atmosphere at a particular point of time in the future. Using science and technology to build smaller representation of climate for forecasting are called weather forecasting models. Accurate weather prediction requires good predictors and best prediction models. Variables in the climate scenario, that can contain accurate information in order to provide proper results while using technology for prediction are called predictors. Predictors are synonymously called as selected features or attributes. There are many related variables that contribute to ISMR, this may be understood by analyzing different theories behind ISMR. Some of the important theories are: difference in the specific heat capacity of land and water giving rise to sea breeze (Bala Subrahamanyam et al., 2001). Low pressure points are formed on land as a result of different heating levels of sea and land, called thermal theory or the differential heating of sea and land theory (Clemens and Prell, 1990). This causes flow of moisture laden winds from the sea to land. The dynamic theory of monsoon explains monsoon on the basis of the annual shifts in the position of global belts of pressure and winds. The jet stream theory (Ramaswamy, 1962) explains monsoon in terms of ridges and troughs caused due to warm and cool air. The air mass theory is based on Inter-Tropical Convergence Zone (ITCZ). The South-East trade winds and the North-East trade winds in the northern hemisphere come together at the equator. The place where they meet is known as the ITCZ. Many more theories have been derived to explain general mechanism of monsoon. All these theories hold their importance in ISMR.

It is well understood that there are number of variables in the climate scenario that can become potential predictors of ISMR. There are many techniques that have been discovered in recent times, for predictor identification. Thus selection of an apt technique for the current scenario becomes an important criterion. The careful selection of predictors and techniques are important for the development of a good prediction model so as to provide accurate forecasting results.

ISMR forecasting has a long history. It all started with the understanding of the reversal of winds over the Arabian sea. This information was utilized by the mariners for their trade. Halley (1686) was the first to investigate the mechanism of monsoon. The scientific studies on ISMR started in the 19^{th} century by the East India commission employees. The outcome of these studies is the discovery of the existence of diurnal variability of pressure over a large part of South Asia and the structure of the tropical cyclones in north Indian Ocean. In 1875 India Meteorological Department (IMD) was established and it systematized the studies on South-West and North-East monsoon. Since then, IMD has discovered different predictors, hemispheric oscillations; leading to the knowledge of teleconnections relating to monsoons, evolution of models for prediction etc. This history shows that the continued effort in this area has led to the understanding of the monsoon mechanism over India, to some extent. While ISMR prediction research has been carried out worldwide, the behavior of the ISMR is only partially understood and has been consistently difficult to predict. Despite all efforts, ISMR prediction model development is considered to be still in its infancy and there is wide scope for the development of new prediction models and predictors (Kashid and Maity, 2012).

1.2 IMPORTANCE OF RAINFALL PREDICTION

Indian Summer Monsoon Rainfall is a complex phenomenon; involving atmosphere, land, ocean and many other domains. The backbone of the Indian economy is agriculture. About three-fourths of the working population in India depends on agriculture for their livelihood. The main source of water for food grain production is from ISMR. ISMR significantly impacts agricultural production in India, thereby affecting India's economy. ISMR is one of the most anticipated weather events for global monsoon researchers. Prediction of Indian monsoon rainfall variability is a crucial factor not only for agricultural output, but also for water resource management, hydro electricity generation, tourism, rain water harvesting plan and so on. In general ISMR has a significant impact on the overall well-being of the subcontinent's residents. In this research we focus on the rainfall patterns in India due to South-Western (summer) monsoon.

Monsoon is a tropical phenomenon, mainly seen in countries that are close to the equator, facilitating deferential heating of land and water bodies. The mechanism of South-West monsoon is during the months of June to September. Moisture bearing winds from the South-West direction of the Indian Ocean blow onto the landmass. They split into two branches; the Arabian sea branch and the Bay of Bengal branch, near the southernmost tip of India. They give maximum rainfall to India. Heavy rainfall basically occurs due to the many water bodies surrounding India. The high landmasses like Western Ghats and Himalayas in the path of the moisture bearing winds are also instrumental in bringing orographic angle to the monsoon. Prediction in this scenario becomes especially difficult because of the complex and unique geographical location of the area considered. The complexity of prediction further increases because of the addition of geophysical, atmospheric, and oceanic components. In spite of all these complexities, attempts have been made to develop good models for the prediction of ISMR.

1.3 MOTIVATION

Literature gives an account of tremendous research being conducted in the field of Indian summer monsoon but, very less research is reported in the scaled down spatial units (like homogeneous rainfall regions). Even with good information on predictors related to ISMR, the efficiency of predictors and their influences may vary due to the changes in climatic phenomena, global warming, deforestation, pollution and various other factors. Hence there is always a need to identify effective predictors from time to time. Since Global Circulation Models (GCMs) are not very effective, researchers have preferred statistical models. However, predictors play an important role in the success of statistical models. Including good predictor and the use of improved statistical techniques are significant for accurate forecasting results.

Based on the available literature, every experiment related to Long Range Forecast (LRF) uses only a few selected predictors. Common among them are the El Niño-Southern Oscillation (ENSO) indicators. However in reality the influence of a very good predictor on ISMR is not constant. It is a fact that some variables show good links to ISMR during some years, while some variables which are completely different, show their relevance to ISMR during other years. The question that arises here is what are the variables that have to be always commonly considered for the prediction of ISMR with good accuracy. The improved techniques in data analytics, increased computing power and cost effective digital storage capacity can be used effectively in accurate prediction of ISMR. Inclusion of every relevant predictor in the prediction of ISMR, is expected to give accurate prediction. However the approach is impractical. Hence, selecting the predictors, based on correlation and their mutual relationships with other predictors in the past, can help selecting the combination of predictors that strongly influence ISMR. The motivating factors mentioned above have encouraged us to carry out research in the area of long range prediction of ISMR. This important endeavor aims to solve some problems related to shortfalls identified in this area.

1.4 APPLICATIONS

Weather forecasts are important warnings as they help prosper life and property. For proper planning, there are many sectors which require rainfall prediction. Weather forecasting and rainfall prediction have their applications in the following areas. In agriculture (Jones et al., 2000), specific weather conditions are desirable at each stage of crop growth, right from sowing seeds to growth of the plants; from harvesting to processing the harvested crops . Rainfall forecasting is highly desirable for other agricultural activities as well. Water security of the livestock can also be taken care off with accurate forecasting assistance. There are many states in India that experience large number of farmers'suicide due to crop failure, as a result of poor monsoon. Local government can announce crop support measures, crop insurance schemes and also undertake artificial rain seeding techniques based on weather forecasts. Flood warning is another important aspect considered under weather forecasting (Pappenberger et al., 2008). Flood warning, given in seasonal forecasts, can alert the government to undertake various measures to prevent loss of life and property. Measures like creating artificial reservoir behind a dam to regulate water flow, diversion of a part of the peak flow to another river or basin, to save the region from significant damage, channel and drainage maintenance and improvement works, which artificially reduce the flood water level etc., can be planned in advance. Drinking water management also benefits by a great deal from LRF of rainfall. Reservoir water level increases in case of excess rains and decreases with drought conditions. Ground water table also depletes during deficit rainfall conditions; leading to drinking water scarcity. Water sharing among states through rivers and distribution of water to public can be planned in advance in accordance to the weather forecasts. Hydro, wind and solar power generation also require accurate rainfall forecasting for proper planning of power distribution (GarcÃŋa-Morales and Dubus, 2007). Wind speed and direction, the possibility of cloudiness and the amount of water that can be collected in the reservoir can give an idea of power security for the following year. Rain water harvesting is crucial in all above mentioned fields of application, thus relating itself to accurate weather forecasting. Recent and earlier studies have shown outbreak of infectious diseases; due to mass generation of microbes and vectors, they exploit the disrupted social and environmental conditions of extreme weather events. When an extreme weather event is expected, prior preparations can be made in order to prevent such outbreaks with the aid of weather forecasts (Thomson et al., 2006). Township maintenance before a rainy season can prevent havocs caused due to water drains filled with silt, open gutters, low lying tunnels etc. The prediction of onset of rainy season plays a major role in township maintenance. With ever increasing air traffic and demand for air space, delay of flights due to weather conditions are not tolerable. In view of this, weather forecasting of different air routes can help plan alternatives for the season. Heavy thunder storms cause severe turbulences and are a problem for an aircraft in flight as strong winds can damage the aircraft on flight. To avoid flight hazards, accurate rainfall predictions are desirable (Steiner, 2009). Marine navigation industry; be it commercial or tourism related, faces problems due to heavy precipitation and storms with strong winds. Fishermen, who use small size boats or ships also come under the threat of violent weather during monsoons. Alternative route plans and warnings can be given in advance with the help of accurate forecasting (Makela et al., 2006). Climate is a key influence on travel planning and travel experience, as, all tourist destinations are climate sensitive. Tourism services are currently available in every climatic zone on the planet, be it deserts, high mountains, tropics and polar regions. Since outdoor activities are severely restricted by heavy rain, snow and wind, forecasts can be used to plan activities around these events. Owing to the heterogeneity in applications, it becomes necessary to incorporate accurate weather forecasting services into decision-making in the various domains mentioned above.

1.5 EARLY HISTORY OF RAINFALL PREDICTION IN INDIA

The severe famine of 1877 due to failure of monsoon encouraged Blanford (1884) to issue tentative forecasts. Based on the relationship between winter and spring snow fall over Himalayas, ISMR forecasts were issued from 1882 to 1885. Later in June of 1886, the first operational forecast was issued for the entire India and Burma. Techniques like analogue and curve parallels were used by Sir John Eliot for LRF. The variables used were based on the weather conditions over entire India and its surrounding regions. Later, the forecast variables used for LRF were (1) Himalayan snow cover (Oct-May), (2) local peculiarities of pre-monsoon weather in India and (3) local peculiarities over the Indian Ocean and Australia. Due to the failure of LRF for the drought years 1899 and 1901, systematic studies were undertaken to understand the complexity in forecasting monsoon rainfall. The first statistically aided LRF of 1909 was the extensive effort of Sir Gilbert Walker (Walker, 1923, 1924). This forecast was based on regression methodology. The concept of correlation was introduced into LRF in this period. Since then new forecasting models have been experimented upon and new revised models have been recommended form time to time. The need for accurate prediction of rainfall has constantly driven the model evolution. This evolution is constantly fueled by; the failure in LRF of a particular year.

1.5.1 Traditional techniques

Though LRF models started with subjective methods like the analogue and curve parallels, statistical and empirical techniques have been dominating the field of ISMR forecasting for decades. Systematic forecasting initially started with regression models, used for forecasting rainfall for entire India. Regression was later used in forecasting for the three homogeneous rainfall regions namely: (1) Peninsula, (2) North-East India and (3) North-West India. These regression models had inbuilt limitations, for example lesser number of predictors. Verification of these forecasts

(1924-87) revealed that only about 63% of these forecasts were correct (according to Monsoon Monograph (Volume 2)). New LRF operational models were built to overcome the limitations faced with regression models. These new models included collective use of large number of predictors, based on new techniques like dynamic stochastic (Thapliyal, 1982), power regression and parametric models (Gowariker et al., 1989, 1991). Subsequently, various other models like principal component regression (Singh and Pai, 1996; Rajeevan et al., 2000), canonical correlation analysis (Rajeevan et al., 1999; Prasad and Singh, 1996), neural network (Navone and Ceccatto, 1994; Goswami and Srividya, 1996; Guhathakurta et al., 1999) and power transfer models (Thapliyal, 2001) were developed. Also the number and variety of variables used in LRF continued to evolve, the spread of variables from different parts of the globe were included in the models. Further evolution of the traditional techniques led to tremendous improvement of prediction capabilities. However each model gave good results for a few years, but subsequently have failed miserably, thus needing improvement in techniques and predictor selection.

1.5.2 Evolution of traditional ISMR forecasting techniques

The reasons for evolution of the rainfall forecasting models both in terms of variables used and the techniques employed are due to the inherent and evolutionary limitations they exhibited from time to time. In the early years, only the peculiarities of the local climate have been used as variables in models. Lack of documentation and maintenance of records posed problems due to the insufficient data available for accurate results. The human analytical knowhow & capability put together was the only processing power available. Huge calculations, with large data analysis were time consuming, and hence on time forecast were almost out of question. Later, due to colonization, global surface and oceanic data were available and proper record keeping was practiced, leading to the discovery of links between global and local weather conditions. Large number of variables were also available, but the limitations still existed in computing powers. Hence small number of most correlating variables were used for LRF. As years passed there has been an improvement in computation power, leading to improvement in forecasting accuracy. However, due to modernization, climate shift and natural calamities, relevance of variables discovered so far, as having links with ISMR, have been losing their correlation and some of them have lost links completely. The accumulated data of variables are in abundance, but the initial reasons of their recording has not been for the purpose of LRF. Though some data was available, it was not suitable to be used directly in LRF. Hence, new techniques for data preprocessing, cleaning etc, had to be developed. Though the processing power of computers has increased, the digital storage capacity were limited, thus posing problems for large data analysis (RAM problems). The latest advancements in computer processing powers, storage revolution, cloud computing platforms and discovery of new techniques in data analytics can be used to an advantage of LRF performance.

1.6 SOME EVOLUTIONARY PROPOSALS FOR PREDICTION OF ISMR

When we speak about a complex phenomenon like monsoon, we are aware that there are many variables involved that have either very strong or weak correlations to it. Sometimes, variables may not be related to monsoon, but are present in the climate scenario. A question in this context is to understand, what variables are to be considered for predicting or formulating association between the predictors, so as to get better results. When there are many predictors, relationships between these predictors and rainfall is not well defined or understood. For accuracy in prediction, few best predictors have to be picked based on some technique. Techniques available to shortlist some of the best predictors are step wise regression (Wilks, 1995), a cross validation scheme suggested by DelSole and Shukla (2002) and association rule mining (Agrawal et al., 1993) etc. Association rule mining has found many applications in the field of feature selection, where many correlated features coexist. Feature selection, based on association rule mining, uses the correlations between variables involved to select only those features that are closely related to the subject of study. This assures a better prediction capability in a good performance bracket.

The literature has reported various predictors affecting ISMR. However, no research has used all of these known predictors for ISMR prediction as there are computational and storage constraints. Use of many predictors provides rich information for any decision, when different categories of the same subject have to be predicted. It provides robustness against some instability that may arise due to less number of predictors. In such cases, different subgroups of predictors may relate to different categories. In addition, many predictors can bring forth nontrivial relationships that may further boost prediction capabilities of the model (Gago and Bento, 1998).

Predictors are the soul of any prediction model. Using apt predictors, suitable to the scenario in hand, is crucial in proper prediction. From literature it can be seen that new predictors of ISMR have been suggested using Correlation Coefficients (CCs) calculations. The advantage of using Correlation Coefficients for detecting predictors, from the whole dump of data available, is that one can quantitatively show strength of relationship between two variables. However, it cannot show cause and effect relationship between variables. Whereas, association rule mining may show cause and effect relationships. Hence, combining CC calculations to unearth possible predictors and then applying association rule mining on the predictors given by CC calculations may be an effective method for suggesting new predictors in areas where unearthing new predictors is necessary.

Since the discovery of ENSO indices it was assumed that these indices have a major role in accurate prediction of rainfall in India. Hence, researchers have been concentrating only on ENSO indicators even in prediction of rainfall of smaller geographical units in India. The major short falls in accurate prediction of rainfall in smaller geographical units can be attributed to this reason (Kashid and Maity, 2012). In general it is seen that there exist a relationship between global weather conditions and the local land surface & oceanic conditions. Hence, it is necessary to include local predictors in prediction of rainfall in any geographical unit. Due to the large geographical diversity of Indian landmass, diverse rainfall behaviors are seen in different geographical units. These rainfall behaviors can be accurately predicted by detecting apt predictors for each geographical units. Hence, there is a need for exploring new predictors whereever necessary.

Prediction of rainfall in ranges (Drought, Deficit, Normal, Excess and Flood) is a commonly followed practice in LRF of rainfall. This task by itself is a difficult and complex goal to achieve. Since there has been improvement in the LRF accuracy in recent years, we can further set our vision on crisp quantitative value prediction of rainfall. For this task it is necessary to find quantitative values of rainfall in between ranges. Whenever the term, in-between two values is herd, Fuzzy logic is the technique that is suggested by experts. Fuzzy logic works, based on human ability of making decisions, by evaluating the membership of an element in the degree of truth. Hence, fuzzy logic is apt for the goal of achieving quantitative value prediction.

1.6.1 Main highlights of research investigations

1. A comprehensive analysis of literature on data sources, predictors, and models of prediction is presented. It features review on types of data sources used in LRF of ISMR. A list of database providing data for research with references and Uniform Resource Locater (URL) links are provided. An account of all the semi permanent systems that bring about monsoon in India is discussed. Possible predictors affecting onset, progress and withdrawal of monsoon is listed. The evolution of statistical models & the available GCMs and their critical review is carried out.

2. The outcome of the literature review is a list of gaps identified, throwing light on the need to find the cause & effect relationship in predictor selection, need for new predictors, improving statistical models and GCMs, requirement of quantitative value prediction and so on.

3. The cause and effect relationship between variables is effectively used in predictor selection, using association rule mining. This becomes necessary to find the combined effect of predictors on the predictend. This works

better than using only CCs for predictor selection, that only depend on the capacity of each predictor relationship with the predictend.

4. Better accuracy of range prediction of ISMR and Peninsular Indian Summer Monsoon Rainfall (PISMR) is achieved by using association rule mining for predictor selection, clustering methods for dimensionality reduction. In the case of PISMR, the influence of both South-West and North-East monsoons in unearthing new predictors is studied. Based on these studies, new predictors are suggested for PISMR and North Interior Karnataka Rainfall (NKSMR) prediction.

5. Fuzzy logic and Neural networks are experimented for quantitative value prediction of rainfall. Fuzzy logic is used for effective decision making and Neural network is used for learning from the data. An application of Neur-Fuzzy system is demonstrated in prediction of NKSMR.

1.6.2 Brief overview of thesis contribution

The major contributions of this thesis are summarized in the following subsection.

A critical review on the types of sources of data, preprocessing techniques required before using them for research, advantage and disadvantages of using these data sources is conducted. It also gives a list of databases along with the climate variables information provided by various organizations. The monsoon mechanism is explained with its teleconnections, to identify climate variables that can become potential predictors, under the onset, progress and withdrawal of monsoon in India. The predictors are categorized, giving a picture of how global and local conditions relate to bringing about monsoon to India. The techniques used in model building along with the different GCMs of the world that are being used for LRF of ISMR and their performance and scope for improvement are discussed in detail. Finally prediction of ISMR covers the traditional techniques, their evolution with reasons behind their failure and success along with the gaps derived from the critical literature survey for better ISMR prediction, is presented in the literature review.

A mix of climate variables, representing different categories like; the oceanic, land surface and atmospheric conditions found in past literature, are employed in ISMR prediction. These climate variables are subjected to data mining technique called closed itemset mining, that find frequent itemsets in order to give hidden causative relationships from the past. These causative relationships are detected by association rules, specifying predictor combinations that are capable enough to predict rainfall with accuracy. Intermediate data processing techniques are used for reducing dimensionality of the data. Finally a classifier is employed for predicting the range of rainfall of a particular year.

A geographical unit called the Peninsular India is influenced by North-East and South-West monsoon. Hence, a study on the influence of pre South-West and pre North-East monsoon weather conditions has been undertaken, to find possible predictors that effect August, September rainfall of this region. Since there is a very less reported work in literature, pertaining to this region, CC calculations were employed to study and suggest new predictors that were later subjected to association rule mining for predictor selection. Feed forward Neural networks have been employed as prediction tool.

An attempt has been made to predict rainfall values instead of ranges using soft computing technique. The changes introduced by the environment are learnt by the Neural networks in order to enforce effective decision making using fuzzy logic. A set of new predictors for North Interior Karnataka Summer Monsoon Rainfall are explored and suggested, based on correlation analysis. Thus an ensemble technique involving association rule mining, clustering and a neuro-fuzzy system is applied on NKSMR data for demonstrating the results of quantitative value prediction.

1.7 OUTLINE OF THE THESIS

The thesis is spread across 6 chapters. The following paragraphs broadly explain the contents of each chapter.

Chapter 1: Introduction : The first chapter of the thesis introduces the problem chosen to address. In brief, this chapter begins with a necessary background of ISMR covering the creation and onset of monsoon for India. The discussion is on importance of ISMR, with a list of applications from various perspectives leading to the motivation for the work carried out. Further the reasons for the difficulties and challenges of ISMR prediction are mentioned. We also have discussed in brief the evolution of ISMR prediction techniques over the years. Chapter 1 ends with clearly articulated research contributions jotted down as thesis outline.

Chapter 2: ISMR prediction: A Review : This chapter contains information about the state-of-the-art literature on Indian summer monsoon specific features in the context of source, predictors and model aspects of forecasting. A detailed list of available database is included. Gap areas discovered as a result of literature review is enumerated and discussed. The general database used in this research work is depicted. The scope of present work derived from review is presented.

Chapter 3: ISMR prediction using data mining and statistical approaches: In brief, the chapter presents a general introduction to ISMR. It mainly concentrates on an algorithm developed for All India Summer Monsoon Rainfall prediction. The association amongst variables are exploited for better predictor selection. The specific database used and the ensemble of techniques used for processing the data and their functionalities are discussed in detail. The assessment of the accuracy in prediction of the ISMR is presented at the end of the chapter.

Chapter 4: Rainfall prediction in homogeneous rainfall region: This chapter presents in detail a model developed for homogeneous rainfall regions rainfall prediction and is applied for prediction of PISMR. The role of correlation analysis and association rule mining for the predictors exploration with meteorological basis for variable inclusion are discussed. The role of newly suggested predictors in PISMR prediction is examined and justification are given in the results and discussion section. This section also presents the performance of Neural networks along with a clustering technique in prediction of PISMR.

Chapter 5: Quantitative long range rainfall prediction - A fuzzy logic approach: This chapter deals in detail, an algorithm developed for quantitative rainfall prediction and its applications on NKSMR data. The chapter describes the use of two main techniques employed for learning and decision making namely, adaptive neural network and fuzzy logic. A few local condition predictors are newly suggested and the performance of the algorithm in NKSMR prediction is presented at the end of the chapter.

Chapter 6: Summary and conclusions: summarizes the contributions of this thesis along with some important conclusions. This chapter also throws light on some of the issues for further research.

CHAPTER 2

LITERATURE REVIEW

In Chapter 1, the general background of ISMR with a brief account of the evolution that has taken place, in the field of LRF of ISMR is addressed. This chapter details the process of evolution of LRF of ISMR with respect to data sources, databases, predictors, models and prediction of ISMR. A list of research gaps are derived from the critical review and are listed at the end of the chapter.

MONSOON is notable at the tropics. It is seen in all the tropical countries and tropical oceans. Most of the countries that come under the influence of monsoon are developing countries and their main occupation still remains agriculture. These countries mainly depend on the monsoon for their agricultural activities. Hence, the prediction of onset of monsoon, the total amount of rainfall per day, per season, is of utmost importance. The slightest variation of monsoon rains can have a greater impact on the life of the population in the tropics. Monsoons are the main sources of water also for drinking, hydro electric power generation, rainwater harvesting programs etc. There has been research in the relationship of summer monsoon rainfall with Rabi and Kharif crops. It is seen that the total food grain yield over India during Kharif (summer) season is directly affected by variations in the summer monsoon. Also summer monsoon precipitation influences the Rabi crop through water and soil moisture availability (Preethi and Revadekar, 2009). Importance of ISMR further becomes evident with the review of literature in the current chapter.

Literature review gives a consolidated account of the published research in the field of interest giving the readers knowledge about the previous and ongoing research in the field. With the consolidated knowledge available the strengths and weaknesses are summarized. This facilitates a researcher to formulate problem that can effectively address the gap areas in the considered field. A combination of a database, which has good coverage with highly influential predictors, where individuals or their combinations can significantly correlate to rainfall and an effective learning model, are crucial for producing correct estimate of rainfall. Therefore, a literature review on the database, predictors and models used for forecasting ISMR is undertaken.

2.1 DATA SOURCES: A REVIEW

Here we discuss different data sources related to climate and ISMR prediction. Sources of meteorological data can be broadly classified as observed data and model data. Observation is an active acquisition of information from the primary source. Observed data are prime data as they are measured in real world scenario. In science, it involves recording observed information using instruments. A model is a small scale imitation of a large phenomenon. Data obtained from such models are known as model data and models are secondary sources of data. Sources of data are important as they need different preprocessing steps and they pose different nature of problems during their practical use.

2.1.1 Observed data

Observed data is collected from a variety of observation systems like polar-orbiting satellite, geostationary satellite, aircraft, radiosonde, baseline air pollution station, satellite ground station, automatic weather stations, automatic river-height and rain gauges, wind profile, observation ships, drifting buoy, meteorological observation station, over the horizon radar (Plummer et al., 2003). The data registered by these systems are in different forms. Depending on these forms, they can be grouped either as images or numeric data. Satellite and radar images are the important catagories of images, whereas sensor based observed data are numeric in nature. Each of these forms can be used separately or in combinations as data sources for forecasting.

Sensor-based observation data

Generally sensor-based observation data provide accurate measurement, but most observing systems may not have been developed with a climate objectives in mind. As a result, tremendous efforts have gone into assessing and reprocessing the data records to improve their usefulness in climate studies (Bosilovich et al., 2013). They suffer from sampling error while representing area means and are not available over most oceanic and unpopulated land area (Plummer et al., 2003). The data faces inconsistencies because of periodical replacement of sensors, maintenance shutdowns, and malfunction due to natural calamities. In such cases, some of the related near by variables are used to calculate the value of unknown variable, causing magnification of point errors in calculation. As a result there are some missing or erroneous data. In addition to this, if an observing station has incomplete data, and longer averages are used to fill the gaps, then, they may contain some kind of bias depending on the weather characteristics of the missing period. Changes in observation station such as station relocation, change in instrumentation or change in observing practice (changing observation time) result in an abrupt step change in the time series resulting in either colder or warmer readings. Changes in the landscape around an observing site also causes non-climatic changes in climate data. Urbanization is one clear example that has caused much concern and uncertainty in climate trends, particularly in terms of temperature (David, 2008). These non-climatic changes are most commonly referred to as inhomogeneities or time-dependent biases. These data lead to increase of uncertainty in any conclusions drawn.

Satellite images and Radar images

Satellite images are not mere photographs. They are the representations of various electromagnetic radiations measured by the sensors on the satellite. Satellites take images on visible spectrum, infra red, near infra red etc. These spectrum sensors are responsible for measuring different radiations coming from the Earth. These are used in differentiating land, sea, thick and thin cloud, height of clouds, cloud top temperature, Sea Surface Temperature (SST), thunderstorm intensity, presence of fog, low clouds etc. Majority of the cloud detection techniques in satellite imagery make use of the property of high reflectance of cloud in the Visible (VIS) spectrum and/or the low temperature in the Infra red (IR) spectrum (D'souza et al., 1990) (Saunders and Kriebel, 1988).

Radar images are formed when Radars send out electromagnetic waves as short pulses, which may be reflected by objects in their path; in part, reflecting back to the radar. For example - if the radar is used to detect precipitation; when the pulses intercept precipitation; part of the energy is scattered back to the radar. From this information the radar estimates the region precipitation and quantity as well. Both Satellite and Radar imagery are characterized by proper deformations and noise due to the different acquisition geometries and processes. These deformations have to be duly taken into account during the Digital Surface and Terrain Models (DSM) generation procedure, in order to fully exploit the potentialities (Capaldo et al., 2012). Satellite imaging systems, collect and downlink large amounts of data. Associated ground systems, further process, store and disseminate this data. Limitations on computer storage, transmission bandwidth, transmission time, and digital display resolution may restrict the amount of data used to represent an image. These issues affect image processing and storage on-board on the satellite (Ellison and Milstein, 1995).

Satellite and radar images are mainly being used in many multi disciplinary ISMR researches. The advantage is that they can be better understood and interpreted due to visual nature. These are used in studying the association between climate and agriculture, climate and traffic monitoring, climate and landscape changes etc. When combined with sensor-based data; image based systems give better results and can span across several domains.

2.1.2 Model data

Climate models are the mathematical representations of the climate. They use quantitative methods to simulate the interactions among the atmosphere, oceans, land surface, and ice. They are used in the study of weather dynamics, weather forecasting etc. Climate models provide solutions that are discrete in time and space, which means that the results obtained represent averages over regions and have a specific time frame. Spatial resolutions may be global, zonal or in terms of numerical grids with less than 100km resolutions. Time may be in terms of days, months or years. The main causes of uncertainties are model trade offs, errors, and the effects of parameterizations (Shackley et al., 1998). Although each climate model has been built to reproduce observational means with some measured compromises, there are many valid reasons for the climate modeling community being far behind in this important endeavor. Each model contains slightly different choices of model parameter values, as well as different parameterizations of under-resolved physics. Model uncertainties are initial condition uncertainties that provide a bias in projection. Initial conditions are often dismissed as a source of uncertainty in climate projections since, the results from model outputs are usually averaged over decades. Hence, because of limited computing power, computer models can represent climate processes and their interactions, only up to a certain spatial and temporal scales (Nychka et al., 2017).

Climate patterns on small scales, like scaled down spatial area, are not depicted by global models which have widely separated grid points. It is still not known whether a particular information is important for climate, and therefore, it is not represented in the model. Models differ in the way they represent climate feedbacks. This is perhaps the most important reason why models have different values of climate sensitivity. Detailed and varying topography and heterogeneous physical process are reduced by the models to a single grid point in the typical resolution (e.g. 100km X 100km) for a global climate model, which has also been the reason for prediction errors.

Table 2.1: Types of meteorological data sources, their advantages and disadvantages

Sl.no	**Type of data source**	**Advantages**	**Disadvantages**	**Remarks**
1	Sensor-based observation data	• Well defined and known pre-processing steps available. • Open source Data available easily for wide variety of domains. • Results obtained are more accurate and close to reality. • A Lot of applications are available that can be considered as baseline examples. • Image processing techniques are available for pre-processing. • Coverage is wide on geographical areas with land, ocean, sea, atmospheric levels related data. • Well documented and well maintained data.	• Observational error must be carefully handled in per-processing steps.	• Measured in real world, hence can be used for deriving hypothesis.
2	Satellite and Radar images	• Image processing techniques are available for pre-processing. • Well defined and known pre-processing steps are available. • Open source Data available easily. • Quite a few applications are available. • Coverage is wide on geographical areas with land, ocean, sea, atmospheric levels related data are available.	• Large amount of memory usage. • Resolution problems- due to limitation in storage requirement. • Target detection problems exists due to similar colour or texture of different objects in the image.	• Measured in real world, hence can be used for deriving hypothesis.

Table 2.1: Types of meteorological data sources, their advantages and disadvantages

3	Model data	• Image processing techniques are available for pre-processing. • Well defined and known pre-processing steps are available. • Open source Data available easily. • Quite a few applications are available. • Coverage is good. • No missing data, models generate data of all variables. • Well documented and well maintained data.	• Distributed in gridded binary format, requires software to visualize and read the data. • Domain knowledge or prior knowledge is required in selecting initial conditions so as to get good results.	• Results may not be accurate due to model uncertainties hence not an option for deriving hypothesis instead can be used for analysis of model.

These models can be advantageous while analyzing several versions of the same model in terms of improvement in successors when compared to their predecessors. Analysis in terms of which part of the global scenario is correctly or wrongly modeled, what are the effects of the wrongly modeled scenarios on other variables in terms of finding some relationships, are of importance but not known. Models can also be used for comparing how well the model works to simulate real world scenario, leading to studying the causes of model errors.

A tabular representation of the meteorological data discussed with their advantages and disadvantages are listed in Table 2.1.

There are many national and international organizations that support researchers through their data in order to improve climate research domain. From literature it may be seen that rainfall data available is either the mean of the rainfall data for the whole of India (Munot and Kumar, 2007) or area weighted means. Both types of means have their own advantages and disadvantages. Parthasarathy et. al. has come up with a spatially coherent monsoon rainfall series for the largest possible area, creating area weighted means to form monthly and seasonal homogeneous Indian monsoon rainfall series for the period 1871-1990 using 306 stations (Parthasarathy et al., 1993). This data is available at : https://data.gov.in/catalog/all-india-area-weighted-monthly-seasonal-and-annual-rainfall-mm. Later Sontakke et. al. created the longest instrumental rainfall series for well spread 316 rain gauge stations on the similar basis till 2006 (Sontakke et al., 2008). These varieties are derived from the observed rainfall.

Reanalysis, systematically produces datasets for climate monitoring and research. A frozen data assimilation scheme and models are used for creating reanalysis data. The raw data from radiosonde, satellite, buoy, aircraft, ship reports etc. are used as inputs to create a dynamically consistent estimate of climate state at each time stamp. For example Climate Prediction Center (CPC) - Merged Analysis of Precipitation (CMAP) is obtained by measuring the observations from rain gauges and are merged with precipitation estimates from several satellite-based algorithms (infrared and microwave) at 2.5° grid points for the duration of January 1979 - December 2002 (Xie and Arkin, 1997)(Xie and Arkin, 1996). Another observational reference for rainfall is taken from the India Meteorological Department (IMD) for the Indian summer monsoon season. The dataset is based on 2,140 stations across the Indian landmass. It has a resolution of 1° X 1° latitude and longitude. The procedure used in the development of this grided data is documented (Rajeevan et al., 2006). This observed rainfall data series is from 1951 to 2008. Indian Institute of Tropical Meteorology (IITM) hosts ISMR data from 1871 to 2014 (Parthasarathy et al., 1995). The IITM also hosts the Longest Instrumental Rainfall Series of the Indian Regions (1813-2006), Homogeneous Indian Monthly Rainfall Data Sets (1871-2014), Homogeneous Indian Monthly Surface Temperature Data Sets (1901-2007), Three hourly Out Going Long Wave Radiation Flux (OLR), Data from Kalpana-1, Very High Resolution scanning Radiometer (VHRR). (starting from May 2004 onwards) and Atlas of Spatial variations of moisture regions and rainfall of India during 19^{th} and 20^{th} Century (Mooley et al., 1981) (Parthasarathy et al., 1987) (Parthasarathy et al., 1993) (Parthasarathy et al., 1995) (Pant and Rupa, 1997). Kalpana-1 satellite is the first dedicated meteorological geostationary satellite launched by Indian Space Research Organization (ISRO). Under ISRO's Prediction of Regional Weather with Observational Meso-Network and Atmospheric Modeling (PRWONAM) programme, Automatic Weather Stations (AWS) were installed through out the country. AWS can continuously record weather data such as temperature, atmospheric pressure, wind speed and direction, rainfall, relative humidity, solar radiation, etc. The data from these AWSs are collected through the Data Relay Transponder on the ISRO's INSAT (Kalpana-1 and INSAT-3A) satellites. The observed data are archived at Meteorological and Oceanographic Satellite Data Archival Centre (MOSDAC) of the Space Applications Centre, ISRO (Kumar et al., 2011). These data are available at: http://www.mosdac.gov.in. These are the main sources of rainfall data that are used by researchers for working with ISMR weather scenario.

In order to provide atmospheric data to the research community, National Centers for Environmental Prediction (NCEP) is involved in two global reanalysis projects:
1. A joint venture with the National Center for Atmospheric Research (NCAR) called NCEP/NCAR Reanalysis (Reanalysis-1) (Kalnay et al., 1996). This is a global reanalysis atmospheric data, spanning from 1948 to the present day.
2. A joint venture with the Department of Energy (DOE) called NCEP/DOE Reanalysis (Reanalysis-2) (Kanamitsu et al., 2002). It is a global reanalysis of atmospheric data, spanning from 1979 to the present day. Both reanalysis are the global sets of grided weather data at a 2.5° X 2.5° horizontal resolution. Reanalysis -1 data is inconsistent due to the changes in the assimilation. The unavailability of satellite images, during the early part of Reanalysis -1 and later inclusion of the inputs from satellite imagery data, has introduced the inconsistency. Inspite of the inconsistency, the data is significantly valid.

The European Center for Medium-Range Weather Forecasts (ECMWF) is an independent intergovernmental organization, supported by most of the nations of Europe. ECMWF maintains an archive of meteorological data known as ECMWF climate reanalysis (Dee et al., 2011) that started with the First Global atmospheric program Global Experiment (FGGE) reanalysis which has been produced in the 1980s, followed by ERA-15, ERA-40 and most recently ERA-Interim. ERA-5 is the fifth generation and the latest of the ECMWF atmospheric reanalysis of the global climate. The model has a horizontal resolution of 31km, and vertical resolution of 137 levels from the surface up to 0.01hPa (around 80km). It contains estimates of atmospheric parameters such as air temperature, pressure and wind conditions at different altitudes, surface parameters such as rainfall, soil moisture content, sea-surface temperature and wave height.

Greenhouse Effect Detection Experiment (GEDEX) was a workshop organized by NASA Earth Science and Applications Division. One of the primary objectives of the workshop was to assemble and document existing data (focusing on temperature) for the analysis of global climate change and to consolidate these selected datasets onto CD-ROMs. The datasets include surface, upper air, and/or satellite-derived measurements of temperature, solar irradiance, clouds, greenhouse gases, fluxes, albedo, aerosols, ozone and water vapor, along with Southern Oscillation Indices and Quasi-Biennial Oscillation statistics. Many of these datasets provide global coverage. The spatial resolutions vary from zonal to 2.5° grids. Temporal coverages also vary. Some surface station datasets cover more than 100 years, while most of the satellite-derived datasets cover data only from 1980 to 1992. Temporal resolution, for most datasets, is available monthly. The data is available at: http://iridl.ldeo.columbia.edu/SOURCES/.GEDEX/.

By combining observations from different platforms like satellite, ships and buoys an analysis has been constructed for Sea Surface Temperature called the NOAA 1/4° daily. Optimum Interpolation Sea Surface Temperature (OISST) is a spatially complete SST map and is produced by interpolation to fill in the gaps (Reynolds et al., 2007). However, minor modifications has been introduced in version 2. The data is available from 1981 to present day at the resolution of 2° X 2°. NOAA's National Centers for Environmental Information (NCEI) is responsible for hosting and providing access to one of the most significant archives on earth, with comprehensive oceanic, atmospheric, and geophysical data. From the depths of the ocean to the surface of the sun and from million-year-old tree rings to near real-time satellite images, the website hosts a variety of data that can be used in climate research. It spans the areas of satellite data, regional products, observational and near real time data, instrument types, model data, ocean profile data and ocean climatology. These data are available at: https://www.ncei.noaa.gov/.

Some of the researches use Global Sea-Ice and Sea Surface Temperature (GISST) (Rayner et al., 2003) dataset. It provides monthly SST data which is infilled using Empirical Orthogonal Function (EOF) interpolation and sea-ice fractions that are based on a mixture of charts, satellite observations and statistical interpolations.

The Met Office, Hadley Centre for Climate Science and Services, hosts data on a variety of domains like agriculture, atmosphere, biosphere, climate indicators, cryosphere, hydrosphere, land surface, oceans, paleo climate, solid earth, spectral/engineering, sun-earth interactions etc., which is available at: http://rda.ucar.edu/.

The Historical Soviet Daily Snow Depth (HSDSD) provides observed snow depth data from 284 World Meteorological Organization (WMO) stations throughout Russia and the former Soviet Union. The area covered is 35° to 75° north latitude and 20° to 180° east longitude. This data product was funded through NOAA's Environmental Services Data and Information Management (ESDIM) program. Data are available on a CD-ROM or for download via FTP. HSDSD Version 2 (Armstrong, 2001) replaces the Version 1. and this new version contains an additional fifteen years of data spanning from 1981 to 1995. It has also improved quality control; comes with the HTML interface and a Java tool for data browsing and extraction. Other parameters included in this product are snow cover percent, snow characteristics, site characterization, and quality flags. Data are in ASCII format and are available at: http://nsidc.org/data/docs/noaa/g01092_hsdsd /index.html.

Climate diagnostic bulletin of the Climate prediction center, USA, is a report that provides near real time ocean/atmosphere monitoring, assessments and prediction. This report is updated in the first week of every month; the data of various time series like Southern Oscillation Index (SOI), Tahiti and Darwin Sea Level Pressure (SLP) anomalies, OLR anomalies, Zonal wind anomalies for different geopotential hight, SST anomalies, depth of the 20° C Isotherm etc. are published in this monthly bulletin, and is available at: http://www.cpc.ncep.noaa.gov/ products/CDB.

The Bureau of Meteorology is, Australia's national weather, climate and water agency. It provides observational, meteorological, hydrological and oceanographic services. They are available at: http://reg.bom.gov.au/reguser/ (Chaudhuri and Pal, 2014).

Table 2.2: Sources of meteorological data used in ISMR research and their links with references

Sl.no	Source	Variables	Link	References
1	Indian Institute of Tropical Meteorology (IITM)	Homogeneous Indian monthly rainfall datasets, Homogeneous Indian monthly surface temperature dataset, Three hourly OLR data.	Http://tropmet.res.in	(Parthasarathy et al., 1995)
2	Ministry of Earth Sciences, India Meteorological Department (IMD)	Area weighted means to form monthly and seasonal homogeneous Indian monsoon rainfall series.	https://data.gov.in/catalog/all-india-area-weighted-monthly-seasonal-and-annual-rainfall-mm	(Parthasarathy et al., 1993) (Sontakke et al., 2008)
3	Climate Prediction Center (CPC)	Merged analysis of precipitation (CMAP).	ftp://ftp.cpc.ncep.noaa.gov/precip/cmap	(Xie and Arkin, 1997) (Xie and Arkin, 1996)
4	Meteorological and Oceanographic Satellite Data Archival Centre (MOSDAC) of Space Applications Centre, ISRO	Temperature, atmospheric pressure, wind speed and direction, rainfall, relative humidity, solar radiation.	http://www.mosdac.gov.in	(Kumar et al., 2011)
5	NCEP/NCAR Reanalysis - 1	Large set of variables spanning pressure level, surface fluxes, other fluxes, spectral coefficients are available.	https://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html	(Kalnay et al., 1996)

Table 2.2: Sources of meteorological data used in ISMR research and their links with references

6	NCEP/DOE Reanalysis (Reanalysis-2)	Atmospheric data.	http://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCEP-DOE/.Reanalysis-2/	(Kanamitsu et al., 2002)
7	ECMWF climate reanalysis	Air temperature, pressure and wind at different altitudes, surface parameters such as rainfall, soil moisture content, sea-surface temperature and wave height.	http://www.ecmwf.int/en/research/climate-reanalysis	(Dee et al., 2011)
8	Greenhouse Effect Detection Experiment (GEDEX)	Surface, upper air, and/or satellite-derived measurements of temperature, solar irradiance, clouds, greenhouse gases, fluxes, albedo, aerosols, ozone, water vapor, along with southern oscillation indices and quasi-biennial oscillation statistics variables.	http://iridl.ldeo.columbia.edu/SOURCES/.GEDEX/	(Schiffer, 1992)
9	The NOAA 1/4° daily Optimum Interpolation Sea surface temperature (OISST)	Sea surface temperature.	https://www.ncdc.noaa.gov/oisst	(Reynolds et al., 2007)

Table 2.2: Sources of meteorological data used in ISMR research and their links with references

10	NOAAs National Centers for Environmental Information (NCEI)	Satellite data, regional products, observational and near real time data, instrument types, model data, ocean profile data and ocean climatology.	https://www.ncei.noaa.gov/	—
11	Global Sea-Ice and Sea Surface Temperature (GISST)	Global sea-ice and Sea Surface Temperature.	http://www.metoffice.gov.uk/hadobs/gisst/	(Rayner et al., 2003)
12	Met Office Hadley Centre for Climate Science and Services	Data from verity of domains like land surface, oceans, sun-earth interactions etc.	http://rda.ucar.edu/	—
13	Historical Soviet Daily Snow Depth (HSDSD)	Snow depth data.	http://nsidc.org/data/docs/noaa/g01092_hsdsd/index.html	(Armstrong, 2001)
14	Climate Diagnostic Bulletin of the Climate Prediction Center, USA	Data of various time series like Southern Oscillation Index (SOI), Tahiti and Darwin SLP Anomalies, OLR Anomalies, Zonal wind anomalies etc.	http://www.cpc.ncep.noaa.gov/products/CDB	—-
15	The Bureau of Meteorology, Australia	Meteorological, hydrological and oceanographic data.	http://reg.bom.gov.au/reguser/	(Chaudhuri and Pal, 2014)

Table 2.2: Sources of meteorological data used in ISMR research and their links with references

16	Climatic Research Unit	Temperature, Precipitation, Pressure, Circulation Indices etc.	http://www.cru.uea.ac.uk/data	(Kumar et al., 2011)

The Climatic Research Unit is widely recognized as one of the world's leading institutions concerned with the study of natural and anthropogenic climate changes. It hosts data of temperature (5° X 5° resolution), precipitation and drought (global land data on 5° X 5° and 5° X 5° resolution for 1900-1998), pressure and circulation indices, etc. They are available at: http://www.cru.uea.ac.uk /data (Kakade and Kulkarni, 2014). The sources of data, their links and important references, are consolidated in Table 2.2.

2.2 PREDICTORS: A REVIEW

When India got its independence in 1947, it started to focus on the all round development of science. Due to the major impact of monsoon on agriculture and its direct contribution to the Gross Domestic Product (GDP) of the country, precise prediction of monsoon became an important challenge. This was a serious motivation for modern research and the beginning of literature period in the field of Indian meteorology, driving the scientific community towards digging meteorological domain for predictors, to improve ISMR forecasting abilities from time to time.

The geophysical spread of South Asia consists of conditions that make high concentrations of rainfall in specific parts at a specific time of the year. This geographical spread contributes to the formation of semi permanent system year after year, giving clues regarding the intensity and spread of rainfall throughout the Indian subcontinent.

Some of the important semi permanent systems that are seen during monsoon are:

1) **Mascarene High (MH)** (Krishnamurti and Bhalme, 1976). Mascarene is located at 30° South latitude and 70° east longitude, and is about 4,000km from the Indian mainland. A high pressure, in this area, drives Indian monsoon current towards India.

2) **Heat Low (HL)** (Rajeevan and Nanjundiah, 2009). Heat Low also known as thermal low is commonly referred to the low surface pressure (3-10hPa) areas in

the northern plains of India, caused by intense heating of the land-surface (and overlying atmosphere) from solar radiation.

3) **Monsoon Trough (MT)** (Anjaneylu, 1869) (Keshavamurty, 1968) (Keshvamurty and Awade, 1970). Monsoon trough is a convergence zone between the winds of the Southern and northern hemisphere, depicting as a line on a weather map. The convergence is caused by the low mean sea level pressure in this area.

4) **Low Level Jet (LLJ)** (Findlatter, 1969)(Joseph and Sijikumar, 2004) (Joseph and Raman, 1996). It is a ribbon of relatively strong winds in the lower part of the atmosphere.

5) **Tibetan High (TH)** (Flohn, 1960)(Koteswaram, 1958a)(Yanai and Song, 1992)(Wu et al., 2004). Tibetan High is the high pressure zone (about 600hPa) created over the Tibetan plateau.

6) **Tropical Easterly Jet (TEJ)** (Koteswaram, 1958b) (Krishnamurti, 1971). Tropical Easterly Jet refers to the easterly winds at upper level (about 15km above Earth's surface).

Monsoon over India is logically studied through three stages namely onset, progress and the withdrawal. In India, onset of monsoon is when the monsoon winds hit the coast of Kerala. It was first investigated by Ananthakrishnan et al. (1967). These monsoon winds that pick up a lot of moisture during their travel across Indian Ocean, divide into two parts at the Peninsular tip and further proceed across India giving rainfall in different quantities depending on the local conditions. For example Western ghats get heavy orographic rains, Cherapunji has a record of receiving the highest rainfall in the world (Dhar and Nandargi, 2006) etc. The Himalayas form a barrier that stops these moisture bearing winds from crossing over into Tibet and further North (Dhar and Nandargi, 2007). This is called the progress of monsoon. During the end of September there is a change in the direction of the wind regime over India. The moist winds of South-West monsoon is replaced by the dry continental winds as it reverses its direction. These winds propagate from North-East towards the Indian Ocean, this phase is called the withdrawal or the North-East monsoon. The semi permanent systems mentioned above may be regarded as the cause of onset, progress and withdrawal of monsoon.

During the onset of monsoon in the southern hemisphere near the Mascarene Island there forms a high pressure zone called Mascarene High. Around this area, the anticyclone air circulation is seen. The monsoon winds are drawn from this high pressure area towards the equatorial troughs (low pressure area). Some of the main indicators of onset include moisture buildup at mid troposphere (Soman and Kumar, 1993), meridional temperature gradient, synoptic scale disturbances like creation of low pressure area, formation of onset vortex, and mid troposphere trough (Bhatla et al., 2016) mini warm pool, where SST 30.5° C and its impact on the formation of Onset Vortex (OV) over the East-Central Arabian Sea (ECAS), 850hPa wind fields from May to June (prior to the onset of monsoon) over the North Indian Ocean (Deepa et al., 2010), humidity and air temperature (Stolbova et al., 2016), characteristic fall in OLR (Pai and Rajeevan, 2007) are other indicators of monsoon activity.

Low Level Jet (LLJ) is a relatively strong wind at the lower level of the atmosphere. These are weaker than the planetary scale upper troposphere jet. They are mainly linked to two functions: 1) transport of moisture 2) momentum. LLJ can give possible clues regarding the onset due to its second function and clues regarding inter annual and inter seasonal monsoon variability through its first function (Puranik et al., 2014). The role of LLJ during the onset is depicted in relation to MH. MH influences northern and southern hemispheric circulation (Xue et al., 2003). Thus, strengthening the cross equatorial flow (Sikka and Gray, 1981) and forming the root cause of monsoon onset over India (Okoola and Asnani, 1981),

Cross equatorial low level jet originates at the MH, and establishes near the Somalia coast, which is near the equatorial Arabian sea. this forms a cyclonic storm called the onset vortex, causing the cross equatorial flow of LLJ (Raju et al., 2005) (Wang, 2006) (Keshavamurthy and Rao, 1992) (Li et al., 2016) and contributes to the moisture build up over the Arabian Sea. A relationship between the meridional gradient of SST anomalies and the co-operation of ENSO and IOD events (Sankar et al., 2011) (Chakravorty et al., 2013) (Liu et al., 2015) is clearly reported in the literature. The onset is characterized by the strengthening of zonal wind at 850hPa (Wang et al., 2009) (Bhatla et al., 2016), formation of SST warm pool in the Bay of Bengal, formation of Arabian Sea warm pool and a large area of convection is seen to its South, which is pulled across the equator and moves North from very low latitudes to Kerala and brings about onset (Joseph et al., 2006).

In the history of monsoon, the earliest monsoon and the delayed monsoon onset date difference is at the most around 46 days (Sahana et al., 2015). Onset and withdrawal of ISMR are partially modulated by the tropical Pacific Sea Surface Temperature (SST) and specifically by the El Nino-Southern Oscillation (ENSO) (Sabeerali et al., 2012). The Sea Surface Temperature anomalies in the western Pacific are negative (positive) during delayed (early) onset (Sankar et al., 2011). The spring time snowpack over the Himalayan Tibetan Plateau (HTP) region and Eurasia has long been suggested to be an influential factor on the onset of the Indian summer monsoon (Datta, 1993)(Mamgain et al., 2010)(Senan et al., 2016). Tibetan Plateau apparently has its effect on ISMR onset as well (Abe et al., 2013).

For the monsoon winds, that come from the MH and hit the coast of Kerala on onset, have to further travel North giving rains to the subcontinent all along their way. The semi permanent system called the Heat Low, plays its role in this process called progress (Hastenrath, 2012). In the subtropical region, during summer, some areas exhibit low pressure due to intense heating, when compared to the surrounding areas. This low pressure areas span from the Sahara across Arabia, Iran, Afghanistan, Pakistan, North-West India; finally ending near Myanmar. This low pressure creates a trough over North India, running from North-West India and ending at the North coast of Bay of Bengal. This trough is called the Monsoon Trough (MT) and is a result of Heat Low. Due to the low pressure at the trough, the monsoon winds that have reached Kerala during the onset, further gust towards the trough, giving rains all over India.

The Heat Low and MT give clues of Inter Annual (IA) and Inter Seasonal (IS) variability of rainfall through the variations of surface temperature, the Mean Sea

Level Pressure (MSLP), the troposphere geopotential height, and wind patterns (Yadav, 2009). The average MSLP of six stations (Jodhpur, Ahmadabad, Bombay, Indore, Sagar and Akola) in the Western Central Indian (WCI) region showed highly significant Correlation Coefficients (CCs) for March-April-May(MAM) and December-January-February (DJF) with ISMR (Parthasarathy et al., 1992)(Bansod et al., 2015). Low Pressure System (LPS) is a major rain-bringing synoptic circulation that forms over the Indian region, including Bay of Bengal and Arabian Sea. LPS plays a vital role in performance of South-West monsoon over the country (Jadhav and Munot, 2009).

Monthly patterns of OLR anomalies, pre-monsoon months, cloud cover and Long Wave Cloud Irradiative Forcing (LWCRF) over the Bay of Bengal region, appears to be good indicators of the forthcoming monsoon rainfall (Munot et al., 2011).

Many monsoon depressions move across the country yearly and their numbers within the Bay of Bengal increase, during July and August of El Niño events (Singh et al., 2002). Monsoon depressions are efficient rainfall producers. Some of the recent studies also suggest the possible influence of ENSO indicators in driving weather conditions (Pradhan et al., 2016) (Nanjundiah et al., 2013) (Kumar et al., 2013) that result in severe drought in India (Neena et al., 2011) (Revadekar and Kulkarni, 2008) (Kumar et al., 2006). It is also seen that ENSO indices influence the Indian Ocean Dipole (IOD), affecting the cross equatorial flow (Chowdary et al., 2006). Internal variability of ISMR can be explained by ENSO indices and IOD (Gadgil et al., 2004) (Pillai and Mohankumar, 2010) (Ashok et al., 2004) (Sajani et al., 2015).

Presence of El Niño like conditions in the Pacific and warming over the equatorial Indian Ocean alter the circulation patterns and weaken the cross-equatorial flow in the drought conditions (Preethi et al., 2011). The low level jet stream drives across equatorial ocean current, the along shore winds induce upwelling, resulting in SST drop in eastern Arabian Sea. This decreased SST has a significant impact on variability of ISMR (Wang, 2006). 850hPa zonal wind gradient anomaly between the region-1 [Equation - 25° N; 30° E 75° E] and region-2 [Equation - 10° S; 30° E - 75° E] is used to represent the cross-equatorial flow (Kakade and Dugam, 2008). The observational studies have (Saha, 1974), established the importance of the role of cross equatorial flow over Indian Ocean with the inter annual variability of ISMR.

Tibetan high plays a role of mechanical barrier and also a heat source (Murakami, 1987) helping anticyclone formation in the upper troposphere above Tibet during the ISMR season. The Tibetan Plateau covers a large area from central to western Asia. It acts as a heat source during the summer and a heat sink during the winter (Ye and Wu, 1998). This barrier stops the moisture bearing winds that come from the MH, from escaping beyond Tibet; ensuring maximum rainfall to the Indian subcontinent. It is found that the Tibetan Plateau Snow Cover (TPSC) is closely linked to the inter annual variations of summer heat waves over Eurasia (Wu et al., 2016). The excessive Eurasian snow cover is associated with a weak monsoon, characterized by higher sea level pressure over India, a weaker Somali jet, weaker lower troposphere westerlies, and weaker upper troposphere easterlies, resulting in weaker summer monsoon (Vernekar et al., 1995)(Liu and Yanai, 2002)(Bamzai and Shukla, 1999). Energy used in melting excessive snow, reduces the surface temperature over a broad region centered around the Tibetan Plateau (Fasullo, 2004). Reduced surface sensible heat flux reduces the mid troposphere temperature over the Tibetan Plateau. The result is to reduce the mid troposphere meridian temperature gradient over the Indian Peninsula weakening the monsoon circulation (Pang et al., 2005). The apparent snow-monsoon relationship generally denotes a very strong El Niño - Southern Oscillation tele connection with winter snow cover and summer monsoon rainfall, rather than a direct influence of the Eurasian snow cover on the Indian monsoon (Peings and Douville, 2010).

TEJ refers to an upper troposphere easterly wind, and is found on the southern periphery of the upper troposphere of the Tibetan plateau (Peings and Douville, 2010). It spans through South-East Asia across Indian Ocean and from Africa to the Atlantic, at a height of 14km (Koteswaram, 1958a) (Koteswaram, 1958b). The

TEJ is week when the monsoon is weak and strong when the monsoon is strong.

Withdrawal or retreating monsoon also known as North-East monsoon starts by the end of September. Towards the end of September the low pressure trough formed in North India, disintegrates, and the low pressure shifts to the equatorial region. As a result winds start blowing towards the equator; away from the northern planes exactly opposite to what happed during the onset of monsoon. Winds also blow from the Tibetan highlands and from further North, towards the equator. These winds while traveling through Bay of Bengal, pick up moisture and precipitate on their way back, giving rains to Orissa, Tamilnadu and some parts of Karnataka. Two possibilities can be followed in finding predictors for North-East monsoon: 1) To determine whether the potential predictor of the summer monsoon can be useful in predicting the North-East monsoon, 2) To find predictors in climate variability in the pre North-East monsoon months. The study conducted by Tomita and Yasunari (1996) shows that northern hemisphere temperature, Indian summer monsoon, North-East winter monsoon, Walker circulation, Sea Surface Temperatures (SSTs) over tropical Indian and Pacific oceans, trade winds, Pacific/North American (PNA) pattern, North Pacific Oscillation (NPO) and Eurasian Pattern are dynamically/thermodynamically linked in a Quasi Biennial Oscillation (QBO) mode. Following this, 20 predictors consisting of sea level pressure, surface air temperature, frequency of westerly wind days and quasi-biennial oscillation in different parts of the years; of different parts of the world, have been studied. This study concluded that the pre monsoon forcing factors are the ones that work on North-East monsoon as well (Singh and Sontakke, 1999). A study on Indian Ocean Sea Surface Temperature and its relationship with North-East Monsoon Rainfall (NEMR) revels that, during the positive SST gradient years, the Inter Tropical Convergence Zone (ITCZ) shifts northwards over the East Indian Ocean. The tropical depressions and the storms and cyclones, formed in the North Indian Ocean, move more zonal, and strike the southern Peninsular India; resulting in excess NEMR. While, in the negative SST gradient years, the ITCZ shifts southwards, over the Indian Ocean. The tropical depressions and the storms and cyclones formed in the North Indian Ocean, move more towards North-westward direction and after crossing 15° N latitude, re-curve to North-East direction towards the Bay of Bengal (BoB) and miss Southern Peninsular India, causing, deficient NEMR (Yadav, 2013). The inter annual variability of North-West India Winter Precipitation (NWIWP), is examined in association with the variability of surface temperature, mean sea level pressure, troposphere geopotential height, and wind patterns over the globe. It is seen that the inter annual variability of NWIWP is Arctic Oscillation/North Atlantic Oscillation (AO/NAO) and El Niño - Southern Oscillation (ENSO) phenomena. Above-normal NWIWP is associated with the positive phase of AO/NAO and the warm phase of ENSO. During the last two / three decades, the influence of ENSO over NWIWP has increased while the influence of AO/NAO has decreased (Pai and Rajeevan, 2007). The Influence of ENSO indices on the North-Eastern summer monsoon rainfall has been studied. Major Oscillations like El Nino Southern Oscillation (ENSO), Indian Ocean Dipole (IOD), and Equatorial Indian Ocean Oscillation (EQUINOO) in the Indian and Pacific oceans show well explained relationships with North-East Monsoon Rainfall (NEMR) (Sreekala et al., 2012). It is seen that the positive phases of ENSO and the Indian Ocean Dipole (IOD) are conducive for normal to above normal rainfall activities during the North-East monsoon (Rajeevan et al., 2012). Inverse relationship of ENSO, with summer monsoon rainfall, has weakened in the recent years, whereas the positive relationship of ENSO with North-Eastern summer monsoon rainfall has strengthened. This is due to epochal changes in the regional circulation features (Kumar et al., 2007). From the above studies, we can see that when a global condition exists, a change occurs in the cross-equatorial flow, which in turn affects inland conditions. Thus, a dependency has been established among these predictors with respect to the summer monsoon season. Each of these studies use different predictors to predict ISMR satisfactorily, with low error rates. Since the efficiency of predictors and their influences may vary, due to the changes in the climatic phenomena, global warming, deforestation, pollution, and other factors, there is always a need to identify effective predictors from time to time. From the Monsoon dynamics depicted above, it can be seen that there is a large number of predictors that can be used for forecasting Indian Monsoon Rainfall (IMR). These predictors can be grouped into regional conditions, ENSO indicators, cross equatorial flow, and global hemispheric conditions (Kumar et al., 1995).

Regional conditions consists of various predictors on surface as well as upper-air meteorological parameters from India and adjoining regions. For example the pre monsoon wind regimes, the heating of the northern landmass, the pressure dip on the land etc.

ENSO phenomenon is an important aspect that influences global weather. There have been many predictors that are derived form its atmospheric and oceanic components. Walker was the first one to develop an index based on ENSO phenomenon, this index was based on pressure and temperature. The Southern Oscillation (SO) is the atmospheric component and El Niño is its oceanic component based on temperature variations.

Cross equatorial flow and its importance in IMR has been established in the 1970s (Saha, 1974). The main aspects of cross equatorial flow are the moisture flux from both Indian Ocean and Arabian sea. The low level jet of East Africa is a significant contributor to cross equatorial flow. In the recent years, the discovery of Indian Ocean Dipole has given good evidence of cross equatorial flow and is an important factor in deciding the inter annual variability of monsoon over India.

The Northern hemispheric surface air temperature (Verma et al., 1985) and the Eurasian / Himalayan snow cover (Shukla, 1987)(Hahn and Shukla, 1976)(Dey and Kumar, 1983) (Dickson, 1984)(Kumar, 1988) have proved its part in prediction of IMR. The link between monsoon and the snow cover has been discussed since 1800.

Table 2.3 gives the list of possible predictors of IMR, grouped according to above mentioned category.

Table 2.3: Predictors of ISMR, pertaining to Onset, Progress and Withdrawal of ISMR, grouped in accordance to (Kumar et al., 1995)

I	**Regional conditions**				
Sl.no	**Name**	**Onset**	**Progress**	**Withdrawal**	**References**
1	Rainfall at Colombo, Minicoy, Thiruvananthapuram, Allapuzha, Kochi, Kozhikode and Mangalore	✓	×	×	(Puranik et al., 2013)
2	Westerly wind circulation up to 600 hectopascal (hPa)	✓	×	✓	(Pai and Rajeevan, 2007) (Singh and Sontakke, 1999)
3	OLR	✓	✓	×	(Pai and Rajeevan, 2007) (Chaudhari et al., 2010)
4	Humidity of air at 500hPa	✓	×	×	(Pai and Rajeevan, 2007)
5	Air temperature at 1000hPa	✓	×	✓	(Stolbova et al., 2016)(Singh and Sontakke, 1999)
6	Vapour up to 300hPa	✓	✓	×	(Soman and Kumar, 1993) (Puranik et al., 2014)
7	SST of Bay of Bengal and Arabian sea	✓	✓	×	(Joseph et al., 2006)(Jadhav and Munot, 2009)(Wang, 2006)

Table 2.3: Predictors of ISMR, pertaining to Onset, Progress and Withdrawal of ISMR, grouped in accordance to (Kumar et al., 1995)

8	Mean sea level pressure of six stations Jodhpur, Ahmedabed, Bombay, Indore, Sagar and Akola	✓	×	✓	(Parthasarathy et al., 1992)(Bansod et al., 2015)(Singh and Sontakke, 1999)
9	Low pressure systems, SST of Bay of Baengal	×	✓	×	(Jadhav and Munot, 2009)
10	Ultra-long waves of low latitudes	×	✓	×	(Bawiskar et al., 2009)
11	The long wave cloud irradiative forcing	×	✓	×	(Munot et al., 2011)
12	Monsoon onset date	×	✓	×	(Rai et al., 2015)
13	Indian Ocean Sea Surface Temperature	×	×	✓	(Yadav, 2013)
II	**ENSO indicators**				
14	SST Tropical Pacific (NINO3, NINO4, NINO3.4)	✓	✓	✓	(Sabeerali et al., 2012)(Sankar et al., 2011)(Liu et al., 2015)(Chakravorty et al., 2013)(Pradhan et al., 2016) (Nanjundiah et al., 2013)(Kumar et al., 2013)(Parthasarathy et al., 1992)
15	Southern oscillation	×	✓	✓	(Revadekar and Kulkarni, 2008)

Table 2.3: Predictors of ISMR, pertaining to Onset, Progress and Withdrawal of ISMR, grouped in accordance to (Kumar et al., 1995)

16	North Atlantic oscillation and Southern oscillation	✗	✓	✗	(Kakade and Dugam, 2006)
III	**Cross-equatorial flow**				
17	ENSO indices and IOD	✗	✓	✓	(Gadgil et al., 2004) (Pillai and Mohankumar, 2010)(Pillai and Mohankumar, 2010)(Sajani et al., 2015) (Pai and Rajeevan, 2007) (Rajeevan et al., 2012)
18	Low-level kinetic energy 850hPa	✓	✗	✗	(Raju et al., 2005)
19	Low-Level Jet (Somali jet) ranging from 1 to 1.5 Km above ground	✓	✓	✗	(Wang, 2006)(Li et al., 2016)(Sengupta et al., 2007)
20	850hPa zonal wind averaged over southern Arabian Sea (5° to 15° N, 40° to 80° E)	✓	✗	✗	(Wang et al., 2009)(Bhatla et al., 2016)
21	evaporation rates over the southern Arabian Sea	✓	✗	✗	(Ramesh Kumar et al., 2009)

Table 2.3: Predictors of ISMR, pertaining to Onset, Progress and Withdrawal of ISMR, grouped in accordance to (Kumar et al., 1995)

22	IOD	✓	✓	✓	(Sankar et al., 2011)(Liu et al., 2015)(Chakravorty et al., 2013)(Gadgil et al., 2004)(Pillai and Mohankumar, 2010)(Pillai and Mohankumar, 2010)(Sajani et al., 2015)(Krishnan et al., 2011)(Preethi et al., 2011)(Sreekala et al., 2012)
IV	**Global/hemisphere conditions**				
23	Springtime snow pack of Himalayan-Tibetan-Plateau (HTP) region and Eurasia	✓	×	×	(Datta, 1993)(Saha et al., 2013)(Vernekar et al., 1995)(Liu and Yanai, 2002)(Bamzai and Shukla, 1999)
24	Starting and ending dates of snowfall	✓	×	×	(Mamgain et al., 2010)
25	Onset of monsoon over South China Sea	✓	×	×	(Puranik et al., 2013)(Shaw, 2009)
26	20 days lag-lead relationship between NAO and MJO	×	✓	×	(Dugam, 2008)
27	Sea-Ice Extent (SIE) of the western Pacific ocean	×	✓	×	(Prabhu et al., 2009)

Table 2.3: Predictors of ISMR, pertaining to Onset, Progress and Withdrawal of ISMR, grouped in accordance to (Kumar et al., 1995)

28	Sea-ice extent over the Bellingshausen and Amundsen Sea Sector (BASS)	✗	✓	✗	(Prabhu et al., 2010)
29	Arctic Oscillation/North Atlantic Oscillation (AO/NAO)	✗	✗	✓	(Pai and Rajeevan, 2007)

With a large number of predictors in the climate scenario, from the discovery of the predictors to using them in prediction models many important tasks like assimilating these predictors, finding their relevance at different times of the year with ISMR, the preprocessing techniques to prepare data for research etc., are involved. The mere task of making the data usable for prediction is a tedious and error prone procedure. These other wise very difficult tasks are currently rather easily handled by computers facilitating new predictor discovery, on time forecasting, helping asses inter relationships between variables so as to facilitate apt predictor selection etc.

2.3 PREDICTION MODELS: A REVIEW

Long Range Forecasting (LRF) of monsoon, has been in demand since the beginning of civilization, the formal start in India was with the inception of India meteorological department in 1875. Since then lot of research has been undertaken in order to find apt predictors and techniques for predicting Indian monsoon rainfall. Two main approaches to LRF have evolved: (1) statistical modeling (empirical) 2) dynamic modeling (Kumar et al., 1995). Statistical models rely on historical relationships (Rajeevan et al., 2004)(Rajeevan et al., 2000)(DelSole and Shukla, 2002) and time series analysis (Goswami and Srividya, 1996), where as dynamic methods use General Circulation Models (GCMs), which are mathematical models that simulate the circulation of the atmosphere and the ocean based on fluid dynamic equations. The GCM's are fed with initial conditions that are processed to forecast rainfall in the future. The use of statistical techniques for LRF are comparatively less complicated than the dynamic model. Notable techniques used in LRF include the Artificial neural networks (Singh and Borah, 2013). Neural Networks (NN) require less formal statistical training, as they have an ability to implicitly detect complex nonlinear relationships between dependent and independent variables, their ability to detect all possible interactions between predictor variables. The availability of multiple machine learning and Neural Network algorithms make it a good choice for building LRF models. In contrast, the black box nature of NN makes it difficult to introduce changes in the intermediate stages to improve the performance of the model. They require greater computational requirements for the process of learning and updating, while familiarizing with the new changing behavior of the predictors, as they are prone to over fitting (Tu, 1996). Genetic Algorithm (GA) is used in combination with neural networks (Kishtawal et al., 2003) (Kashid and Maity, 2012) (Dwivedi and Pandey, 2011). GA helps in finding an optimal NN design; but, has short falls like parameter tuning and computational complexity. Regressions have been extensively used in ISMR prediction (Rajeevan et al., 2007)(Sivakumar et al., 2015) ever since the beginning of LRF in India. They work by finding relationship between several predictors and one predictend. The inherent problem with using regression in prediction is; regardless of the type of regression used, it can give seriously wrong results, if there are severe outliers or influential cases. These must be identified and dealt with accordingly. From 1988 to 2002, the IMD issued forecast for ISMR using a power regression model (Gowariker et al., 1991) with 16 predictors. The forecast failure in 2002 prompted IMD to introduce several new models (Rajeevan et al., 2004). However, the new models failed in 2004 (Gadgil et al., 2005). Some of the prediction attempts, using the above mentioned techniques, include NNs (Guhathakurta et al., 1999) (Singh and Borah, 2013), genetic algorithms (Kashid and Maity, 2012) (Dwivedi and Pandey, 2011), and data mining (Dhanya and Nagesh Kumar, 2009). While these techniques have produced good models with better prediction capabilities, their performances must be further tuned, to improve the prediction of all India monsoon rainfall and homogeneous regional monsoon rainfall, by careful selection of predictor subset in combination with other techniques. Since statistical techniques use historical data, selection of predictors is a very important aspect for these models to perform well in forecasting. But from literature, it is seen that majority of the predictors are active only for some time period (Rajeevan et al., 2007) and later loose their importance with respect to rainfall. Hence it becomes necessary to choose techniques that can dynamically select predictors (Tripathi and Govindaraju, 2008) for forecasting with good accuracy. Some of the inherent problems that come with statistical models are epochal variation in the predictand - predictor relationship, inter-correlation between the predictors, changing predictability, etc. (Rajeevan et al., 2004). Statistical models can be improved by including some minor changes like changing the model size, the use of new predictors, changing the combination of predictors, changing the length of model training period, including ensemble techniques, including apt learning mechanisms with decision making techniques etc, (Rajeevan et al., 2007).

Dynamic methods use General Circulation Model (GCM), and are built from mathematical equations representing atmospheric, oceanic, land and sea ice components using concepts like the fluid dynamics, thermodynamics etc. The GCMs are fed with a state of climate at a particular instant called the initial condition; the GCM processes the equations and forecasts the possible state of climate in the future based on the provided initial conditions. This aspect of GCMs is widely believed to be advantageous in simulating global scale climate changes, as compared to simpler models, which do not calculate the large-scale processes from the first principles. Some insights can be gotten from such GCMs so as to help understand ISMR dynamics like the one explained by Bergen Climate Model Version (Furevik et al., 2003), ABOM POAMA version 2.4 (Hudson et al., 2011), GFDL CM version 2.1 (Delworth et al., 2006), FRCGC SINTEX-F model (Luo et al., 2005) (Wang et al., 2015), RegCM (Pattnayak et al., 2016)(Dash et al., 2006). According to the results of Bergen Climate Model Version, volcanic eruptions typically lead to strengthened Indian Summer Monsoon (ISM) circulation and ISM rainfall in the 3rd year after volcanic eruptions (Cui et al., 2014) . Simulating the inter annual ISMR variability, even with SSTs given as direct input, is a difficult task for Atmospheric General Circulation Model (AGCM), leave alone simulating SSTs (Sperber et al., 2001). Problems may relate to many factors, such as local effects of the formulation of physical parameterization schemes, while common model biases that develop elsewhere within the climate system may also be important. For example cold Sea Surface Temperature (SST) biases develop in the northern Arabian sea in the CMIP5 multi model ensemble (Levine et al., 2013) (Meehl et al., 2007). Such biases have previously been shown to reduce monsoon rainfall in the Met Office Unified Model (MetUM) by weakening moisture fluxes incident upon India. The simulated rainfall climatology on the Community Atmosphere Model (CAM) is found to be affected by biases over certain regions like the Arabian sea, the West coast of India and the central Indian Ocean region. By introducing a correction component, computed by using the Tropical Rainfall Measuring Mission (TRMM) climatology, these biases can be removed (Das et al., 2012). In an experiment, two dynamical AGCMs namely ECHAM3 (Version 3.6, Max Planck Institute) and the NCEP (MRF9) models (Klimarechenzentrum, 1992) (Kumar et al., 1996) have been tested for their skills in forecasting ISMR. Model outputs confirm that there are inter seasonal variations in the skill of the models, and this skill is also dependent on the location and the parameter being simulated (Grimm et al., 2006). A ported version of Climate Forecast System (CFS) (Saha et al., 2006), on retrospective prediction has provided good results, whereas, the 2009 prediction has had a large error, this is due to the inability of the model to capture positive Indian Ocean Dipole (IOD) phase. This suggests that, the error could be reduced with improvement of the ocean model over the equatorial Indian Ocean (Janakiraman et al., 2011). The Earth System Science Organization and Indian Institute of Tropical Meteorology have launched a National Monsoon Mission (NMM). The mission aims at improving the Indian monsoon forecast using Coupled Forecasting System developed by the National Centers for Environmental Prediction (NCEP), USA. A detailed analysis of sensitivity, to the initial condition, for the simulation of the Indian summer monsoon using Climate Forecast System version-2 (CFSv2) reveals that, in general spatial correlation bias either increases or decreases as forecast lead time decreases. Improvement of the physical processes in the CFSv2 (Saha et al., 2014) may enhance the overall predictability (Pokhrel et al., 2016). In a 2-tier modeling system, the predicted Sea Surface Temperatures (SSTs) are prescribed to an atmospheric GCM and in a 1-tier system, the SSTs evolve as a part of coupled ocean-atmosphere system (Kar et al., 2012). For example, the experimental seasonal forecast of ISMR on Community Atmosphere Model (CAM) (a

1-tire model), in the validation exercise, has pointed out a number of shortcomings in the forecast system (Das et al., 2015). Inspite of good skill in producing SST forecast, the GCMs are still unable to reproduce the observed remote response of SST and rainfall. The availability of several GCMs can be used to explore Multi-Model Ensembles (MMEs), enhancing the quality of prediction (Peng et al., 2002) (Palmer et al., 2004) (Hagedorn et al., 2005). Even though multi-model prediction systems may have better skills in predicting the Inter Annual Variability (IAV) of Indian Summer Monsoon (ISM), the overall performance of the system is limited by the skill of individual models (single model ensembles). This can be seen in the following ensemble based projects which have been studied by Ankita Singh et. al. ("http://www.sciencedirect.com/science/article/pii/S0377026512000425#"). Six GCM's namely: fully coupled versions of IRI models, referred to as MOM3AC1, and MOM3DC2, atmospheric component of European Centre Hamburg Model (ECHAM version 4.5) coupled with the Modular Ocean Model (version 3), IRI's mixed layer coupled model, referred to as ECHGML, the CFS, which is the National Center for Environmental Prediction (NCEP)'s climate forecast system version-1 model and the last two models, referred to as ECHcasst and ECHcfssst, are two-tier versions of ECHAM4.5, were used to analyze the predictability of All India Summer Monsoon Rainfall (AISMR) and its dependence on lead time using General Circulation Model (GCM) output. All these models are also discussed in detail (Kar et al., 2012). From this study, it is found that these model outputs do not have significant skills either over all the homogeneous regions or for the country as a whole. To improve the results Canonical Correlation Analysis (CCA) was applied on the outputs of GCM's (Singh et al., 2012). The Florida State University's, Coupled ocean-atmosphere, General Circulation Model (FSUCGCM) has been found to be successful in producing a deterministic forecast, superior to individual member models and a better than the multi-model ensemble mean forecast (Mitra et al., 2005). Whereas DEMETER is the acronym of the EU-funded project entitled, "Development of a European Multi model Ensemble system for a seasonal to an inter annual prediction". An ensemble model, when used to predict ISMR, resulting in poor performance due to the inability of individual models to capture very long breaks-drought relationship (Joseph et al., 2010). Hence, it sill becomes a necessity to strengthen the ability of individual models. GCMs have varying skills, in simulating climatology and inter-annual variability of observed rainfall, because of the large bias (systematic and random) in simulating the ISMR (Kar et al., 2012) (Acharya et al., 2011). However, even a very good GCM built will have some inbuilt short falls that result in the difference of observational analyses and their estimates are still limited due to incomplete and inaccurate representation of climate process, mainly due to limited computing power, which inturn leads to some simple assumptions made. The human understanding of the climate system in terms of physical, chemical and biological processes is not adequate (Stone and Risbey, 1990)(Office, 1995). Coarse resolution of GCMs also effects their ability to predict detailed rainfall variability (Acharya et al., 2011). As suggested by Acharya et al. (2014) and Palmer et al. (2004), ISMR is a complex phenomenon; thus its inherent uncertainty can be better predicted using probabilistic forecasts. Some statistical bias correction methodologies can be used to act on model outputs, making the statistical properties of the corrected data, match those of the observations (Acharya et al., 2013). This leads us to a question, whether current GCMs are in fact superior to simpler models for simulating temperature changes, associated with global scale climate change? These models have had varied degrees of success in their prediction capabilities. They exhibit limited skills in simulating intra seasonal variations in the ISMR (Sperber and Palmer, 1996). The GCM developed by (Sperber and Palmer, 1996) exhibits limited skills in simulating intra seasonal variations in the ISMR. The analysis of GCMs indicates that, the precipitation anomaly patterns of model ensemble predictions are substantially different from the observed counterparts in this region. However, the summer monsoon circulations are reasonably predicted (Zhu et al., 2008). The long-range prediction capabilities of GCMs are not satisfactory, therefore, statistical models are preferred. While statistical techniques have produced good models with better prediction capabilities, their performances must be further fine tuned, to improve the prediction of all India monsoon rainfall and homogeneous regional Summer Monsoon Rainfalls (SMR)s, by careful predictor subset selection in combination with other techniques. The revolution in the computer field, in terms of parallel processing, improved storage capacity leading to the discovery of new areas of technology (eg. big data analytics), distributed computing etc. helping complex tasks to be carried out in no time. The gap between predicted output and the reality has been narrowing since a few decades. Hence automation may be used to improve prediction accuracy, by making the statistical models and GCM more robust, by fine tuning and equipping models with sophistication.

From the various studies available, we know that prediction accuracy for all India monsoon rainfall and scaled down level monsoon rainfall is yet to be improved. Hence, is necessary to design better prediction models and to select more suitable predictors.

2.4 PREDICTION OF ISMR: A REVIEW

The 1877 famine was the turning point that laid a foundation for the monsoon forecasting in India. Sir H.F. Blanford used the relationship between winter and spring snow falls over Himalayas and the previous seasonal ISMR, to issue forecasts from 1882 to 1885. These were successful tentative forecasts. This successful forecast encouraged the first operational forecast for monsoon rainfall of 1886 (Blanford, 1884). Sir John Eliot applied subjective methods such as analogue and curve parallels on 1) Himalayan snow cover (Oct-May) 2) Local peculiarities of pre monsoon weather in India and 3) Local peculiarities over the Indian Ocean and Australia (Thapliyal, 1987). This model failed to predict droughts of 1899 and 1901. The short fall may be attributed to the subjective methods, without solid reasoning and the selection of predictors due to lack of knowhow of monsoon mechanism. Influenced by Hildebrandsson (1897) and Lockyer and Lockyer (1904); Walker, discovered Oscillations in Northern Hemisphere known as North Atlantic Oscillation (NAO) and North Pacific Oscillation (NPO) and one in the Southern Hemisphere Southern Oscillation (SO) (Walker, 1918). These discoveries were made in an attempt to include global conditions that can influence ISMR. The techniques of forecasting are also improved by eliminating subjectivity and by including concept of correlation. Regression technique was employed and Sir Gilbert Walker's first official forecast was made in 1909. Subsequently the country has been divided into 3 homogeneous subdivisions namely 1) Peninsula 2) North-East India 3) North-West India. Since then, forth the forecast has been facilitated by one regression equation for each region (Walker, 1924). Verification of these forecasts (1924-87) reveal that about 63% of these forecasts were correct (Thapliyal, 1982). The limited success of these forecasts may be due to the predictors only representing ENSO indices only. In reality ENSO indices alone cannot explain the entire ISMR variation (Pokhrel et al., 2016). The 1980s saw attempts to overcome the limitations of earlier multiple regression based Long Range Forecasting (LRF) models; by collective use of large number of predictors. This led to the development of a few operational LRF models, based on techniques like dynamic stochastic (Thapliyal, 1982), power regression and parametric (Gowariker et al., 1989)(Gowariker et al., 1991), principal component regression(Singh and Pai, 1996) (Rajeevan et al., 2000), canonical correlation analysis (Rajeevan et al., 1999)(Prasad and Singh, 1996), neural networks (Navone and Ceccatto, 1994)(Goswami and Srividya, 1996)(Guhathakurta et al., 1999) and power transfer (Thapliyal, 2001).

Amongst these, the 16 parameter power regression and parametric models were used for official IMD forecast from 1988 to 2002. This model made decisions about rainfall, based on favorability of 16 parameters. If 70% of the parameters were favorable then the forecast was a wet monsoon and on the other hand, when less than 30% of parameters were favorable, there was 83% probability that monsoon rainfall would be deficient(Gowariker et al., 1989)(Gowariker et al., 1991). In view of the large prediction errors in 1997 and 1999 (Rajeevan et al., 2000), the model has been revised. Though this model includes many more predictors, other than only ENSO conditions, it failed miserably. It can be seen that, these 16 parameters include some of the out dated influences of ISMR like the April 500hPa ridge position (Thapliyal and Rajeevan, 2003). From 2003 IMD issued forecasts in 2 stages.

The first stage was an April - forecast and second stage at End June. The second stage forecasts consisted of update on April forecast along with seasonal rainfall forecast for the geographical sub regions of the country. In contrast, prior to 2003, the forecast was given by the end of May. This step may be regarded as a logical step towards improved forecasts because, many of the semi permanent systems seen during monsoon, strengthen or lose scope during June. The status of some oscillations like IOD is correctly known by the end of June. Thus, helping better forecasting results. During 2003 to 2006, forecasts were issued, using the 8 and 10 parameter models based on power regression and probabilistic discriminant analysis techniques (Rajeevan et al., 2004). The 8 and 10 parameter models worked better in comparison with the 16 parameter models (Rajeevan et al., 2004). This may be partly due to the two stage forecast and also the change in the techniques used. In 2004 the country was divided into 4 new homogeneous rainfall sub division, namely 1) North-West India (NWI), 2) North-East India (NEI), 3) Central India (CI) and 4) South Peninsular India (SPI). The verification of the forecasts for the period 2001 - 2008 shows that CI and SPI have greater root mean square error. This shows that using same set of predictors and techniques for both ISMR and Sub Division Rainfall (SDR) prediction does not yield accurate forecasts. There is a need of more research in the SDR area in both predictor selection and also model implementation (Kashid and Maity, 2012) is to be studied in detail. A new statistical forecasting system based on the ensemble technique was introduced in 2007. This model consisted of three major points; a) use of a new smaller predictor dataset b) use of a new non-linear statistical technique along with conventional multiple regression technique c) application of the concept of ensemble averaging. The ensemble forecasting system, operates by taking the mean of the two forecasts prepared from two separate set of model anomalies; Multiple linear Regression (MR) and Projection Pursuit Regression (PPR) techniques (Rajeevan et al., 2007). This model has apparently given good results. In 2014 the forecast was issued in 3 stages and the same model with replaced predictors was employed and the 3rd stage forecast was based on Principal Component Regression (PCR) model for prediction (Pai and Bhan, 2015). Further the same ensemble model was used in both 2015 and 2016 forecasts. The ensemble method uses multiple learning algorithms to improve prediction capabilities(Rokach, 2010). In the recent times extensive research has been carried out by individual researchers using latest technologies, such as, neural networks(Guhathakurta et al., 1999)(Singh and Borah, 2013), (Sahai et al., 2000), genetic algorithms (Kashid and Maity, 2012)(Dwivedi and Pandey, 2011), Multiple regression model (Sadhuram and Ramana Murthy, 2008)(Munot and Kumar, 2007) and data mining (Dhanya and Nagesh Kumar, 2009). Though the forecasting accuracy is continuously improving it has still a long way to go. The main reason being, there are hundreds to thousands of variables in the climate scenario, of which some may strongly influence ISMR, some may not influence at all and some may have a weak influence and also, the influence of these variables may epochally vary. For example the SST and its relationship with ISMR varies (Varikoden and Babu, 2015). Therefore, it is advised not to use the same predictors for forecasting year after year, as the efficiency of the predictors and their influence may vary, due to the changes in the climatic phenomena (Vathsala and Koolagudi, 2017). Instead a routine may be designed, for finding most likely predictors that influence ISMR. This, may further be used with statistical techniques for better forecasting. The literature shows various predictors affecting ISMR. To date, no research has used the maximum number of these known predictors for ISMR prediction (Vathsala and Koolagudi, 2016) . This may be in anticipation of the memory and CPU time constraint the data may impose. With all the recent developments in the field of high performance computing; including maximum number of known predictors of ISMR, nontrivial relationships that may boost prediction capabilities of the model (Gago and Bento, 1998) can be brought forth. Forecasting with many predictors, provides rich information. It provides robustness against some instability that may result due to a low number of predictors, when different categories of the same subject have to be predicted (Vathsala and Koolagudi, 2016). Statistical models have many inherent limitations (Kumar

et al., 1995) (Rajeevan, 2001). The correlations between monsoon rainfall and the predictors can never be perfect. They may undergo changes from time to time and there may be cross-correlations between the parameters (Hastenrath and Greischar, 1993) (Parthasarathy et al., 1991). Hence there has been research in the field of General circulation models for LRF, as the 2003 experimental forecasts from Atmospheric GCM have been parallely verified. GCMs may require more time to stabilise, due to their complexity in implementation and/or in customizing it to Indian scenario. The difficulty in implementation may be also due to limited understanding of the monsoon dynamics and their incorporation in programming GCMs. Once The GCMs are in place they can provide accurate LRF (Rajeevan et al., 2004). Hence, a general understanding can be derived from the long history of research in LRF, They are :
1) The predictor influence on ISMR is not constant
2) Selecting apt predictors for forecasting helps in better forecast
3) Statistical forecasting models are largely used, post processing techniques help in improving the accuracy of forecast
4) Dynamic models are not up to the mark as of now. However, with a few improvements, they can prove as better forecasters infuture, when compared to statistical models.

2.5 RESEARCH GAPS

The following research gaps are identified form the above review.

Predictor selection has become a major concern in the field of ISMR forecasting. The influence of predictors may vary or the correlation may cease to exist after a few years. Using the same predictors, year after year, until the failure in forecasting is observed, seems to be an unscientific practice. Instead, using some recent techniques for the fresh selection of predictor, for every new forecast, seems to be a good alternative. The approaches may be explored to identify new correlations among prediction and use fresh set of predictors for ISMR prediction.

Many of the recent models by individual researchers, mainly concentrate on

ENSO indices. Since some studies have shown that ENSO alone is not indicative of IA and IS variations of ISMR and also ISMR showing different behaviors during neutral ENSO activity years, there is a necessity to concentrate on different predictor combinations.

So far, none of the researches, have used all known predictors of ISMR for forecasting. This is mainly due to the anticipated memory and CPU time constraint while using huge data. Now, new efficient techniques can be used to handle such tasks, so as to improve forecasting ability. Assuming different predictor groups may be helpful in predicting different ranges of rainfall (example - floods, droughts etc).

Owing to differences in Homogeneous Rainfall Regions (HRR), it also calls for predictor exploration, for a better forecasting of SMR for each HRR. Due to very less work in exploring model building for HRR, this area of SMR forecasting has suffered. Each of the HRR seems to be very different from every other HRR. This gives us an opportunity to explore and build different models for different HRRs.

With the introduction of big data analysis, a system can be developed that simply takes all available meteorological data and processes them to select good predictors and to find any new peculiarities in daily, monthly trends that may contribute for better forecasting.

ISMR forecasting in terms of ranges like flood, droughts and normal rainfall, is commonly seen in history. Range prediction is a good initiative, but, the quantitative rainfall prediction can prove a better base to plan some of the important activities. Hence, quantitative value of rainfall forecasting may be explored.

Ensemble techniques used in statistical modeling can be explored for improving the statistical models for forecasting ISMR.

India being a big country in terms of the land area, even its states have different patterns of summer monsoon rainfall within them. For example Karnataka coast experiences heavy rainfall along the coast line whereas North interior Karnataka is one of the driest parts of India. Hence, forecasting as such, both, in terms of predictor exploration and model building for such a scaled down area can be a challenge that has to be addressed.

Indigenous development of GCM can be explored so as to meet the requirement of developing components that specifically suit the Indian scenario. Though the number of GCMs available today world wide is large, they are not built with local Indian conditions. Already available GCM results can be compared and contrasted with observed data, so as to provide clues in customising the GCMs to improve ISMR prediction.

In conclusion ISMR forecasting has a long history, that has provided us good amount of understanding of this complex phenomenon. The exploration is still not enough owing to the continuous change in the climatic conditions of the world due to; global warming, deforestation, green house effects etc. This area, as a research, seems to have no end to it. Thus, giving us an opportunity to explore new models, new predictors etc. decade after decade.

2.6 PROBLEM STATEMENT

Based on the research gaps listed above the research problem for current work is defined as follows.
Facilitating Indian Summer Monsoon Rainfall prediction through -
a. Carefully selecting the predictors for ISMR.
b. Predicting the rainfall of Homogeneous Rainfall Region (HRR), choosing Peninsular India as a case study.
c. Predicting quantitative values of rainfall rather than the rainfall ranges for scaled down geographical area, choosing North interior Karnataka as a case study.

The defined research problem has been elaborated with little insights below:
Forecasting with many predictors always provides a richer base of information and robustness against instabilities that may arise due to fewer predictors when there are different categories of the subject being predicted. The inclusion of many predictors can also bring forth nontrivial relationships that may further boost the model's prediction capabilities (Gago and Bento, 1998). Establishing

associations between predictor influences, there by taking advantage of these associations in selecting combinations of predictors, is to be carried out to exclude non-influential predictors. Individual predictors may exhibit limited incite for prediction, whereas, a combination of predictors may combinedly exhibit new behavior that facilitates prediction with better accuracy. Combination of predictors can capture the changing climatic conditions better, there by improving prediction capabilities. Hence, the first objective aims at using all known climate variables that can become predictors for ISMR during development of Long Range Forecasting (LRF) model. Association rule mining is used to extract useful predictors.

India is divided into four Homogeneous Rainfall Regions. Different homogeneous rainfall regions experience different rainfall behavior. HRR rainfall prediction is important for local applications like; tourism, crop insurance etc. The second objective aims at developing a model for prediction of HRR rainfall. Since the HRRs exhibit different rainfall behavior, their local conditions may also play an important role in prediction. Hence, local condition predictor exploration, based on correlation coefficient analysis, is essential. Combining these new local condition predictors, with global predictors in a model and the ensemble techniques, to achieve better accuracy, is to be tested on Peninsular Indian rainfall prediction.

Quantitative value prediction has lot more advantages, when compared to range prediction. The third objective aims at building a crisp quantitative rainfall value prediction model, using soft computing approach known as fuzzy logic. Fuzzy logic imitates the ability of human decision making in computing problems. This model with soft computing techniques is to be demonstrated on the North Interior Karnataka rainfall data.

2.7 GENERAL DATABASE AND DATA PREPROCESSING STEPS USED FOR THE CURRENT WORK

The data used in the current work includes sensor based observed data and reanalysis data (model data). For authenticity, rainfall and temperature data hosted by Indian government organization called the IITM is used; this data is obtained from IMD. Data of other climate variables are obtained from NCEP-NCAR, which is a combination of observations and reanalysis data. The following list provides the general description of database used. The task of preparing data includes a range of calculations from the raw data obtained. Although the data are available, they have to be brought to a format that will indicate their interrelationship with ISMR. The format here means; fetching trends, averages between months, picking maximum and minimum, negating values, and so on. Each of the selected variables involves different calculations that must be performed, to make it the predictor for ISMR.

1) Darwin Sea Level Pressure (DSLP)

DSLP is a Mean Sea Level Pressure (MSLP) between grid points 130.0° E and 12.5° S (Rasmusson et al., 1982). According to Shukla and Daniel A (1983), DSLP tendency has a good ISMR prediction capacity and is calculated as the difference between the March-April-May average and the December-January-February average.

$DSLP = MSLP_{Mar-Apr-May} - MSLP_{Dec-Jan-Feb}$

where DSLP is Darwin mean sea level pressure, and MSLP is a DSLP between grid points 130.0° E and 12.5° S.

(Source: `http://www.cpc.ncep.noaa.gov/data/indices/darwin`.)

2) ENSO-SO index (ENSO-SOI)

The NINO3, NINO4, and NINO3.4 indexes are considered.

Ihara et al. (2006) reported that the NINO3 index is calculated using monthly gridded Sea Surface Temperature (SST) anomalies, averaged over the NINO3 region ($5°N$ to $5°S, 150°W$ to $90°W$) and then averaged over the summer monsoon season, specifically June, July, August, and September (JJAS).

$NINO3\_index\_int = NINO3_{SST_(Mar-Apr-May)}$

$NINO3\_index = NINO3\_index\_int_{(Jun-Jul-Aug-Sep)}$

According to Kumar et al. (1995), the difference between the NINO4 area SST ($5°N$ to $5°S, 160°E$ to $150°W$) of the March-April-May average and the December-January-February average has a good correlation with ISMR.

$NINO4\_index = NINO4_{SST_(Mar-Apr-May)} - NINO4_{SST_(Dec-Jan-Feb)}$

NINO3.4 calculations are conducted based on Rayner et al. (2003) as the average of the March-April-May Pacific SST over the NINO3.4 area ($5°N$ to $5°S, 170°W$ to $120°W$).

$NINO3.4\_index = NINO3.4_{SST_(Mar-Apr-May)}$

(Source: http://www.cpc.ncep.noaa.gov/data/indices/sstoi.indices.)

ENSO-SOI (standard Tahiti MSLP − standard Darwin MSLP) is prepared as a predictor in accordance with Chakravarty (2011), by calculating March-April-May averages.

$SOI = MSLP_{standardTahiti} - MSLP_{standardDarwin}$

(Source: http://www.cpc.ncep.noaa.gov/data/indices/soi.)

3) Equatorial Indian Ocean Oscillation (EQUINOO)

EQUINOO is the oscillation of the Indian Ocean with an enhanced convection over the West Equatorial Indian Ocean (WEIO) (50°to 70°E, 10°S to 10°N) (Francis and Gadgil, 2013) and a reduced convection over the East Equatorial Indian Ocean (EEIO) (90°to 110°E, 10°S to EQ) (Francis and Gadgil, 2013)(positive phase) along with the anomaly of the opposite sign (negative phase). It has indexes based on the Indian Ocean SST and the zonal component of the surface wind at the equator (60°E to 90°E, 2.5°S to 2.5°N) called Equatorial Wind (EQWIN). Ready-made data are available in Charlotte et al. (2012). In this work, the same dataset has been directly selected without any further changes.

4) March-April-May minimum (min), maximum (max), average air temperature of Jodhpur, Ahmedabad, Bombay, Indore, Sagar, and Akola.

Observed data provided by the Indian Institute of Tropical Meteorology website have 1° X 1° longitude and latitude resolution of averaged March-April-May values for max, min, and average temperatures. Hence, the exact temperatures of the specific places mentioned above cannot be selected from the observed data because of the resolution. Thus, the temperature values from approximately the nearest longitude and latitude to the places considered, are computed.

(Source: http://cccr.tropmet.res.in/cccr/getUI.do?dsid=id-a4ea271c42&
varid=\tolerance9999\emergencystretch3em\hfuzz.5\p@\vfuzz\hfuzzwinter_

mean_temp-id-a4ea271c42&auto=true.)

5) March-April-May minimum (min), maximum (max) air temperature of South Indian region (20°N - 7.5°N, 60°E - 100°E).

Observed data as provided in Indian Institute of Tropical Meteorology (IITM) website has 1° X 1° (longitude and latitude) resolution of averaged March-April-May values for max and min temperatures. This is a local condition variable that is used in this study to find its possible relationship with PISMR and NKSMR.

(Source:http://cccr.tropmet.res.in/cccr/getUI.do?dsid=id-a4ea271c42&varid=winter_mean_temp-id-a4ea271c42&auto=true.)

6) MSLP - Average MSLP of Jodpur, Ahmedabad, Bombay, Indore, Sagar, and Akola.

Observed data downloaded from the National Centers for Environmental Prediction (NCEP) and the National Center for Atmospheric Research (NCAR) reanalysis are in 2.5° X 2.5° longitude and latitude resolution, which is quite coarse to interpolate and produce 0.5° X 0.5° resolution grid. The observed data are saved in the Network Common Data Form (NETCDF) format and an interpolation is taken, to obtain the missing data. Thus, the temperature values from approximately the nearest longitude and latitude, to the places considered from interpolated 0.5° X 0.5° resolution, are obtained.

(Source: http://iridl.ldeo.columbia.edu/maproom/Global/Atm_Circulation/Sea_Level_Pres.html.)

7) Mean Sea Level Pressure (MSLP) - Average Mean Sea Level Pressure of South Indian region and Indian Ocean above the equator (20°N - 5°N, 60°E - 100°E).

Observed data downloaded from NCEP and NCAR is with 2.5° X 2.5° (longitude and latitude) resolution. This is a local condition variable and is used to explore it possibility of becoming a predictor of PISMR and/or NKSMR.

(Source:http://iridl.ldeo.columbia.edu/maproom/Global/Atm_Circulation/Sea_Level_Pres.html.)

8) 200hPa meridional component of wind for May in Bombay, Delhi, Madras, Nagpur, and Srinagar.

Observed data downloaded from NCEP-NCAR reanalysis are in 2.5° X 2.5° longitude and latitude resolution, which is quite coarse. In this case, data are collected in the same way as explained in case of minimum and maximum air temperature. (Source: `http://iridl.ldeo.columbia.edu/maproom/Global/Atm_Circulation/Wind_SST.html`.)

9) 200hPa meridional component of wind, for the month of May in the South Indian region and the Indian Ocean, above the equator (20°N - 5°N, 60°E - 100°E). Being a local condition variable, it is used in the current study, to find its possible contribution in predicting NKSMR and/or PISMR.

Observed data downloaded from NCEP-NCAR is with 2.5° X 2.5° (longitude and latitude) resolution.

(Source:`http://iridl.ldeo.columbia.edu/maproom/Global/Atm_Circulation/Wind_SST.html`.)

10) Rainfall data for the whole of India.

ISMR Rainfall data is used in the current work (Dhanya and Nagesh Kumar, 2009).

(Source: `ftp://www.tropmet.res.in/pub/data/rain/iitm-regionrf.txt`.)

11) Rainfall data for Peninsular India region.

The observed data of Peninsular India is used in accordance to Dhanya and Nagesh Kumar (2009) in the current work.

(Source:`ftp://www.tropmet.res.in/pub/data/rain/iitm-regionrf.txt`.)

12) Rainfall data for North Interior Karnataka region. This monthly averaged data is provided by the Indian government department, for research. This data is grouped by homogeneous rainfall sub divisions of India.

(Source:`ftp://www.tropmet.res.in/pub/data/rain/iitm-regionrf.txt`.)

Since the data availability of the predictors used in all the individual chapters in this thesis is limited to the 1969-2005 (37 years) period, each of the study presented in the thesis is similarly restricted. The terms variables, predictors and features are interchangeably used in this thesis.

2.8 SUMMARY

This chapter critically reviews 1) Meteorological databases 2)Predictors of ISMR relative to monsoon mechanism 3) Models used for prediction of ISMR and 4) Prediction mechanisms of ISMR.

Meteorological database section covers the sources, the mechanisms of recording, inherent shortfalls and preprocessing techniques, that may be required to collect data, reliable for forecasting Indian Summer Monsoon Rainfall. A list of organizations, around the world, with proper references and links that provide meteorological data for research has been given. All possible variables involved in ISMR, that can become possible predictors, are depicted relatively, to the monsoon mechanism under the section "predictors of ISMR". The process of onset, progress and withdrawal of Indian monsoon and the variables that possibly support these mechanisms are shown. These variables that are considered as predictors in different researches relating to ISMR forecasting are grouped into 4 categories, namely: regional conditions, ENSO indicators, cross equatorial flow, and global hemispheric conditions. The techniques used for developing forecasting models, are included along with their strengths and shortfalls in "prediction models review". The prediction techniques and predictors used in the history of ISMR prediction, their success, their failures, the reasons of their failures and the outcome of the research, are depicted in "prediction for ISMR review". The research gaps are listed and the problem statement of the current work is defined at the end of the chapter. Details of the general data preparation considered in the current research is also given.

CHAPTER 3

ISMR PREDICTION USING DATA MINING AND STATISTICAL APPROACHES

3.1 INTRODUCTION

In Chapter 2, the process of evolution in the field of Indian Summer Monsoon Rainfall prediction in terms of predictors and models has been discussed. It gives a detailed study, that outlines the data sources used by various researchers in the field. Knowledge about the general conditions that prevail during the onset, progress and withdrawal of monsoon and their relationship with the predictors is discussed. In the current chapter, we present a model developed, based on ensemble techniques including data mining and statistical approaches for better prediction of Indian Summer Monsoon Rainfall (ISMR).

ISMR and its Long Range Forecast (LRF) have been of prime importance to India; it accounts for almost 75% of annual rains experienced by the subcontinent. There has been a lot of effort put in by the scientists for accurate prediction of ISMR. The achieved success in LRF of ISMR is limited, but improving day by day. The statistical techniques with different predictors have given limited success in this endeavor. The limitations in predictor selection and statistical techniques may be the cause of limited success in this field. Considering only the effects of individual predictors than their combination. Heavy dependence on ENSO indicators, random choice of predictors instead of following systematic approach and influence of climate variations on prediction are the basic problems of predictor selection. In general, a majority of the statistical techniques have given fairly good results, compared to GCMs, but they still require fine tuning.

The monthly and seasonal variation of rainfall is predictable to some extent, due to the slowly varying boundary forcing conditions such as, Sea Surface Temperature, snow cover, soil moisture etc. The semi permanent system setup during monsoon involves many known predictors and also gives some knowledge of the monsoon mechanism. The fact, that, this setup happens every year with some variations, is made use off to help forecast ISMR, depending on state of the pre monsoon boundary forcing conditions.

In this work we propose to use a large number of known predictors of ISMR. The result of the combination of predictors is used, to effectively employee selected predictors for prediction. Including large number of predictors in the study requires a lot of memory and CPU time. To deal with the constraints, ensemble techniques involving association rule mining, cluster membership and simple logistic regression are used in the process of prediction.

3.2 DATABASE

For the purpose of prediction of Indian Summer Monsoon Rainfall, a total of 36 variables have been used; the list of main categories and the climate variables considered under them are as follows:

1) Darwin Sea Level Pressure (DSLP)

2) ENSO-SO index (ENSO-SOI) - NINO3, NINO4, NINO3.4 and ENSO-SOI indices are considered.

3) EQUINOO - EQUINOO phase and EQUWIN index.

4) March-April-May minimum (min), maximum (max), average air temperature of Jodhpur, Ahmedabad, Bombay, Indore, Sagar, and Akola (Indian cities).

5) MSLP - Average MSLP of Jodpur, Ahmedabad, Bombay, Indore, Sagar, and Akola (Indian cities).

6) 200hPa meridional component of wind for May at Bombay, Delhi, Madras, Nagpur, and Srinagar (Indian cities).

7) Rainfall data for the whole of India.

The sources, paper references and the data preprocessing steps involved to prepare the data for the experiment have been discussed in the section 2.7 of the previous chapter. The climate variables and their representations are given in Table 3.1.

Table 3.1: 36 chosen ISMR predictors from existing literature and their representation.

Sl.no	Predictors	Representation
1	Darwin Sea Level Pressure in millibar	DSLP
2	Equatorial Wind speed (EQWIN) in meters per second	EQWIN
3	Equatorial Indian Ocean Oscillation (EQUINOO)	EQUINOO-Phase
4	Jodhpur Mean Sea Level Pressure in millibar	Jd-Mslp
5	Ahmedabad Mean Sea Level Pressure in millibar	Ah-Mslp
6	Bombay Mean Sea Level Pressure in millibar	Bb-Mslp
7	Indore Mean Sea Level Pressure in millibar	Id-Mslp
8	Sagar Mean Sea Level Pressure in millibar	Sa-Mslp
9	Akola Mean Sea Level Pressure in millibar	Ak-Mslp
10	NINO 3.4 index	NINO3.4
11	NINO 3 index	NINO3
12	NINO 4 index	NINO4
13	SOI	Soi
14	Maximum Temperature Jodhpur in °C	Jd-Maxtmp
15	Minimum Temperature Jodhpur in °C	Jd-Mintmp
16	Mean Temperature Jodhpur in °C	Jd-Meantmp
17	Maximum Temperature Ahmedabad in °C	Ah-Maxtmp
18	Minimum Temperature Ahmedabad in °C	Ah-Mintmp
19	Mean Temperature Ahmedabad in °C	Ah-Meantmp
20	Maximum Temperature Bombay in °C	Bb-Maxtmp
21	Minimum Temperature Bombay in °C	Bb-Mintmp

Table 3.1: 36 chosen ISMR predictors from existing literature and their representation.

22	Mean Temperature Bombay in °C	Bb-Meantmp
23	Maximum Temperature Indore in °C	Id-Maxtmp
24	Minimum Temperature Indore in °C	Id-Mintmp
25	Mean Temperature Indore in °C	Id-Meantmp
26	Maximum Temperature Sagar in °C	Sa-Maxtmp
27	Minimum Temperature Sagar in °C	Sa-Mintmp
28	Mean Temperature Sagar in °C	Sa-Meantmp
29	Maximum Temperature Akola in °C	Ak-Maxtmp
30	Minimum Temperature Akola in °C	Ak-Mintmp
31	Mean Temperature Akola in °C	Ak-Meantmp
32	Average Wind Speed Bombay in meters per second	Bb-Wnd
33	Average Wind Speed Delhi in meters per second	Di-Wnd
34	Average Wind Speed Madras in meters per second	Ma-Wnd
35	Average Wind Speed Nagpur in meters per second	Ng-Wnd
36	Average Wind Speed Srinagar in meters per second	Sr-Wnd

3.2.1 Association rule mining definitions and process used

Association rules are derived from historic data by analyzing the frequency in which two or more events occur together, this frequency is helpful in deriving if/then rules. The concepts of Support and Confidence in data mining are useful to understand the algorithms presented in this and subsequent chapters, and are defined below.

Definition 1. Let $I = I_1, I_2, I_3, ...I_n$ be a set of n attributes called items. Let $D = T_1, T_2, T_3, ...T_m$ be a set of m transactions where each transaction has a transaction ID T_i, such that each transaction contains a subset of items in I. An association rule is an implication of the form $X \Rightarrow Y$ such that $X, Y \subseteq I$ and $X \cap Y = \emptyset$. X is called the antecedent and Y is called the consequent (Aggarwal, 2015).

Support and Confidence are the two parameters to determine the interestingness of a rule. Different measures, such as, lift, f-measure and cosine can be used to estimate the interestingness of the rule. Support and Confidence are well known in the data mining community for this task.

Definition 2. Support is defined as the proportion of transactions that contains X.

$Support(X) = \frac{X.count}{m}$ where m is the total number of transactions.

Definition 3. Confidence of $X \Rightarrow Y$ is defined as the measure of how often items in Y appear in transactions that contain X.

$Confidence(X \Rightarrow Y) = \frac{(X \cup Y).count}{X.count}$

The concepts of super set and subset can be defined as.

Definition 4. Set C is a superset of set D if set C contains all of the elements of set D. When all elements of D are also elements of C, D is called the subset of C. for example Let $C = \{A, B, C\}$ and $D = \{A, B\}$. Here set C is a super set of set D and set D is the subset of set C.

3.3 METHODOLOGY

In total, 36 known ISMR predictors are used in this work. The aim of the algorithm is to exploit the effect of predictor combinations by identifying associations between them, so as to achieve good accuracy in prediction. For exploiting associations between variables, association rule mining is employed. Association rules are based on frequent itemsets. For identifying frequent itemsets the repetition of the combinations of variables are counted in the entire dataset used. The processing (CPU) time for performing these comparisons & calculations and the memory requirement, increases as the dimensionality and size of the dataset increases. To process huge data with large number of predictors, the algorithm should be able to handle large data without posing memory and CPU time constraints. To achieve these requirements, the algorithm, depicted below, uses both data mining and statistical approaches. Feature selection and dimensionality reduction are achieved by employing data mining techniques, whereas, the statistical technique called regression is used for prediction. The preprocessing task on dataset is essential to prepare data for further usage. The general database discussed in section 2.7 give various data preparation steps employed for each considered predictor.

3.3.1 Algorithm steps

The overall steps of the algorithm involved in the process of ISMR prediction is given below, these steps are explained in detail, in the following algorithm.

```
Algorithm 1: An algorithm for long range prediction of ISMR
  Input: Prepared dataset, Output: Predicted rainfall range.

    1. Identify closed itemsets
    2. Derive association rules
    3. Select top association rules for each rainfall range specified in Table 3.2
    4. Use antecedent variables as selected features
    5. Convert numerical data to nominal data for dimensionality reduction
    6. Apply cluster membership using Expectation Maximization (EM) on
       selected features
    7. Apply simple logistic regression function on the cluster membership data,
       for predicting the rainfall range.
```

Table 3.2: Rainfall categories of ISMR and their value ranges. Z represents the average rainfall value of June, July, August and September (JJAS) in mm.

Sl.no	**Category**	**JJAS Standardized Rainfall (SR)and Rainfall Value (RV) ranges**
1	Flood	$\geq$ SR=1, RV=957.9628364
2	Excess	SR=0.5, RV=915.6773641 $\leq$ Z $<$ SR=1, RV=957.9628364
3	Normal	SR=-0.5, RV=831.1064196 $<$ Z $<$ SR=0.5, RV=915.6773641
4	Deficit	SR=-1, RV=788.8209474 $<$ Z $\leq$ SR=-0.5, RV=831.1064196
5	Drought	$\leq$ SR=-1, RV=788.8209474

Step 1. An itemset is closed if none of its immediate supersets has the same

support as this itemset. For example, if $\{Bread, Milk\}$ is an itemset that has $support = 4$, and all of its supersets have $support < 4$, then $\{Bread, Milk\}$ is a closed itemset. Closed itemset mining is a technique to overcome large frequent itemset problems (An itemset whose support is greater than or equal to specified minimum support threshold is called frequent itemset.) that cause memory and CPU time constraints, without leaving out important association rules and compromising on the quality of the rules (Aggarwal, 2015).

Step 2. The basic concept of association rule mining (Agrawal et al., 1993) is to find the set of all itemsets that frequently occur (Repeat) in the database. Furthermore, these frequent itemsets are used to frame rules that tell us how a set of items influence the presence of another set (Aggarwal, 2015).

Step 3. Rules with high confidence are made out of highly correlated items. The degree of dependence between items in a transactional dataset is described by the confidence. Hence, to ensure that the predictors in the rule have the highest degree of correlation with rainfall behavior, only rules with the highest confidence in the categories of *Flood*, *Excess*, *Normal*, *Deficit*, and *Drought* rainfall are extracted. Table 3.2 depicts the range values of the mentioned rainfall categories, on the basis of the calculations presented in Dhanya and Nagesh Kumar (2009).

Step 4. The left-hand side proposition of the rule is called the antecedent and the right-hand side is known as the consequent. An association rule states that the antecedent occurs with the consequent ($X \Rightarrow Y$). Thus, designating antecedent items, as selected features, give well related predictors that imply a particular rainfall category, which is a consequent, resulting in better forecasting.

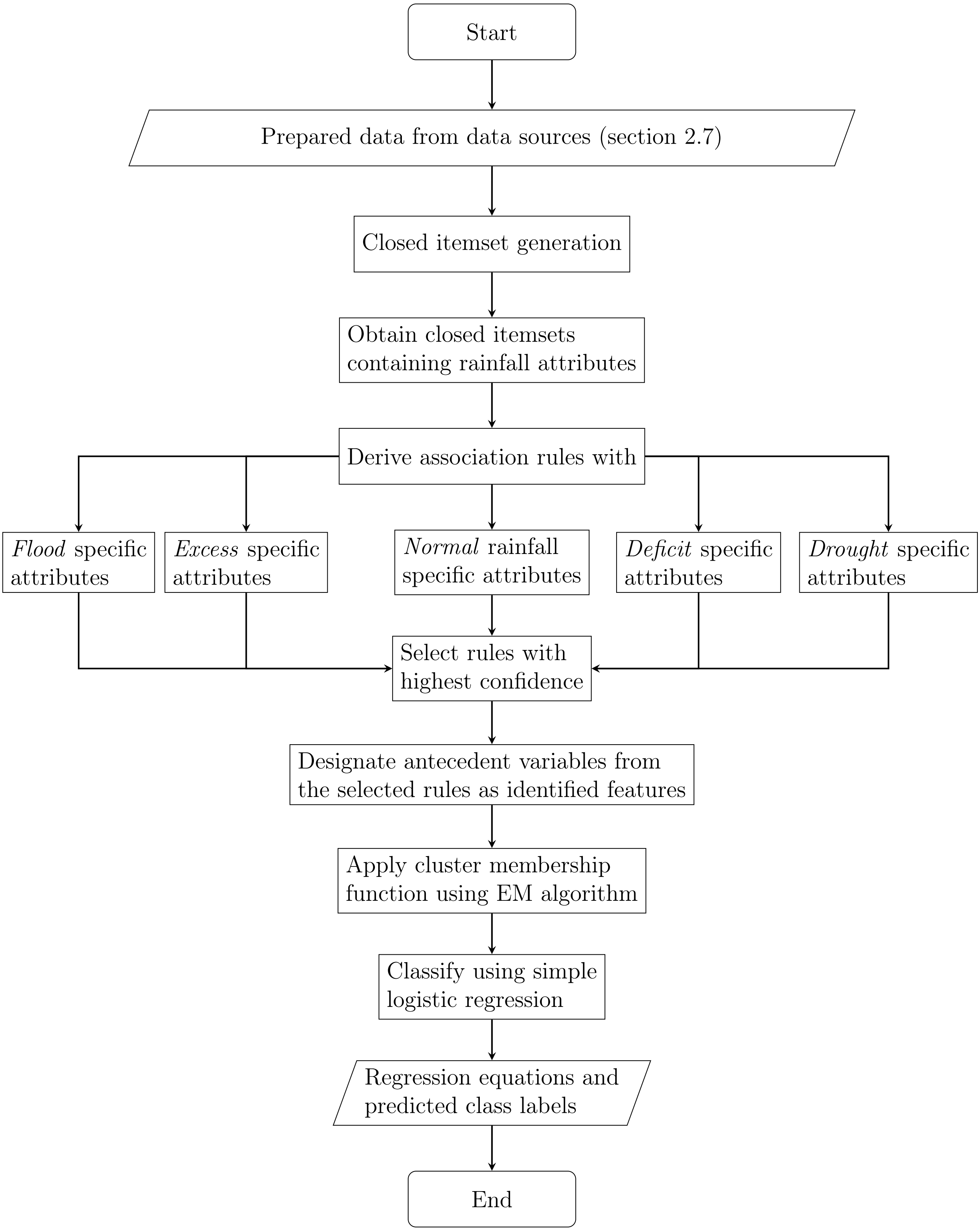
Step 5. Converting numerical data to nominal: Nominal variable is also called categorical variable, it takes the name of the category. Nominal variables can be divided into two or more categories. For example, gender is a nominal variable having two categories male and female. This data has an advantage, namely; reduction in dimensionality, where a range of data falls into one category, thereby reducing the dimension of the problem, resulting in reduction of computational cost.

Step 6. Clustering is a process of grouping a set of data into meaningful subclasses based on data similarity, leading to the reduction in dimensionality of the problem. The capability of the EM technique to consider hidden effects brought in by unobserved or absent variables, is used for better prediction (Jin and Han, 2011).
Step 7. Logistic regression function is used as a prediction tool. Being a probabilistic classifier, it predicts a probability distribution over a set of classes, rather than only outputting the most likely class, the sample should belong to. Probabilistic classifiers provide classification with a degree of certainty, including rejection of the option (make a decision if sufficiently confident) (Lu and Getoor, 2003).

Figure 3.1 shows the flow diagram of the algorithm. The set of input feature vector is used for generating association rules. The algorithm processes the data and generates closed frequent itemsets with respect to the minimum support supplied. The derived closed frequent itemsets are high in number with some combinations that do not contain the required attributes, such as rainfall variables. These unwanted combinations, are to be removed by cutting down the number of closed frequent itemsets, to a smaller set, this is done by removing closed itemsets that do not contain rainfall variable. These closed frequent itemsets are then used for association rule mining, without specifying the minimum confidence threshold. All association rules in each category of rainfall are grouped together. In each group, the rules with the highest confidence are chosen. This step is conducted after sorting rules in descending order of the confidence values. The selected rules of all groups are combined, and the unique antecedent attributes (selected features) of rules are used for further processing.

Data of selected features are further clustered for dimensionality reduction. The considered cluster membership function employs the EM mechanism. EM clustering algorithm computes probabilities of cluster memberships based on one or more probability distributions. The goal of the clustering algorithm then is to maximize the overall probability or likelihood of the data. The EM algorithm

Figure 3.1: Flow chart of proposed ISMR prediction model.


Start
Prepared data from data sources (section 2.7)
Closed itemset generation
Obtain closed itemsets containing rainfall attributes
Derive association rules with
Flood specific attributes
Excess specific attributes
Normal rainfall specific attributes
Deficit specific attributes
Drought specific attributes
Select rules with highest confidence
Designate antecedent variables from the selected rules as identified features
Apply cluster membership function using EM algorithm
Classify using simple logistic regression
Regression equations and predicted class labels
End

gives cluster membership values to the data, thereby reducing the dimension of the data to the number of clusters produced by the EM algorithm. Finally, simple logistic regression function is applied on the data, with cluster membership for classification, serving as a prediction tool.

3.3.2 Applying the algorithm on ISMR data

The value ranges are obtained by dividing rainfall data into five ranges; namely: *Flood*, *Excess*, *Normal*, *Deficit*, and *Drought*. The threshold values between the categories are determined by identifying the values at ±1 and ±0.5 Standard Deviation (SD) of the standardized data (Rajeevan et al., 2006) (Prabhu et al., 2010). This felicitates ranging the rainfall; rainfall values that fall beyond 1(SD) are grouped into the *Flood* range, rainfall values that fall in between 1(SD) and 0.5(SD) are grouped into the *Excess* range, whereas rainfall values that fall in between 0.5(SD) and -0.5(SD) are grouped into the *Normal* range. Similarly the value ranges for *Deficit*, and *Drought* are decided.

The algorithm is intended to predict a particular year's rainfall, as *Flood*, *Excess*, *Normal*, *Deficit*, or *Drought*, based on related climate attributes. The range thresholds can be more precisely defined by including data from more number of years. Rainfall may also be divided into seven ranges instead of five. A total of 36 predictors, and data of 37 years (1969-2005) are considered in this study. Each predictor is divided into five ranges, namely, high, max, normal, min, and low.

To use the available data for data mining, similar instances are clubbed into one range and these ranges are given names. Association mechanism generally includes simple predictive rules, which work well with binned/ranged data (Ordonez and Zhao, 2011) called discretized data. Discretization leads to information loss, but however, the amount of data loss is proportional to the number of bins/ranges created. A few number of bins leads to a huge loss of information. The amount of information loss reduces as the number of bins increase, thereby smoothening the curve, although, it is not a linear one (Holcombe and Paton, 1997). In accordance with this observation, experiments are carried out by gradually increasing the number of bins. Finally, dividing the data into five ranges resulted in good model results, indicating a tradeoff between information loss and model performance. In the current work, data are divided into five ranges. The threshold values are determined by identifying the values at ±1 and ±0.5 Standard Deviation (SD) (Dhanya and Nagesh Kumar, 2009). These ranges are named in a pattern. For example, DSLP has five ranges and is represented as DSLP-high, DSLP-max, DSLP-normal, DSLP-min, and DSLP-low. Table 3.3 depicts the predictors and their representations with value ranges, calculated as explained above. This binned data is further subjected to association rule mining.

Association rule-based feature selection gives a feel of the data before selecting the features. This ability, allows the chosen method to perform better than the other techniques that consider a subset of data for feature selection. A comparison of feature selection methods has been carried out previously. The experimental results indicate that, the feature selection algorithm, based on association rules, yields significant reduction in the number of influential features required for classification and simultaneously maintains acceptably high classification accuracy (Xie et al., 2009). However, frequent itemset-based association rule mining, with low minimum support threshold, may result in large memory shortage for high-dimensional datasets. Unfortunately, the response time of highly optimized frequent pattern mining algorithms ranges from minutes to hours, based on the size of the dataset, the threshold of minimum support, and the minimum confidence used to specify rule quality (Aggarwal et al., 2014). Consequently, the closed frequent itemset mining-based association rule mining technique can deal with memory constraint and Central Processing Unit (CPU) time constraint without affecting the efficiency of prediction.

Table 3.3: Predictors and their value ranges used for ISMR prediction. Z represents a predictor value. The words shown in the brackets in the column headings indicate the representation in the feature list.

Sl.no	**Variable**	**Low (low)**	**Minimum (min)**	**Normal (*Normal*)**	**Maximum (max)**	**High (high)**
1	DSLP	≤ 1.703636	$1.703636< Z \leq 2.266232$	$2.266232< Z < 3.391425$	$3.391425 \leq Z < 3.954022$	≥ 3.95402
2	EQWIN	$\leq$-1.20244	$-1.20244< Z \leq -0.80041$	$-0.80041< Z <0.00365$	$0.0036\leq Z < 0.405679$	≥ 0.405679
3	EQUINOO-Phase	Negative	Neutral	Positive		
4	Jd-Mslp	≤ 1005.4329	$1005.4329< Z \leq 1005.8741$	$1005.8741< Z < 1006.7565$	$1006.7565\leq Z < 1007.1978$	≥ 1007.1978
5	Ah-Mslp	≤ 1007.7913	$1007.7913< Z \leq 1008.1542$	$1008.1542< Z < 1008.88$	$1008.88\leq Z < 1009.2429$	≥ 1009.2429
6	Bb-Mslp	≤ 1008.1825	$1008.1825< Z \leq 1008.5597$	$1008.5597< Z < 1009.3141$	$1009.3141\leq Z < 1009.6913$	≥ 1009.6913
7	Id-Mslp	≤ 1006.90475	$1006.90475< Z \leq 1007.29112$	$1007.29112< Z < 1008.06384$	$1008.06384\leq Z < 1008.4502$	≥ 1008.4502
8	Sa-Mslp	≤ 1005.53062	$1005.53062< Z \leq 1005.99054$	$1005.99054< Z <1006.91037$	$1006.91037\leq Z < 1007.37028$	≥ 1007.37028
9	Ak-Mslp	≤ 1006.81531	$1006.81531< Z \leq 1007.21126$	$1007.21126< Z < 1008.00316$	$1008.00316\leq Z < 1008.39911$	≥ 1008.39911
10	NINO3.4	≤ 26.90884	$26.90884< Z \leq 27.23609$	$27.23609< Z < 27.89058$	$27.89058\leq Z <28.21782$	≥ 28.21782
11	NINO3	≤ -0.97977	$-0.97977< Z \leq -0.54651$	$-0.54651< Z <0.320022$	$0.320022\leq Z <0.753287$	≥ 0.753287
12	NINO4	≤ -0.24013	$-0.24013< Z \leq -0.03196$	$-0.03196< Z <0.38439$	$0.38439\leq Z <0.592564$	≥ 0.592564
13	Soi	≤ -0.67453	$-0.67453< Z \leq -0.30078$	$-0.30078< Z <0.446723$	$0.446723\leq Z <0.820472$	≥ 0.820472
14	Jd-Maxtmp	≤ 36.36947	$36.36947< Z \leq 36.87379$	$36.87379< Z <37.88243$	$37.88243\leq Z <38.38675$	≥ 38.38675
15	Jd-Mintmp	≤ 20.26153	$20.26153< Z \leq 20.69752$	$20.69752< Z <21.56951$	$21.56951\leq Z <22.0055$	≥ 22.0055
16	Jd-Meantmp	≤ 28.36508	$28.36508< Z \leq 28.81045$	$28.81045< Z <29.70118$	$29.70118\leq Z < 30.14654$	≥ 30.14654
17	Max Ah-Maxtmp	≤ 36.27315	$36.27315< Z \leq 36.69293$	$36.69293< Z <37.53248$	$37.53248\leq Z <37.95225$	≥ 37.95225
18	Ah-Mintmp	≤ 20.61638	$20.61638< Z \leq 20.95684$	$20.95684< Z <21.63776$	$21.63776\leq Z <21.97822$	≥ 21.97822
19	Ah-Meantmp	≤ 28.48823	$28.48823< Z \leq 28.84661$	$28.84661< Z <29.56339$	$29.56339\leq Z <29.92177$	≥ 29.92177
20	Bb-Maxtmp	≤ 33.1398	$33.1398< Z \leq 33.38044$	$33.38044< Z < 33.86172$	$33.86172\leq Z <34.10236$	≥ 34.10236
21	Bb-Mintmp	≤ 22.07126	$22.07126< Z \leq 22.33712$	$22.33712< Z <22.86883$	$22.86883\leq Z <23.13468$	≥ 23.13468
22	Bb-Meantmp	≤ 27.63142	$27.63142< Z \leq 27.87173$	$27.87173< Z <28.35233$	$28.35233\leq Z <28.59263$	≥ 28.59263
23	Id-Maxtmp	≤ 37.44705	$37.44705< Z \leq 37.80866$	$37.80866< Z <38.53188$	$38.53188\leq Z <38.89349$	≥ 38.89349
24	Id-Mintmp	≤ 21.23	$21.23< Z \leq 21.50865$	$21.50865< Z <22.06594$	$22.06594\leq Z <22.34459$	≥ 22.34459
25	Id-Meantmp	≤ 29.38287	$29.38287< Z \leq 29.68083$	$29.68083< Z <30.27674$	$30.27674\leq Z <30.5747$	≥ 30.5747

Table 3.4: Predictors and their value ranges used for ISMR prediction. Z represents a predictor value. The words shown in the brackets in the column headings indicate the representation in the feature list.

Sl.no	Variable	Low (low)	Minimum (min)	Normal (Normal)	Maximum (max)	High (high)
26	Sr-Maxtmp	$\leq$36.96793	36.96793$<$ Z $\leq$37.39045	37.39045$<$ Z $<$38.2355	38.2355$\leq$ Z $<$38.65802	$\geq$38.65802
27	Sr-Mintmp	$\leq$20.61844	20.61844$<$ Z $\leq$20.94882	20.94882$<$ Z $<$21.60956	21.60956$\leq$ Z $<$21.93994	$\geq$21.93994
28	Sr-Meantmp	$\leq$28.83476	28.83476$<$ Z $\leq$29.19042	29.19042$<$ Z $<$29.90174	29.90174$\leq$ Z $<$30.2574	$\geq$30.2574
29	Ak-Maxtmp	$\leq$38.68746	38.68746$<$ Z $\leq$39.05576	39.05576$<$ Z $<$39.79235	39.79235$\leq$ Z $<$40.16065	$\geq$40.16065
30	Ak-Mintmp	$\leq$22.84749	22.84749$<$ Z $\leq$23.15618	23.15618$<$ Z $<$23.77355	23.77355$\leq$ Z $<$24.08224	$\geq$24.08224
31	Ak-Meantmp	$\leq$30.79263	30.79263$<$ Z $\leq$31.11854	31.11854$<$ Z $<$31.77038	31.77038$\leq$ Z $<$32.09629	$\geq$32.09629
32	Bb-Wnd	$\leq$-0.7891201	-0.7891201$<$ Z $\leq$0.5224062	0.5224062$<$ Z $<$3.1454587	3.1454587$\leq$ Z $<$4.456985	$\geq$4.456985
33	Di-Wnd	$\leq$-0.6416854	-0.6416854$<$ Z $\leq$0.8186305	0.8186305$<$ Z $<$3.7392621	3.7392621$\leq$ Z $<$5.199578	$\geq$5.199578
34	Ma-Wnd	$\leq$-0.0969792	-0.0969792$<$ Z $\leq$0.8794324	0.8794324$<$ Z $<$2.8322557	2.8322557$\leq$ Z $<$3.8086673	$\geq$3.8086673
35	Ng-Wnd	$\leq$-0.992905	-0.992905$<$ Z $\leq$0.7429549	0.7429549$<$ Z $<$4.2146747	4.2146747$\leq$ Z $<$5.9505346	$\geq$5.9505346
36	Sr-Wnd	$\leq$-4.6976358	-4.6976358$<$ Z $\leq$-2.3603306	-2.3603306$<$ Z $<$2.3142798	2.3142798$\leq$ Z $<$4.651585	$\geq$4.651585

Dimensionality reduction is the process of retaining the most important attributes and removing the noisy or irrelevant attributes in order to reduce computational cost. To achieve good results in classification, only those attributes that are likely to affect the behavior under study, should be considered. The effect of hidden and unknown variables must also be considered. A filter with this combination, when used in selecting attributes for classification, is expected to yield the best results. Clustering, using EM, is particularly apt in the current scenario because of the hidden effect of the unknown predictors on the rainfall variable. EM is an algorithm for maximizing a likelihood function, when some of the variables in the model are latent variables and works on the basis of parameter estimation. The algorithm starts with an initial estimate of what each parameter might be, it computes the likelihood that each parameter produces the data point (expectation). Calculate weights for each data point, based on the likelihood of it being produced by a parameter. Combine these weights, together with the data, to compute a better estimate for the parameters (maximization). Repeat expectation and maximization steps, until the parameter estimate converges. The application of the EM-based clustering mechanism has grouped 37 years of data into eight clusters, clearly separating *Drought*, *Deficit*, *Excess*, and *Flood* behavior of rainfall. *Normal* has four separate clusters, specifying four types in itself. EM is a widely used technique for density estimation and model-based clustering. EM calculates the maximum likelihood estimates of the parameters for each candidate of the model (Dempster et al., 1977). The algorithm can consider the hidden effect brought in by unobserved or absent variables on the candidates of the model (Dempster et al., 1977). It can optimize a large number of variables simultaneously and provide good estimates, for any missing information.

Simple logistic algorithm is a classification algorithm that uses logistic regression and tree induction for supervised learning tasks. The algorithm constructs a logistic regression model by stepwise fitting, to select relevant attributes, naturally (Landwehr et al., 2005). This classification algorithm is probabilistic in nature and is used as a predictor classifier in the current research. The main advantage of logistic regression is that explicit class probability estimates are produced rather than merely a classification (Niels et al., 2005; Sumner et al., 2005). Thus, it can output a confidence value. It refrains from outputting a result when its confidence of choosing any particular output is too low. The model is tested in three cases with ten fold cross validation, five fold cross validation and 70% training & 30% testing set.

3.4 RESULTS AND DISCUSSION

In total, 25 out of 36 features are selected from the closed itemset-based association rule mining technique, which are useful for ISMR prediction. The selected attributes and association rules are listed in Table 3.5.

Theories and conclusions of research can be used for justifying many of the rules derived in the current study. The land and ocean temperature gradient initially drives monsoon flow in the early Indian summer monsoon season. Once the monsoon is set, convective heating maintains the tropospheric temperature gradient and drives monsoon flow throughout the season (Taniguchi et al., 2010). This phenomenon is evident from the association rules obtained. In all categories of rainfall, the rules are composed of land, ocean, and atmospheric attributes. For instance the association rules when observed, contain rules such as EQWIN-normal, NINO4-high, EQUINOO-Phase-neutral ⇒ *Drought*; DSLP-normal, EQWIN-min, EQUINOO-Phase-negative, NINO4-normal ⇒ *Drought*; EQUINOO-Phase-neutral, NINO3.4-normal, Ma-Wnd-high ⇒ *Excess.* Warm SSTs in central and eastern parts of the equatorial Pacific region are associated with low monsoon rainfall. Trade winds from South America normally blow westward towards Asia during the summer monsoon. Increased SSTs of the Pacific Ocean weakens these winds. Therefore, the moisture content decreases, resulting in reduction and uneven distribution of rainfall across the Indian subcontinent (Angell, 1981; Khandekar and Neralla, 1984). This phenomenon is also depicted by NINO4 and NINO3 in the rules specifying *Drought.* El Niño, ENSO, EQWIN, and EQUINOO can explain much of the ISMR variability (Gadgil, 2003). The

Table 3.5: Association rules with their confidence and the selected attributes for prediction of ISMR.

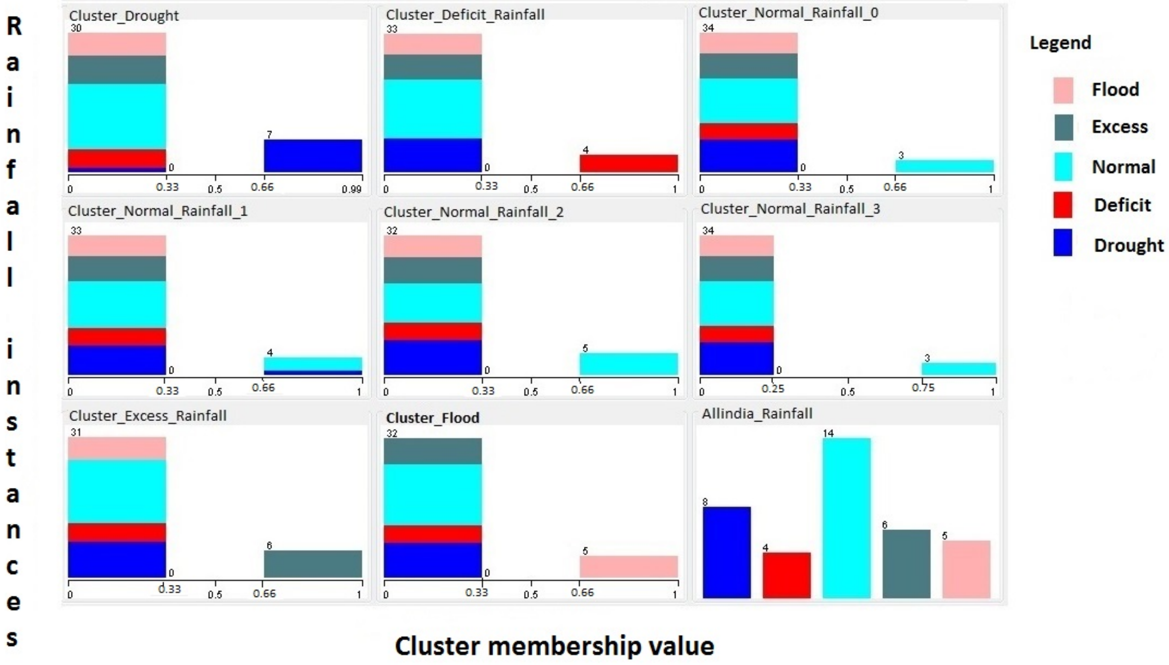
All India Rainfall ranges	Rule	Confidence	Selected attributes
Drought	Sa-Mslp-min Ah-Mslp-min Jd-Mslp-min ⇒ *Drought* EQWIN-normal NINO4-high EQUINOO-Phase-neutral ⇒ *Drought* NINO3-normal DSLP-normal Soi-normal ⇒ *Drought* DSLP-normal EQWIN-min EQUINOO-Phase-negative NINO4-normal ⇒ *Drought* Ah-Maxtmp-normal EQWIN-normal Id-Maxtmp-normal EQUINOO-Phase-neutral Ah-Meantmp-normal ⇒ *Drought*	0.7272 0.7272 0.7272 0.7272 0.7272	Sa-Mslp Ah-Mslp Jd-Mslp EQWIN NINO4 EQUINOO-Phase NINO3 DSLP Soi Ah-Maxtmp Id-Maxtmp Ah-Meantmp
Deficit	Sr-Wnd-normal Bb-Maxtmp-normal NINO4-normal ⇒ *Deficit*	0.7272	Sr-Wnd Bb-Maxtmp NINO4
Normal	NINO3-normal Bb-Mintmp-normal ⇒ *Normal* NINO3-normal Jd-Maxtmp-normal EQUINOO-Phase-neutral Id-Meantmp-normal ⇒ *Normal*	0.8750 0.8750	NINO3 Bb-Mintmp Jd-Maxtmp EQUINOO-Phase Id-Meantmp
Excess	EQUINOO-Phase-neutral NINO3.4-normal Ma-Wnd-high ⇒ *Excess* Jd-Mintmp-normal Id-Mslp-min Jd-Mslp-min ⇒ *Excess*	0.7272 0.7272	EQUINOO-Phase NINO3.4 Ma-Wnd Jd-Mintmp Id-Mslp Jd-Mslp
Flood	EQWIN-high ⇒ *Flood* Sa-Meantmp-normal Ak-Maxtmp-normal Sa-Maxtmp-normal NINO4-normal Ak-Meantmp-normal ⇒ *Flood*	0.5 0.5	EQWIN Sa-Meantmp Ak-Maxtmp Sa-Maxtmp NINO4 Ak-Meantmp

rules for the extremes (*Droughts* and *Floods*) are in accordance with the previous studies conducted by Gadgil et al. (2004). A positive EQUINOO phase is associated with a suppression of convection over the eastern EEIO and an enhancement of convection over the WEIO, leading to heavy moisture-bearing cloud formation producing heavy rains over India. A negative phase is associated with its reverse phenomenon. Although enhanced convection over the WEIO is associated with an easterly (East to West) anomaly of the equatorial surface wind, resulting in above normal rainfall over India. Enhanced convection over the EEIO is associated with a westerly (West to East) anomaly of the equatorial surface wind, resulting in *Drought* over India. When global conditions are normal, EQWIN can decide the fate of ISMR (Gadgil et al., 2004). Considering the rule DSLP-normal, EQWIN-min, Euinoo-Phase-negative, NINO4-normal ⇒ *Drought*, DSLP and NINO4 are normal global conditions and the EQWIN index is lesser than normal, specifying negative EQUINOO phase resulting in *Drought.* Another rule EQWIN-high ⇒ *Flood* shows, high EQWIN specifying the positive EQUINOO phase resulting in *Floods.* Both rules logically justify the theory concluded in previous research (Gadgil et al., 2004). High-temperature and low-atmospheric pressure regions that gradually develop over the Indian subcontinent during premonsoon months March, April, and May, lead to large-scale influx of maritime air from the South Indian Ocean into India. Research on temperature and pressure of West Central India, shows a negative correlation with ISMR (Parthasarathy et al., 1992), justifying the rule Jd-Mintmp-normal, Id-Mslp-min, Jd-Mslp-min ⇒ *Excess.* Association rules also have shown interesting instances of atmospheric conditions, such as the combination of normal NINO3.4 condition with high Madras winds giving rise to *Excess* rainfall. Madras has a general lack of cloud cover or rain during the summer months, and land heating influences a strong sea breeze component in the lower troposphere (Krishnamurti and Kishtawal, 2000). When ENSO conditions and Indian Ocean Dipole conditions are neutral, the land conditions may become crucial in deciding the variation in monsoon rainfall. The rule EQUINOO-Phase-neutral, NINO3.4-normal, Ma-Wnd-high ⇒ *Excess* is an illustration to this condition. Associative prediction greatly expands the area that is searched for clues. Much of regression analysis deals with discerning, in which independent clues are helpful for better prediction (Daellenbach et al., 2012).

Better classification performance, depends on the way the data are prepared, so as to incorporate the inter and intra relationships between the attributes in the model and their relationships with the rainfall pattern. The latter is taken care of by feature selection, whereas the former is taken care of by the cluster membership mechanism, with the aid of EM algorithm. The EM cluster mechanism reduces the dimensionality from 25 selected attributes to eight clusters, conducted by cross validation for automatically determining the number of distinct groups in the data. The instances are then given membership value, which is the probability of it belonging to a particular cluster. Figure 3.2 displays the cluster information. In the first cluster, seven instances with membership value between 0.663 and 0.994 (higher probability), fall under the *Drought* category. Instances having high membership values grouped under this cluster, are *Drought* instances, making it easy for the classifier in prediction. Hence, this cluster is named Cluster_Drought. The same follows with other clusters as well. Among the eight clusters, Cluster_Drought, Cluster_Deficit_Rainfall, Cluster_Excess_Rainfall, and Cluster_Flood clearly show their association with *Drought*, *Deficit*, *Excess* and *Floods*, respectively. The other four clusters are associated with *Normal* rainfall, without having uniform features between instances, to clearly mark them into a single cluster.

Simple logistic regression takes a single parameter and produces probabilities. Predicting class probabilities is better than predicting classes. The simple logistic regression model with regression equation specifying eligibility to a particular class of rainfall is shown in Table 3.6. The regression equation shows that Cluster_Drought is related to *Drought* and *Normal* rainfall, with different slopes and intercepts. The regression equation is of the form $y = mx + c$ here m is the slope and c is the intercept. The Cluster_Deficit has a clear-cut membership value difference between other instances (0 to 0.33) and all four *Deficit*

Figure 3.2: Clusters of ISMR data from 1969-2005 using Expectation-Maximization (EM algorithm).



rainfall instances (0.665 to 0.998); hence, this cluster can be uniquely related to *Deficit* rainfall. Cluster_Excess_Rainfall has a clear-cut membership value difference between other instances (0 to 0.33) and all six *Excess* rainfall instances (0.667 to 1); hence, this cluster can be uniquely related to *Excess* rainfall. Cluster_Flood can be uniquely related to *Floods* because of a clear-cut membership value difference between other instances (0 to 0.33) and all five *Flood* rainfall instances (0.667 to 1). Cluster_Normal_Rainfall_0, Cluster_Normal_Rainfall_1, Cluster_Normal_Rainfall_2, and Cluster_Normal_Rainfall_3 are not considered in the classification scenario, as these clusters do not show clear or generalized association with a particular class of rainfall. The association of Cluster_Drought with both *Drought* and *Normal* rainfall may lead to incorrect classification. In spite of Cluster_Normal_Rainfall_0, Cluster_Normal_Rainfall_1, Cluster_Normal_Rainfall_2, and Cluster_Normal_Rainfall_3 relating to *Normal* rainfall, Cluster_Drought is considered the predictor of *Normal* rainfall because the former four clusters do not cover all instances of *Normal* rainfall. In addition, Cluster_Drought contains all instances of *Normal* rainfall and has only one instance of *Drought* with almost the same membership values. One instance of *Drought* has similar features to three *Normal* rainfall instances, and can be found in Cluster_Normal_Rainfall_1. Cluster_Drought has only one instance of *Drought* with all instances of *Normal* rainfall in the same membership range, whereas, Cluster_Deficit, Cluster_Excess_Rainfall and Cluster_Flood have all instances of *Drought* with all instances of *Normal* rainfall in the same membership range. Hence, Cluster_Drought is the optimal choice as a predictor for *Normal* rainfall. The slopes and intercepts define the linear relationship between two variables and can be used to estimate an average rate of change. The greater the magnitude of slope the greater the rate of change. From simple logistic equation, $Drought : -0.82 + [ClusterDrought] * 3.67$ and $Normal : 1.02 + [ClusterDrought] * -2.3$ from Table 3.6 illustrate that the performance of rainfall as *Drought* or *Normal* can be predicted with Cluster_Drought as a parameter. The value -0.82 indicates the average cluster membership value for an instance to be predicted as *Drought.* The value 1.02 indicates the average cluster membership value for an instance to be predicted as *Normal* rainfall. The value of slope in *Drought* equation means that, for every increase of 1 in cluster membership value of Cluster_Drought, there is a 3.67 (from Table 3.6) times increase in chance of that instance being predicted as a *Drought.*

Table 3.6: Simple logistic regression equations for prediction of ISMR.

Class	**Equations**
Drought	-0.82 + [Cluster_Drought] * 3.67
Deficit	-0.99 + [Cluster_Deficit] * 3.67
Normal	1.02 + [Cluster_Drought] * -2.3
Excess	-1.01 + [Cluster_Excess_Rainfall] * 3.52
Flood	-1 + [Cluster_Flood] * 3.44

Similarly, the slope in *Normal* equation indicates that for every increase of 1 in cluster membership value of Cluster_Drought, there is a 2.3 (from Table 3.6) times decrease in chance of that instance being predicted as a *Normal* rainfall. To

Figure 3.3: Plot of *Drought* and *Normal* rainfall model, of regression equation given in Table 3.6 for ISMR prediction.

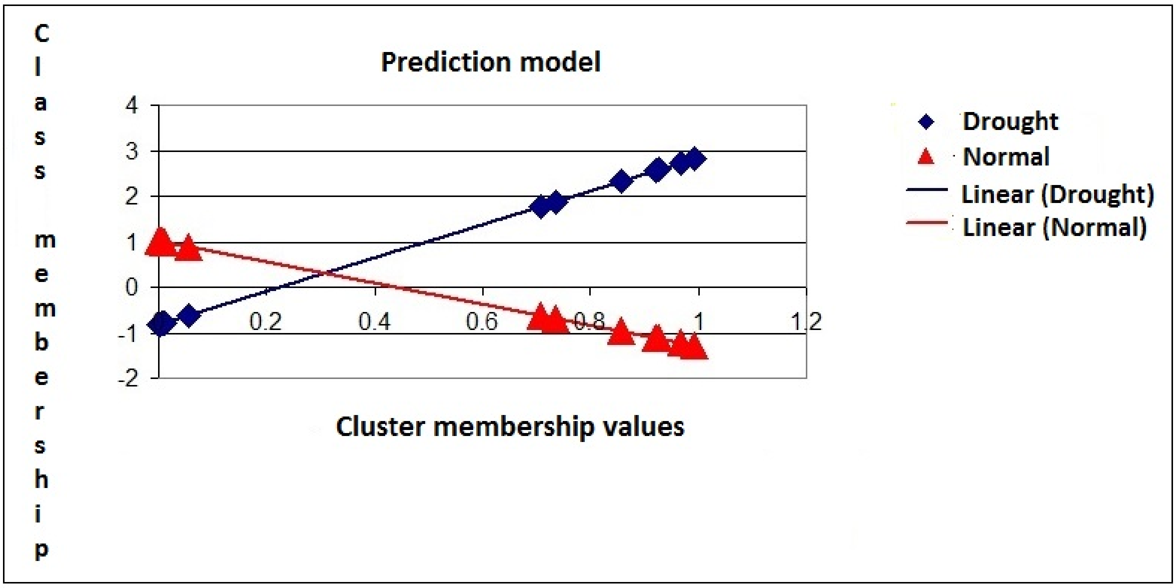


Table 3.7: Classification results using simple logistic regression in ISMR prediction.

Measures	10 Fold	5 Fold	70% Training set
Correctly Classified Instances in %	97.29	97.29	91.00
Incorrectly Classified Instances in %	2.70	2.70	9.00
Kappa statistic	0.96	0.96	0.86
Mean absolute error	0.13	0.12	0.12
Root mean squared error	0.18	0.19	0.18
Coverage of cases (0.95 level) in %	100.00	100.00	100.00
Total Number of Instances	37	37	11

Table 3.8: Confusion matrices of ISMR prediction results.
Legend: Dr - *Drought*, De - *Deficit*, No - *Normal*, Ex - *Excess*, Fl - *Flood*.

10 and 5 fold cross validation						**70% Training set**				
	Dr	De	No	Ex	Fl	Dr	De	No	Ex	Fl
Dr	7	0	1	0	0	1	0	1	0	0
De	0	4	0	0	0	0	2	0	0	0
No	0	0	14	0	0	0	0	5	0	0
Ex	0	0	0	6	0	0	0	0	1	0
Fl	0	0	0	0	5	0	0	0	0	1

avoid cluttering, this phenomenon is shown in Figure 3.3 where only *Drought* and *Normal* equations are plotted.

Table 3.7 shows the results and statistics ten fold and five fold cross validation along with 70% training set. The accuracy of the model is measured in, the Kappa coefficient, it is a more robust measure, than the simple percent agreement calculation, as it takes into account the agreement occurring by chance. The mean absolute error is also used to measure, how close the forecasts are to the actual outcomes. The Root Mean Squared Error (RMSE) represents the sample standard deviation of the differences between the predicted and observed values. The coverage of cases, specifies whether all classes available in the training set are tested using the test set. From the statistics, the model performance is considerably good with around 97% accuracy in ten and five fold cross validation and 91% with partition of training and test set.

In ten, five fold cross validation and 70% test set, as shown in the confusion matrix in Table 3.8, one instance of *Drought* has been incorrectly classified as *Normal* rainfall, because of the association of Cluster_Drought with both *Drought* and *Normal* rainfall, with one instance of *Drought* having similar features or membership value as that of *Normal* rainfall.

The presented model on an average exhibits good prediction performance. From the literature, IMD's operational model failed to predict the deficient monsoons of 2002 and 2004 (Rajeevan et al., 2007) and GCMs listed in Nanjundiah et al. (2013) have wrongly predicted the *Normal* monsoon of 1997 and *Excess* monsoon of 1983. By contrast, the model presented in the current study accurately predicts normal rainfall of 1997 despite the strongest El Niño event (usually associated with Droughts), high rainfall of 1983, and deficient monsoon of 2002 and 2004. Table 3.9 provides a comparison of proposed model predictions with state-of-the-art models.

Table 3.9: Comparison with state-of-the-art predictor models of ISMR.

Sl.no.	**State-of-the-art models**	**Prediction results of state of the art models**	**Prediction results of proposed model**
1	7 General circulation models listed in Nanjundiah et al. (2013)	Normal monsoon season of 1997 and the Excess monsoon season of 1983 were wrongly predicted	Correctly predicts the normal monsoon of 1997 and Excess monsoon of 1983
2	IMD operational statistical model (Rajeevan et al., 2007)	Failed to predict deficient rainfall of the year 2002	Correctly predicts the deficient rainfall of the year 2002
3	IMD new operational statistical models introduced in 2003 (Rajeevan et al., 2007)	Failed to predict deficient rainfall of the year 2004	Correctly predicts the deficient rainfall of the year 2004
4	Statistical Linear Discriminant Analysis (LDA) model (Wilks, 1995; Rajeevan et al., 2000)	For test data of 1998 to 2002, 8 parameter model gave 68% correct classification, whereas the 10 parameter model showed 78% correct classification	Model predicts rainfall of 1995 to 2005 correctly.

The proposed method accurately predicts all extreme ISMR events in the study period, except for the *Drought* of 1982. 1982 and 1997 have the strongest El Niño events (usually associated with *Droughts*). Upon further examination of the dataset, NINO3, NINO4, MSLPs, and temperatures of selected predictor values of 1982 and 1997, fall in the same ranges. However, the 1997 ISMR has been correctly predicted as *Normal*, because its predictor values match the predictor values of *Normal* rainfall in more than two instances, thereby satisfying the minimum support threshold. Therefore, 1997 ISMR is grouped under *Normal* rainfall. In the case of 1982, although the predictor values match the *Normal* rainfall predictor values, 1982 is actually a *Drought* year; such an instance occurs only once in the dataset. Hence, it does not satisfy the minimum support threshold, so it is not grouped under *Drought* but under *Normal* rainfall. In view of ISMR being a complex phenomenon, experimenting and adding some more likely predictors of ISMR and including more 1982-like instances to the dataset to satisfy the minimum support threshold for grouping under the *Drought* category, may correctly predict ISMR of 1982.

3.5 SUMMARY

This chapter presents an algorithm with the following three components: 1) feature selection, 2) dimensionality reduction, and 3) classification of features for predicting ISMR. The motivational scenarios, pertaining to feature selection and dimensionality reduction, are discussed. Techniques used, include closed itemset generation-based association rule mining for feature selection, cluster membership function for dimensionality reduction, and simple logistic function for classification. The algorithm is applied on precipitation data with 36 variables obtained from 37 years. The association rules derived for *Flood*, *Excess*, *Normal*, *Deficit*, and *Droughts*, leading to 25 selected features, are listed. Cluster membership reduces dimensionality from 25 to eight, by grouping closely related instances into one cluster. A clear-cut and generalized cluster relationship, to a class of rainfall is evident in simple logistic equations. The algorithm predicts, with an accuracy of around 97% in ten, five fold cross validation and an accuracy of 91% in 70% training set and 30% test set. A single cluster related to *Drought* and *Normal* rainfall has resulted in an incorrect classification of one instance. It can be seen that combination of predictors can practically provide good prediction accuracy. The Ensemble of data mining techniques with statistical techniques has greatly aided in the performance of the model. Performance improvement of the prediction model, through comparisons between classification algorithms, application of the prediction model to different homogeneous rainfall regions, and testing the model with increased data, may be considered for improving the proposed model. In conclusion, it can be noted that, including a proper mix of ISMR predictors, as suggested by association rules and the use of ensemble techniques can complement each other and help in accurate predictions.

CHAPTER 4

RAINFALL PREDICTION IN HOMOGENEOUS RAINFALL REGION

4.1 INTRODUCTION

As seen in Chapter 3, an ensemble approach, that makes use of predictor combinations to exploit associations resulting in a better prediction of ISMR, was presented. The success of the presented model has encouraged us to further develop models for smaller geographical units. Although all India level monsoon forecasts are useful at the national level, the same features and approaches may not work efficiently, for smaller regions. However, the sub regional scale, summer monsoon rainfall prediction is essential. This chapter presents a model for summer monsoon rainfall prediction for Homogeneous Rainfall Region (HRR).

A cluster of smaller geographical units that experience similar monsoon rainfall behavior is known as homogeneous rainfall region. Walker was the first to divide India into sub regions based on rainfall behavior (Walker, 1924), using regression technique for building prediction model for each of the Homogeneous Rainfall Regions. Since then, the prediction of HRR summer monsoon rainfall has emerged as an important research event. Currently to facilitate ISMR prediction, the country has been classified into four geographically homogeneous rainfall regions, namely; North-West India, Central India, North-East India, and South Peninsular India.

It is difficult to predict the rainfall of the sub divisions for various reasons, like; the data sources available for research community are provided in grid sizes of bigger resolutions, making it inaccurate for HRR rainfall prediction. During rainfall prediction research, importance is given to continent and country level monsoon forecasting, limiting the research on predictors for scaled down geographical units like HRR. Predictor identification for sub regional rainfall forecasting mainly depends on the local conditions as an essential factor. For example, in case of Peninsular India, predictors depend on the water bodies surrounding the landmass. Similarly for North-West India, predictors essentially depend on the Thar desert. However the availability of the data for local condition predictor exploration is scares. All these factors make it difficult to predict the rainfall in HRR.

Predictors of all India Summer monsoon rainfall & ENSO indicators have been used for HRR rainfall prediction thus far. A number of important studies have been conducted and have employed; (a) multivariate Principal Component Regression (PCR), (b) Neural Networks (NN), (c) Linear Discriminant Analysis (LDA) and so on. The models based on these techniques have experienced large errors in their 1991, 1994, 1997, and 1999 rainfall predictions and have shown deteriorating predictive performance after 1988, both for North-West and Peninsular India (Rajeevan et al., 2000). It is seen from literature that the predictions of HRR rainfall have had limited success (Thapliyal, 1982). Hence, attempts for prediction in a new way is essential.

In the current work, Peninsular India is selected as a HRR, as Peninsular India has two sources of rainfall, namely; the South-West Monsoon and the North-East monsoon. The possibility of the influence of the pre North-East monsoon condition is explored for prediction of Peninsular Indian Summer Monsoon Rainfall (PISMR). This study explores the local variables of both pre South-West and pre North-East monsoon, along with, global variables for prediction of PISMR. The data is carefully selected and taken from various local and international sources, to support and suit our requirement.

4.2 CHOOSING VARIABLES TO PREDICT PISMR

To explore new variables that are possible predictors of PISMR, a statistical technique called Correlation Coefficient (CC) is used. CC calculation gives quantitative measure of the dependence between variables. Though it does not give information about causative relationship, it can give a broad picture about the relationship between variables. The set of variables considered for predictor exploration are based on some studies that are reported in the literature. The following variables having meteorological basis are considered for deciding final set of predictors. The El Niño Southern Oscillation (ENSO) effect refers to the changes in SSTs over the tropical Pacific Ocean. In view of the importance of the ENSO phenomenon with respect to its effect on climate variability in various regions of the globe, many predictors have been developed based on ENSO. Their relationship with the South-West monsoon has been studied extensively during the previous two decades. Some of the most studied predictors include, Darwin Sea Level Pressure (DSLP), NINO3.4, NINO4, and the Southern Oscillation Index (SOI). There is evidence in literature that, Peninsular Indian Summer Monsoon Rainfall prediction is not accurate using ENSO and EQUINOO predictors alone. This might be due to the fact that it is not the chief rainy season for Peninsular India. Since the summer monsoon rainfall season ranges from June to September and the winter monsoon rainfall season is from October to December, it can be seen that the end of summer monsoon rainfall and the beginning of winter monsoon rainfall is adjacent. Hence it might not be sufficient to consider pre summer monsoon months data alone, for exploring possible predictors. Instead, including pre North-East monsoon months data, in the exploration, looks to be a wise approach. It is a well known fact that the monsoon is set up by a combination of temperature gradient and the pressure system, which influences the wind systems of lower and upper atmosphere (Krishnamurthy and Kinter, 2003) (Krishnamurti and Ramanathan, 1982) (Yasunari, 1985) (Sathiyamoorthy et al., 2007). Peninsular India, being the immediate landmass bounded by the Indian Ocean, experiences depressions in pressure, variations in the wind components, variations in regional temperatures, during the months of March-April-May, which may give clues regarding the monsoon process in this region. Hence, in this study, these parameters, of the southern Indian region, to some extent the Bay of Bengal, Arabian Sea, and northern Indian Ocean above the equator, are explored to identify possible predictors of the PISMR. Thus, CCs of regional weather conditions, including Mean Sea Level Pressure (MSLP), maximum and minimum temperatures, and wind speeds in the 20°N to 5°N, 60°E to 100°E area are calculated. The CCs are calculated for the data of each local variable specified above; for each month namely March, April, May, June, July and August against PISMR values. CCs are also calculated against averages of pre South-West monsoon months; March-April-May and averages of pre North-East monsoon months June-July-August. It is seen from this study that, only temperature averages, MSLP of pre South-West monsoon months and May month values of wind speed, show correlation with PISMR. However pre North-Eest monsoon months do not show any correlation with PISMR; their CC values are seen to be approximately equal to zero. The highest CCs seen with the variable maximum temperature, minimum temperature, MSLP and wind speed are in the ranges of 0.29 to 0.31, 0.41 to 0.43, -0.065 to -0.12 and 0.24 to 0.28 respectively. Form this we understand that, the pre North-East monsoon conditions are not potential predictors for PISMR prediction. The reason for this may be that, the wind reversal takes place only towards the end of September. Thus summer monsoon conditions prevail till end of September, and till August, no significant changes are seen in the weather and thus no correlation can be found during this period.

Table 4.1 lists the chosen predictors and their labels, whereas Figure 4.1 pictorially shows the same.

4.2.1 Database

Based on the correlation experiment explained in previous subsection, a total of 27 predictors are used in the current work. The predictors are identified, considering the previous research reports and Correlation Coefficient values of MSLP, temperature and wind. The data preprocessing steps involved in preparing the

Table 4.1: Chosen predictors for predicting PISMR and their labels.

Sl.no	**Predictors**	**Label**
1	Darwin Sea Level Pressure in millibar	DSLP
2	Equatorial Wind (EQWIN)	EQWIN
3	NINO 3.4 index	NINO3.4
4	NINO 4 index	NINO4
5	Southern oscillation index	SOI
6	Maximum temperature (12.5N, 76.5E) in °C	Max1
7	Maximum temperature (12.5N, 77.5E) in °C	Max2
8	Maximum temperature (12.5N, 78.5E) in °C	Max3
9	Maximum temperature (13.5N, 78.5E) in °C	Max4
10	Maximum temperature (11.5N, 76.5E) in °C	Max5
11	Maximum temperature (10.5N,77.5E) in °C	Max6
12	Maximum temperature (9.5N, 77.5E) in °C	Max7
13	Minimum temperature (20.5N, 83.5E) in °C	Min1
14	Minimum temperature (20.5N, 84.5E) in °C	Min2
15	Minimum temperature (19.5N, 82.5E) in °C	Min3
16	Minimum temperature (19.5N, 83.5E) in °C	Min4
17	Minimum temperature (18.5N, 83.5E) in °C	Min5
18	Mean sea level pressure (5N, 77.5E) in millibar	Mslp1
19	Mean sea level pressure (5N, 80E) in millibar	Mslp2
20	Mean sea level pressure (5N, 82.5E) in millibar	Mslp3
21	Mean sea level pressure (5N, 85E) in millibar	Mslp4
22	Mean sea level pressure (5N, 92.5E) in millibar	Mslp5
23	Average wind speed (7.5N, 92.5E) in meters per second	Wnd1
24	Average wind speed (7.5N, 95E) in meters per second	Wnd2
25	Average wind speed (7.5N, 97.5E) in meters per second	Wnd3
26	Average wind speed (7.5N, 100E) in meters per second	Wnd4
27	Average wind speed (5N, 97.5E) in meters per second	Wnd5

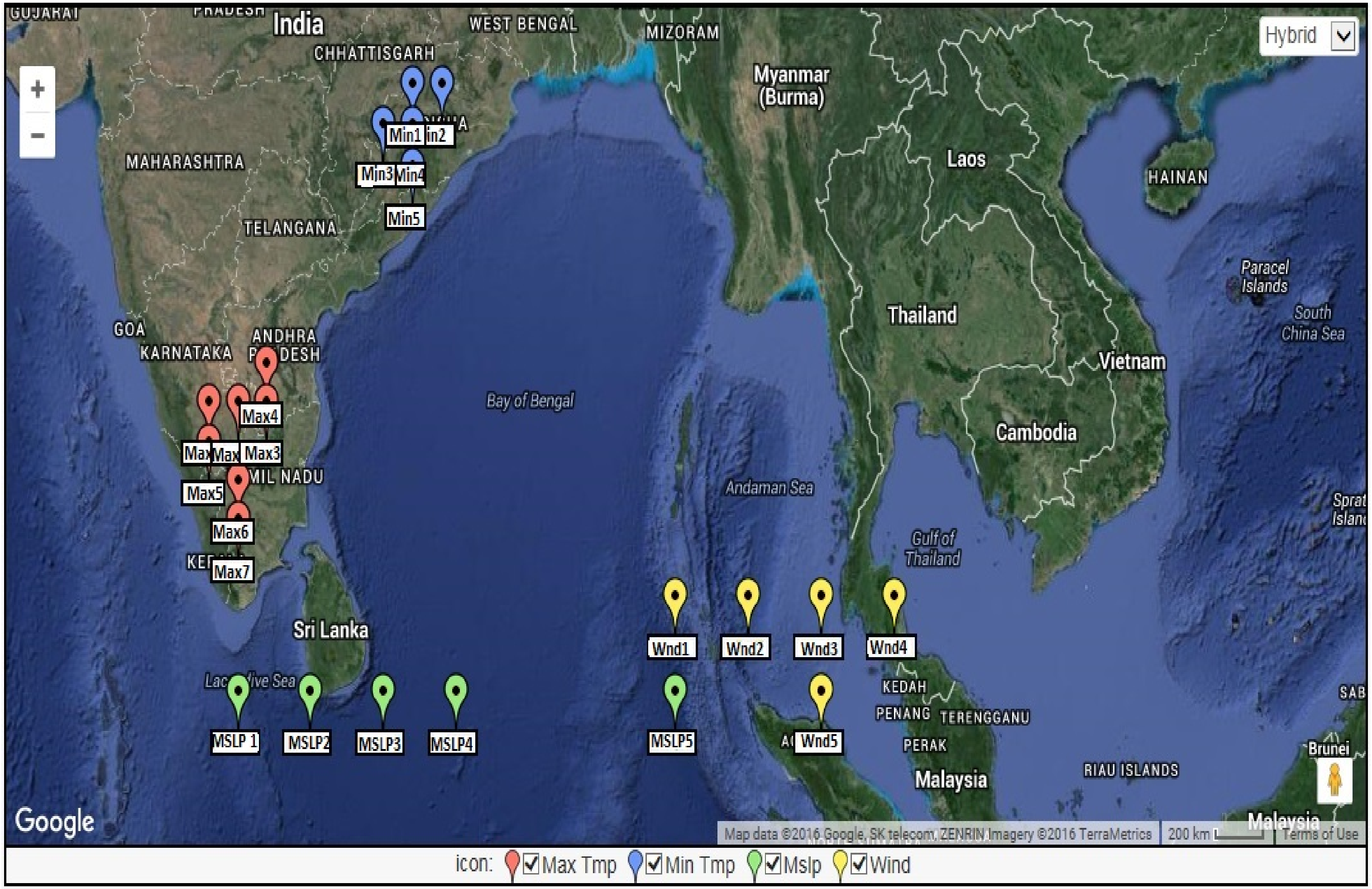


Figure 4.1: Longitude and Latitude coordinates with the highest CCs for Maximum Temperature, Minimum Temperature, Wind speed, and MSLP considered in prediction of PISMR.

data for this experiments have been addressed in the section 2.7.

1) The global variables namely DSLP, NINO4, NINO3.4 and SOI are considered as predictors.

2) March-April-May minimum (min), maximum (max) air temperature of the South Indian region (20°N - 7.5°N, 60°E - 100°E) are considered as part of local variables.

The observed data as provided by the Indian Institute of Tropical Meteorology (IITM) website has a 1° X 1° (longitude and latitude) resolution for the averaged March-April-May values of the maximum and minimum temperatures. After performing correlation analysis, $(9.5°N, 77.5°E), (10.5°N, 77.5°E), (11.5°N, 76.5°E),$ $(12.5°N, 76.5°E), (12.5°N, 77.5°E), (12.5°N, 78.5°E), (13.5°N, 78.5°E)$ are considered as latitude longitude coordinates for maximum temperature and $(18.5°N, 83.5°E),$ $(19.4°N, 83.5°E), (19.5°N, 82.5°E), (20.5°N, 84.5°E), (20.5°N, 83.5°E)$ are considered as latitude longitude coordinates for minimum temperature.

4) Mean Sea Level Pressure (MSLP) - Average MSLP of the 20°N - 5°N, 60°E - 100°E latitude longitude span is considered for predictor exploration as part of local variables.

The observed data are downloaded from the National Centers for Environmental Prediction (NCEP) and the National Center for Atmospheric Research (NCAR) and have a 2.5° X 2.5° (longitude and latitude) resolution. Top five coordinates with highest CC amongst the calculated coordinates are considered. The selected coordinates are (5°N, 77.5°E), (5°N, 80°E), (5°N, 82.5°E), (5°N, 85°E), (5°N, 92.5°E).

5) 200hPa meridional component of wind for the month of May in the South Indian region and Indian Ocean above the equator (20°N - 5°N, 60°E - 100°E) are considered as part of local variables.

The observed data are downloaded from NCEP-NCAR with a 2.5° X 2.5° (longitude and latitude) resolution. Top five coordinates with highest CC amongst the calculated coordinates are considered. The selected coordinates are $(7.5°N, 92.5°E),$ $(7.5°N, 95°E), (7.5°N, 97.5°E), (7.5°N, 100°E), (5°N, 97.5°E).$

6) Rainfall data for Peninsular India region. The data are downloaded from the IITM website.

4.3 PISMR PREDICTION

A total of 27 features are used to predict the PISMR. A data mining approach, suitable for handling the memory and CPU time constraints while processing high-dimensional data, is proposed for feature selection and dimensionality reduction. For the prediction, a statistical technique known as back propagation; Neural Network (NN) is used. The rainfall ranges for Peninsular India are given in Table 4.2. The correlation analysis is the first step for finding existence of possible relationships between variables in a bigger phenomenon. The causative aspect of a relationship within variables is necessary for finding apt predictors. Data mining approach, specifically association rule mining, is used to find the variable groups that can provide good clues about rainfall. The algorithm used for prediction of PISMR is the same as the one depicted in Chapter 3 under section 3.3. In accordance, the first four steps derive closed itemsets, subsequently, association rules are extracted for predictor selection. The clustering and classification mechanisms, giving good accuracy, in predicting PISMR are simple K-means and multilayer perceptron function respectively.

For clustering, a simple k-means algorithm is used, it partitions n observations into k clusters, where each observation belongs to the cluster with the nearest mean. This method produces exactly k different clusters of greatest possible distinction. The best number of clusters k is not known apriori, so it must be computed from the data. By applying the simple k-means clustering mechanism, 37 years of data are grouped into nine clusters, and are separated into the *Drought*, *Deficit*, *Normal*, *Excess*, and *Flood* according to behavior of rainfall.

The Multi Layer Perceptron (MLP) function can be viewed as a classifier, consisting of multiple layers of neurons; where the first layer is the input layer. the number of neurons depends in this layer, on the number of input variables, and the last layer is the output layer. The intermediate layers are known as hidden layers. With the MLP, the input undergoes non-linear transformation, whereby the input data is projected into a space and thus it becomes linearly separable. This intermediate layer is referred to as the hidden layer, which can be either one or many, depending on the application. The function of the hidden layer is to encode the input and map onto the output. It has been shown that an MLP with only one hidden layer can fairly approximate any function that connects its input with its outputs (Csáji, 2001). The number of neurons in the hidden layer is determined in accordance with Equation 4.1 (Bishop, 2002).

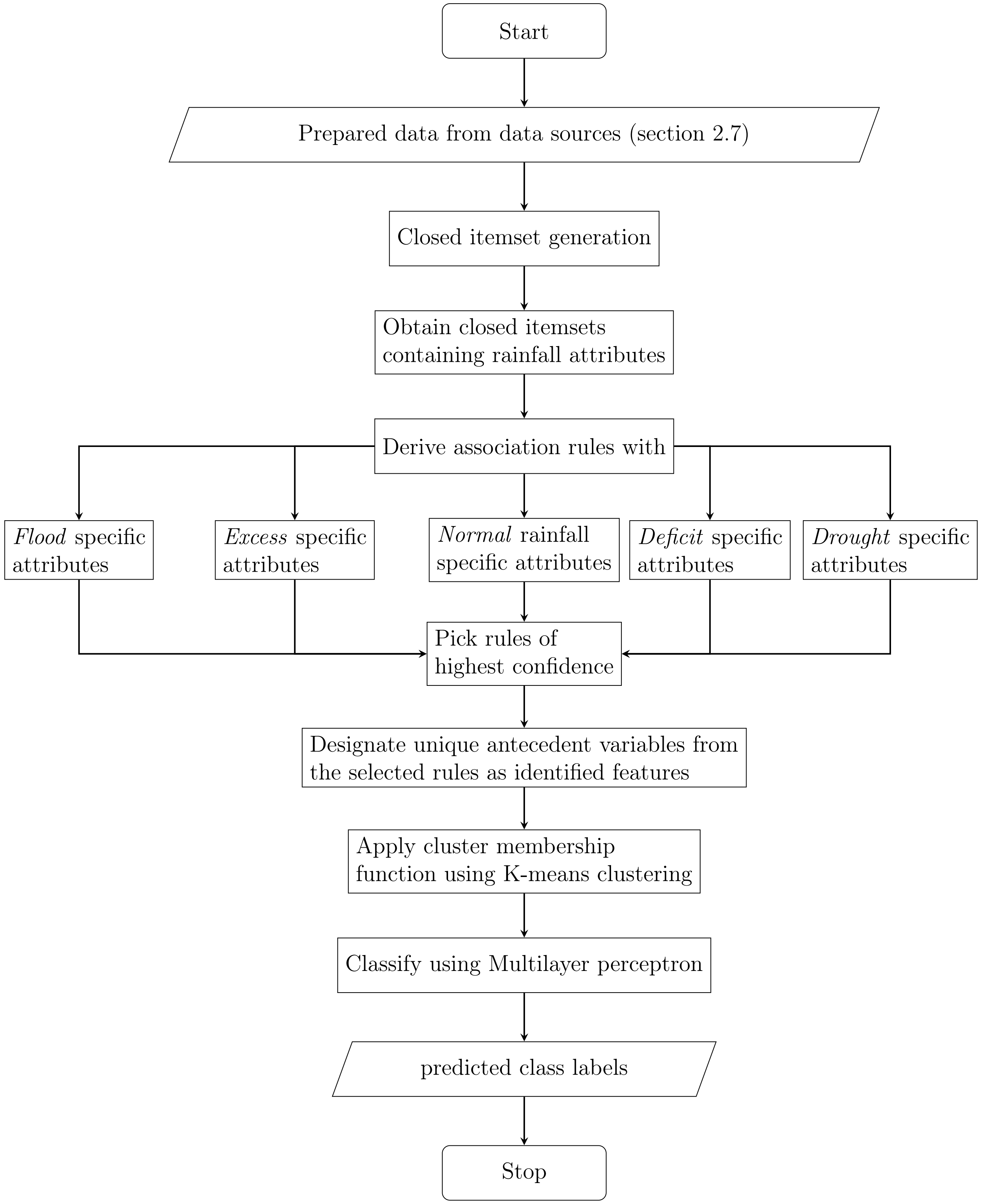
Table 4.2: Rainfall categories and their value ranges for PISMR prediction. Z represents an average rainfall for the season.

Sl.no	**Category**	**June, July, August, September rainfall average in mm**
1	Flood	$\geq$ 1914.44
2	Excess	1782.33 $\leq$ Z $<$ 1914.44
3	Normal	1518.11 $<$ Z $<$ 1782.33
4	Deficit	1385.99 $<$ Z $\leq$ 1518.11
5	Drought	$\leq$ 1385.99

$$Hidden\, neurons = \frac{(Number\ of\ input\ attributes\ +\ number\ of\ classes)}{2} = \frac{(9+5)}{2} = 7 \tag{4.1}$$

The number of neurons in the hidden layer is determined based on the input variables and the output classes. Here, the number of inputs is nine (as derived from the simple-k-means clustering) and the number of outputs, or classes, is five (*Drought*, *Deficit*, *Normal*, *Excess*, and *Flood*). As seen in Equation 4.1, the hidden layer has seven neurons. The output layer is the one that indicates the outcome of the network. The model is tested with three cases of the datasets namely ten, five fold cross validation, and 70% training and 30% testing datasets. Figure 4.2 shows a flow diagram of the system.

Figure 4.2: Flow chart of proposed model for predicting Peninsular Indian summer monsoon rainfall.


Start
Prepared data from data sources (section 2.7)
Closed itemset generation
Obtain closed itemsets containing rainfall attributes
Derive association rules with
Flood specific attributes
Excess specific attributes
Normal rainfall specific attributes
Deficit specific attributes
Drought specific attributes
Pick rules of highest confidence
Designate unique antecedent variables from the selected rules as identified features
Apply cluster membership function using K-means clustering
Classify using Multilayer perceptron
predicted class labels
Stop

4.4 RESULTS AND DISCUSSION

The association rule based feature selection resulted in 20 out of 27 variables being selected as attributes for predicting PISMR. Table 4.3 lists the association rules, their confidence levels, and the selected attributes. The rules shown in the table can be justified based on previous studies in this field. From the literature, it is known that El Niño, ENSO, EQWIN, and EQUINOO together can explain much of the PISMR variability (Gadgil, 2003). The land and ocean temperature gradients initially drive the monsoon flow in the early Indian summer monsoon season. Once the monsoon is set up, convective heating maintains the tropospheric temperature gradient and drives the monsoon flow throughout the season (Taniguchi et al., 2010). This is evident from the obtained association rules, as in all categories of rainfall, it is seen that the rules consist of land, ocean, and atmospheric attributes (local conditions). This indicates that global conditions like ENSO indices do not necessarily explain the rainfall variations in Peninsular India (Kashid and Maity, 2012). Almost all rules are dominated by conditions like the MSLP, the maximum & minimum temperatures (land condition wind), and wind (ocean conditions). When there is an MSLP drop in Northern India, the moisture bearing winds gush into low pressure regions, dumping rain on parts of India all along its path. Land conditions and global conditions together explain variations in rainfall of Peninsular India. According to Gadgil (2003), it is known that lower values of EQWIN indicate drought and an above normal DSLP with other land conditions, indicate a flood. High temperature and low atmospheric pressure regions that gradually develop over the Indian subcontinent during the pre-monsoon months of March, April and May lead to large-scale influxes of maritime air from the South Indian Ocean, into India. This justifies rules like (1) Mslp1- max ⇒ *Drought*, (2) DSLP- min Mslp4- low Mslp3- low Mslp2- low ⇒ *Flood*, and (3) Wnd3- high Wnd2- high ⇒ *Flood*. The above rules show that the MSLP values are inversely related to the PISMR. If MSLP drops in this region (5°N latitude between 77.5°E and 92.5°E longitude), the winds carrying moisture become concentrated near 5°N latitude and do not enter the other interior parts of India, which leads to good rainfall in

Peninsular India. In addition, if the MSLP is high at 5°N latitude between 77.5°E and 92.5°E longitude, the winds carrying moisture, enter the interior landmass, which leads to normal to less rainfall in Peninsular India. The relationship between Mslp1, Mslp2, Mslp3, Mslp4, and the MSLP of different parts of India can give a better understanding in this regard.

Table 4.3: List of association rules, their confidence levels, and selected attributes for predicting rainfall in peninsular India region.

Peninsular India Rainfall	**Rules**	**Confidence**	**Selected attributes**
Drought	Mslp1- max ⇒ *Drought* Eqwin- normal Wnd4- max ⇒ *Drought* Min1- low NINO4- normal Min2- low ⇒ *Drought*	0.4536 0.4536 0.4536	Mslp1, Eqwin, Wnd4, Min1, NINO4, Min2, Max5, Wnd5, Soi, DMSLP, Mslp5, Wnd3, NINO3.4, Wnd2, Wnd1, Min4, Max2, Mslp4, Mslp3, Mslp2
Deficit	Max5- min ⇒ *Deficit* Wnd5- normal Soi- normal ⇒ *Deficit* DMSLP- normal Soi- normal ⇒ *Deficit*	0.8750 0.8750 0.8750	
Normal	Mslp5- max NINO4- normal ⇒ *Normal* Wnd3- normal wnd4- normal NINO3.4- normal Wnd2- normal ⇒ *Normal* Wnd1- normal NINO3.4- normal ⇒ *Normal* Min4- normal Max5- normal Min4- normal ⇒ *Normal*	0.7857 0.7857 0.7857 0.7857	
Excess	No rules derived, Short of available data		
Flood	Wnd3 - high Wnd2 - high ⇒ *Flood* DMSLP- min Mslp4- low Mslp3- low Mslp1- low ⇒ *Flood*	0.7857 0.7857	

Association rules have also shown interesting instances of atmospheric conditions, such as the fact that a combination of low minimum temperatures, with normal Nino4, yields a drought. Most of the rules generated in the PISMR pre-

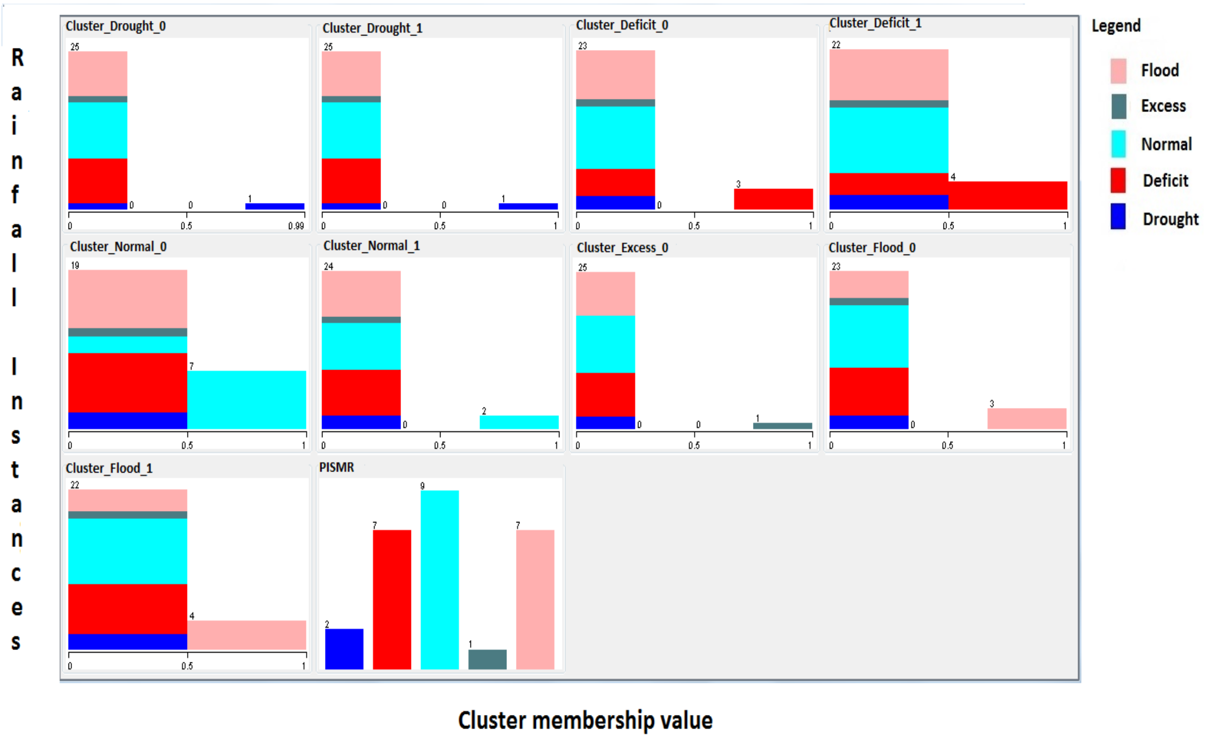


Figure 4.3: K-means cluster membership results for 26 training instances of PISMR prediction.

diction can be justified by the differential heating theory of Halley (1686). Other theories and conclusions from previous research may also be used for justifying many of the rules derived in this study.

The k-means clustering algorithm is a partition-based cluster analysis method, in which k is the initial number of cluster centers. The distances between each cluster center and each instance are calculated and then assigned the instances to their nearest cluster. This process is repeated until the criterion function is converged. Membership values are assigned to these instances, which is its probability of belonging to a particular cluster. Figure 4.3 and Figure 4.4 show the cluster information of the training and test sets, respectively. Figure 4.3 shows that, the simple-k-means clustering technique has grouped the 26 input instances of the training set into nine groups, and clearly partitioned each rainfall behavior into two clusters. Only *Excess* behavior is grouped into a single cluster. In this dataset, there are two instances of *Drought*, seven instances of *Deficit*, nine instances of *Normal*, one instance of *Excess*, and seven instances of *Flood*. k-means

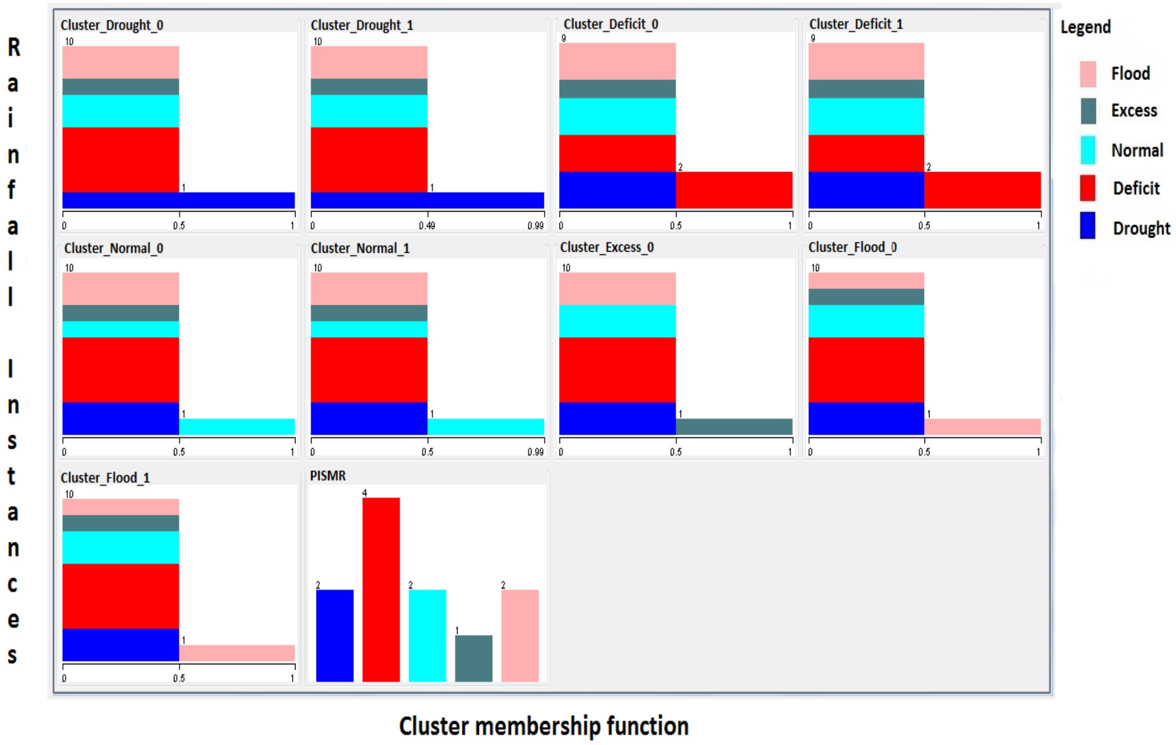


Figure 4.4: K-means cluster membership results for 11 test instances of PISMR prediction.

algorithm divides the input instances of each rainfall category into two clusters. Clusters are derived by giving each instance a probability based on the predictor values. For example, Figure 4.3 presents ten graphs in three rows and four columns. In Figure 4.3 and Figure 4.4 the graph in second column of the third row (the last one) shows the total instances of all rainfall categories and the other graphs represent clusters. The graph on first row first column, shows that one instance of the *Drought* rainfall category has a probability range of 0.76 to 0.99, and all other rainfall category instances have a 0.00 to 0.24 probability of belonging to this cluster. This cluster is called Cluster_Drought_0. Similarly, the graph in the first row second column shows that the other instance of *Drought* rainfall has a probability range of 0.75 to 1, thus belonging to the second cluster, as compared to all other instances, which have 0.00 to 0.25 probability, hence this cluster is named Cluster_Drought_1. Similarly, for each category of rainfall two clusters are formed (except the *Excess* rainfall category, due to the presence of only a single instance in it), which uniquely identify a few instances belonging

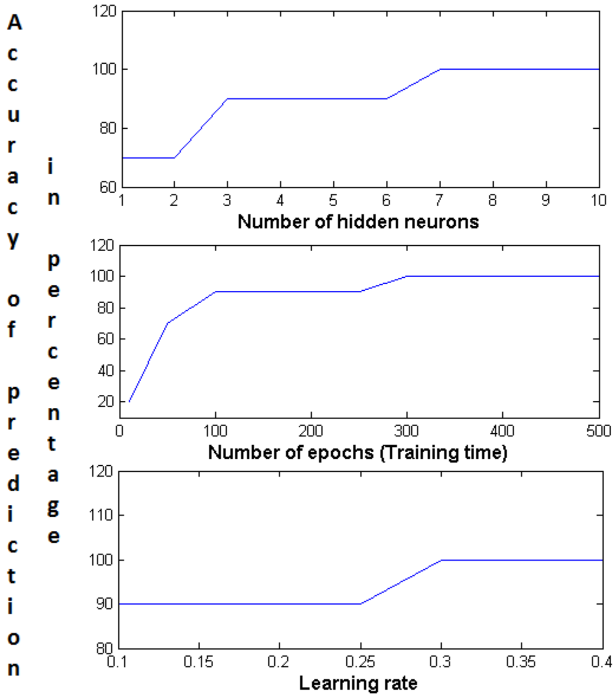


Figure 4.5: Variation in the PISMR prediction capability of the model with varying numbers of (a) hidden neurons, (b) number of epochs, and (c) learning rate.

to a particular cluster. After clustering, the dimensionality of the dataset being used is reduced from 20 selected attributes to nine clusters. These variables can now be visualized as nine in number with 26 probability values each. Similarly, in Figure 4.4, the simple-k-means clustering technique has grouped the 11 input instances of the test set into nine groups following the training set, clearly partitioning each rainfall behavior into two clusters, except for *Excess* behavior, which is grouped into a single cluster. The difference between the training and test set clusters is that, in the training set the probability ranges are narrow, so there is a strong chance that the instances will belong to a particular cluster. In the test set, however, the probability of a particular instance belonging to a particular cluster, ranges widely, thus indicating more chances for this particular instance to belong to another cluster, but with lesser probability.

As described above, MLP consists of an input layer, hidden layer(s), and an output layer. The model parameters i.e., the number of hidden neurons, the learning rate, and the number of epochs, are determined on an empirical basis. Figure 4.5 shows the relationship of the model parameters to the predic-

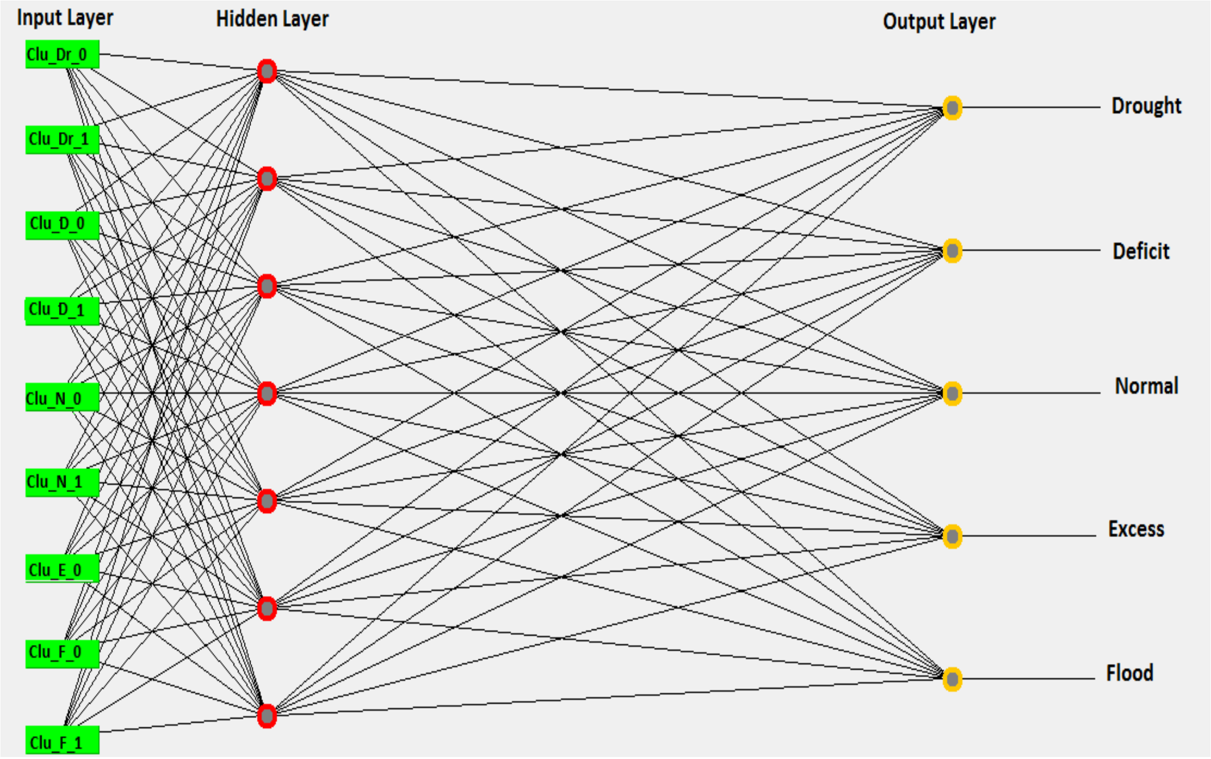


Figure 4.6: PISMR prediction model with nine input neurons, one hidden layer with seven neurons, and five output neurons.

tion accuracy. The Figure, shows that the model performs with considerably good accuracy with seven or more number of neurons in a hidden layer, a 0.3 as a learning rate, and 300 or more epochs. Figure 4.6 shows the artificial Neural Network model used in this study. In the Figure 4.6 Clu_Dr_0 represents Cluster_Drought_0, Clu_Dr_1 represents Cluster_Drought_1, Clu_D_0 represents Cluster_Deficit_0, Clu_D_1 represents Cluster_Deficit_1, Clu_N_0 represents Cluster_Normal_0, Clu_N_1 represents Cluster_Normal_1, Clu_E_0 represents Cluster_Excess_0, Clu_E_1 represents Cluster_Excess_1, Clu_F_0 represents Cluster_Flood_0, Clu_F_1 represents Cluster_Flood_1. Table 4.4 shows the PISMR prediction results and statistics for the ten fold, five fold cross validations and with a 70% training set. From these statistical results, it is seen that the model performance is very good, with around 94.5% accuracy in the ten fold and five fold cross validations and 90% accuracy with the mutually exclusive training and test sets. Good results are seen in the 70% training set case. In the cross validation cases, the results are better with a comparatively small amount of errors. In the confusion matrix shown in Table 4.5, for the ten fold cross validation, it is seen that two instance of *Excess* have been incorrectly classified as *Normal* rainfall. During the division of instances in the ten fold cross validation, two available instances of *Excess* might have been put into two different groups. Although the predictor values match those of *Normal* rainfall, they are actually *Excess* rainfall. This instance occurred once in our cross validation group, where the instance was not grouped under *Excess*, but was grouped under *Normal* rainfall, which has similar features and lower probability. A similar reason can be given for the faulty prediction of *Excess* as *Deficit* and *Normal* in the five fold cross validation, and for predicting *Excess* as *Normal* in the 70% training set case seen in Table 4.5.

Table 4.4: Classification results using the multilayer perceptron function for PISMR prediction.

Measures	**10 Fold**	**5 Fold**	**70% training set**
Correctly Classified Instances in %	94.59	94.59	90.90
Incorrectly Classified Instances in %	5.40	5.40	9.09
Kappa statistic	0.92	0.92	0.87
Mean absolute error	0.04	0.05	0.05
Root mean squared error	0.14	0.14	0.16
Coverage of cases in %	97.29	94.59	90.90
Total Number of Instances	37	37	11

On an average, the proposed model performs well during prediction. Table 4.6 presents a comparison of the proposed model predictions with those of the state-of-the-art models results. In previously reported studies, researchers have experimented with Neural Network models using eight and six parameters. These models predicted erroneous results in 1997 and 1999. However, all the models demonstrate deteriorating predictive performance after 1988 for Peninsular India (Russell and Robert, 1999). In contrast, the proposed model accurately predicts

Table 4.5: Confusion matrix of PISMR prediction results. Legend: Dr-*Drought*, De-*Deficit*, No-*Normal*, Ex-*Excess*, Fl-*Flood*

Tenfold cross validation						**Fivefold cross validation**						**70% training set**					
	Dr	De	No	Ex	Fl		Dr	De	No	Ex	Fl		Dr	De	No	Ex	Fl
Dr	4	0	0	0	0	Dr	4	0	0	0	0	Dr	1	0	0	0	0
De	0	11	0	0	0	De	0	11	0	0	0	De	0	3	0	0	0
No	0	0	11	0	0	No	0	0	11	0	0	No	0	0	3	0	0
Ex	0	0	2	0	0	Ex	0	1	1	0	0	Ex	0	0	1	0	0
Fl	0	0	0	0	9	Fl	0	0	0	0	9	Fl	0	0	0	0	3

Table 4.6: Comparison of proposed model with those of state-of-the-art models for PISMR prediction.

Sl.no	**Performance of models from literature**	**Performance of proposed model**
1	1997 and 1999 rainfall were wrongly predicted by Russell and Robert (1999) and Rajeevan et al. (2000)	Correctly predicts the *Excess* and *Drought* of 1997 and 1999 respectively
2	IMD 2003 prediction had a large difference when compared to observed rainfall - Monsoon Monograph (Volume 2)	Correctly predicts the *Deficit* of 2003

all extreme PISMR events in the study period. The model accurately predicted the *Excess* PISMR of 1997 and the *Drought* PISMR of 1999. According to the IMD's Monsoon Monograph (Volume 2), the predicted and actual values of the PISMR of 2003 were very different. The proposed model presented here, correctly predicts the *Deficit* PISMR of 2003. Compared with the other models that exhibit deteriorating predictive skills after 1988, the performance of the proposed model is better, with 94.5% accuracy in cross-validation schemes.

4.5 SUMMARY

In this chapter, an attempt for accurately predicting the ranges of Peninsular Indian Summer Monsoon is undertaken. The approach includes exploring new PISMR predictors, dimensionality reduction through clustering and use of Neural Networks for prediction. It is seen that the combination of Correlation Coefficients, with association rule mining, is used for predictor selection. In this work, 20 likely PISMR predictors are identified after association rule mining. From the identified rules, it can be seen that, land conditions along with global conditions explain the variations of PISMR. Predictors based on land conditions such as the MSLP, temperatures, and wind speeds are useful in predicting PISMR. The multilayer perceptron model, with a single hidden layer containing PISMR seven hidden neurons, yielded good accuracy in predicting. The approach is applied to precipitation data with 27 variables identified over a period of 37-years. The model predicts with 90% accuracy, when total data are partitioned into test and training sets. It yielded an accuracy of around 94% in ten and five fold cross validations.

In conclusion, it can be seen that the combination of Correlation Coefficients and association rule mining, has suggested predictors that have given good results in prediction of PISMR. An insufficient number of *Excess* instances resulted in an incorrect classification of two instance, in five fold cross validation. A relationship is found between the PISMR and the MSLP at 5°N latitude between 77.5°E and 92.5°E longitude, which is an interesting problem that requires further investigation. The PISMR is a complex phenomenon. To improve PISMR prediction skills, it is recommend that further experimentation be carried out by the inclusion of more PISMR predictors related to North-East monsoon. Consideration of more number of *Excess* rainfall instances in the datasets to satisfy the minimum support threshold for correctly grouping it under the *Excess* rainfall category can give better results.

CHAPTER 5

QUANTITATIVE LONG RANGE PREDICTION OF RAINFALL - A FUZZY LOGIC APPROACH

5.1 INTRODUCTION

In Chapter 4, a model for prediction of homogeneous rainfall region, taking a case study of Peninsular India was presented, where, a statistical technique called the Correlation Coefficient has been employed to detect variables that can become predictors of PISMR from the dump of climate data. The causative relationships that are not established by Correlation Coefficient analysis has been analyzed using association rule mining. The combination has resulted in a set of predictors that have given good accuracy. In this work K-means clustering and Neural Networks are used for dimensionality reduction and prediction; respectively. In the current chapter a model for prediction of quantitative values of rainfall is presented. The model performance is demonstrated by applying it on North Interior Karnataka Summer Monsoon Rainfall (NKSMR). In this work, Neural Networks are used to capture the changing information of climatic parameters and fuzzy logic is used to make human like heuristic decisions.

Range prediction includes prediction intervals of rainfall, a prediction interval is an estimated scope that has a start point and an end point in which the future prediction falls. For example in PISMR prediction, if a year's prediction is Excess rainfall, then the rainfall value falls within the range of Excess i.e. between 1782.33 mm $\leq$ Z $<$ 1914.44 mm. This interval is decided based on some statistical calculations. Range prediction gives a broader scope for the rainfall value to group under one of the categories. It only predicts the group into which the rainfall value of a year belongs, in other words, this merely gives general information regarding the actual rainfall value. This general range information may not be sufficient for planning many applications. One major problem encountered in range prediction is the crisp boundaries of the rainfall categories identified, based on some statistics from previous data. For example in the case of Excess ($1782.33 \leq Z < 1914.44$), a value 1914.44 mm of rainfall is Excess, whereas, just 1 mm less i.e, 1913.44 mm, falls into normal rainfall range, which is wrong according to human perception. This is true with all boundary values. This problems can be solved by involving degrees of truth, as in the case of membership function of fuzzy logic, to help better prediction.

Applications of rainfall prediction benefit more from quantitative value perdition, than just range prediction. For example, for a reservoir of medium capacity that has to get prepared (costly affair) prior to the rainy season; for water outlet in cases of excess rainfall. Instead of forecasting the range as Excess rainfall (where the lower bound of Excess range is upper bound of Normal rainfall range), it would be better to predict quantitative value. If the value is a little more than normal then the vast expense involved in preparing canals etc. can be reduced. If the rainfall is in the upper limits of Excess range then the outlet preparation is worth investing on. Similarly, almost all applications of rainfall precipitation can benefit from quantitative rainfall prediction than the range prediction.

Range prediction accuracy has consistently improved over the decades. Since consideration of Quantitative Rainfall Prediction (QRP) can further boost the accuracy, attempts are under progress for developing models for QRP. In the current work North Interior Karnataka Summer Monsoon Rainfall (NKSMR) is taken as a case study for the QRP model. Karnataka is a state in India which is located in the Peninsular part of India. Karnataka is divided into 3 homogeneous rainfall regions namely, Coastal Karnataka, South Interior Karnataka and North Interior Karnataka. In India , coastal Karnataka is known for its heavy rainfall; in contrast North Interior Karnataka is one of the driest parts of India. Majority of the rains experienced by this region is form the South-West monsoon. By and large the land of North Interior Karnataka being dry, it mainly depends on South-West monsoon for agriculture and drinking water. The economy of Karnataka is based on agriculture, hence failure of rainfall can adversely affect her economy. The state also regularly experiences large number of farmers suicide due to crop failure as a result of poor monsoon. The population of the state mainly depends on rivers for drinking water and power production. Poor monsoon also brings up the problems of scarcity of drinking water and electricity in the state. To help take precautions and arrange for alternatives, it is a necessity to predict the quantitative value of summer monsoon of Karnataka with reasonable accuracy. The contents in this chapter addresses the challenges in predicting summer monsoon rainfall of North Interior Karnataka, not in terms of ranges but in terms of crisp values.

The contributions of this work are:

1) Local condition predictors are newly suggested based on Correlation Coefficient calculations and association rule mining.

2) Use of Neuro-Fuzzy system for prediction of North Interior Karnataka Rainfall (NKSMR) by devising a mechanism to obtain effective membership functions for fuzzy "If-then" inference rules used in predictions.

3) NKSMR value prediction using the details mentioned in step 1 and 2.

This approach uses soft computing techniques namely fuzzy logic in combination with neural networks. Neural network helps the system evolve in the changing environment. Fuzzy logic helps knowledge representation, in the form of membership functions performing effective inferencing and thus taking intelligent decision.

5.2 VARIABLES CONSIDERED FOR PREDICTOR EXPLORATION TO PREDICT NKSMR

Since North Interior Karnataka is situated in the Peninsular part of India, the same meteorological conditions are assumed for considering climate variables (mean sea level pressure, maximum & minimum temperature, wind speed) (section 4.2). The predictors considered for NKSMR prediction, based on Correlation Coefficients analysis, are presented in the Table 5.1, The map in Figure 5.1 shows the correlating coordinates. The highest CCs seen with the variable maximum & minimum temperature, MSLP and wind speed are in the ranges of 0.31 to 0.34, 0.21 to 0.23 and 0.35 to 0.39 respectively. Negative correlation specify inverse relationships (when the considered variable value goes up, the rainfall value goes down). Since negative correlations also show strong relation between variables, in the current case MSLP and wind in some parts show some very negative relationship (with a CC of -0.0065 for MSLP and two negative CCs for wind speed -0.065 and -0.035) and these grid points are included for further work. The database used for the predictor exploration is as specified in subsection 4.2.1. The NKSMR data is specifically taken from IITM website, as specified in section 2.7 of chapter 2.

Table 5.1: Grid points with highest CCs of local climatic conditions considered for NKSMR prediction.

Sl.no	**Predictor**	**Coordinates of grid points**
1	Maximum temperature	$(12.5°N, 76.5°E), (9.5°N, 77.5°E),$ $(8.5°N, 76.5°E), (8.5°N, 77.5°E)$
2	Minimum temperature	$(19.5°N, 82.5°E), (19.5°N, 76.5°E), (9.5°N, 77.5°E),$ $(8.5°N, 76.5°E), (8.5°N, 78.5°E)$
3	MSLP	$(15°N, 67.5°E), (15°N, 72.5°E), (17.5°N, 70°E),$ $(17.5°N, 72.5°E), (17.5°N, 80°E)$
4	Wind speed	$(7.5°N, 60°E), (7.5°N, 92.5°E), (7.5°N, 95°E),$ $(10°N, 92.5°E), (17.5°N, 77.5°E)$

5.3 QUANTITATIVE RAINFALL PREDICTION

The algorithm presented in this section is designed for quantitative rainfall prediction. The crisp rainfall values are predicted, based on human decision making power mimicked by the fuzzy logic. Other techniques ensemble, used in this algorithm include data mining techniques with Adaptive Neural Networks (ANN).

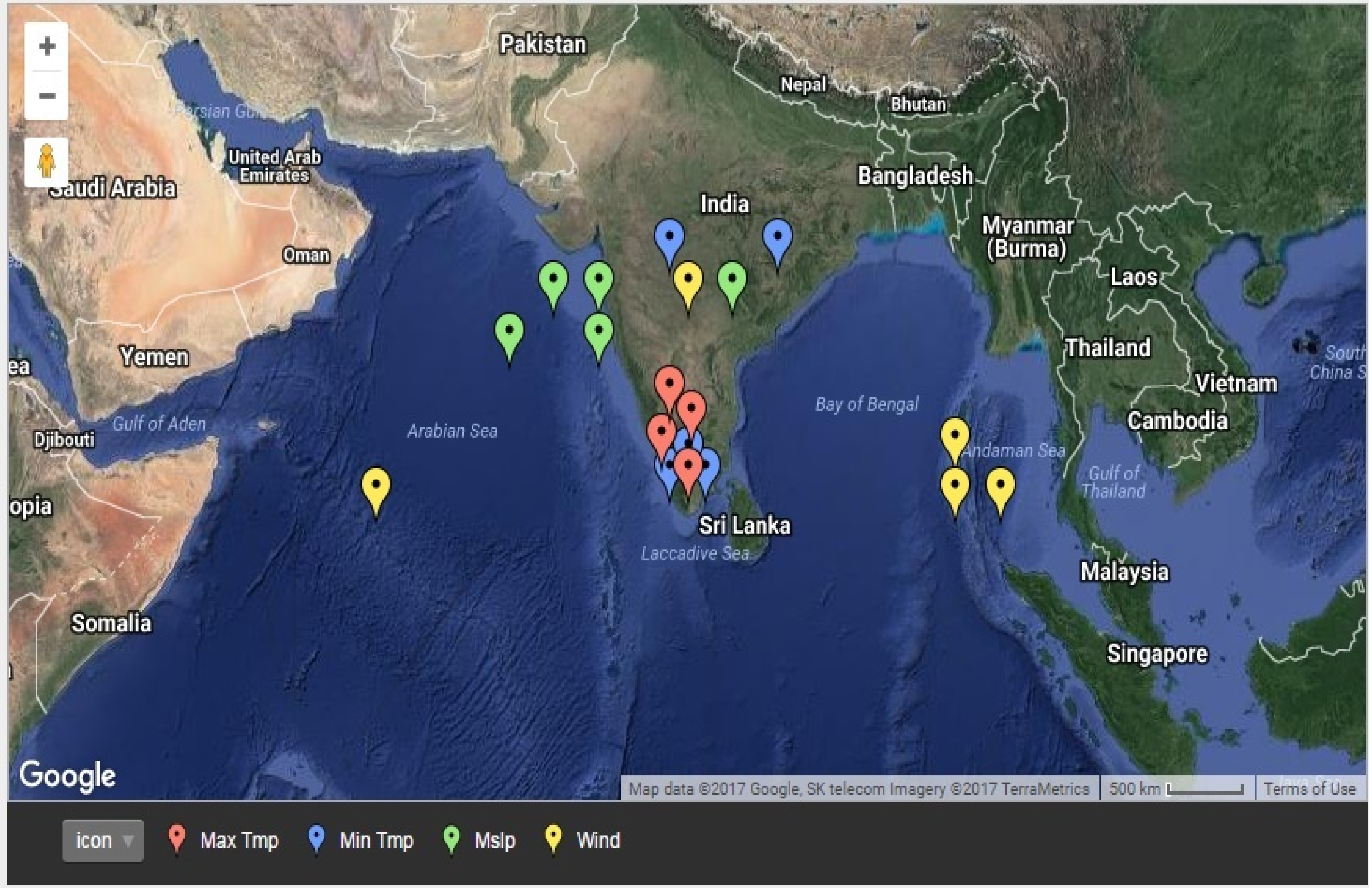


Figure 5.1: Longitude and Latitude Coordinates with the highest CCs for Maximum Temperature, Minimum Temperature, Wind speed, and MSLP considered in NKSMR prediction.

5.3.1 Algorithm steps

Algorithm 2: Neuro-Fuzzy algorithm for quantitative prediction of NKSMR

```
Input: Raw predictor data considered in this work as described in the
section 5.2, Output: Predicted quantitative values of rainfall
    1. Calculate CC, at each of the grid points for local variables MSLP, Max &
       Min temperature and wind speed values.
    2. Consider all global variables (DSLP, NINO4, NINO3.4, SOI) and highest
       CC grid points of local variables for further computation.
    3. Derive frequent itemsets using association rule mining.
    4. Derive association rules.
    5. Pick top association rules based on highest confidence measure for each bin
       of rainfall data.
    6. Use antecedent variables as predictors.
    7. Apply subtractive clustering on selected predictors to decide on the
       number and type of membership functions for a fuzzy system.
    8. Set Sigmoid membership function for Mslp5 & Wnd5 and Gaussian
       membership function for Min1, DSLP, NINO3.4 & Wnd2.
    9. Apply adaptive neural networks for fine tuning the number of "If-then"
       rules obtained from the membership functions.
   10. Use the "If-then" rules for quantitatively predicting the amount of rainfall
       for the test data.
```

The overall steps of the algorithm involved in the process of quantitative rainfall prediction is given below, these steps are explained in detail, in the following algorithm. In the algorithm, Step 1 and Step 2 give information about the predictors to be considered. For exploring the most influential predictors, variables are subjected to CC calculations against NKSMR data. Data from the grid points with highest CC are considered for further experimentation. The selected coordinates under maximum & minimum temperature, MSLP, and wind speed are given in Figure 5.1. The selected points for Max temperature are 4, whereas; for all other variables it is 5. this is because, only 4 points are associated with the highest CC in the case of maximum temperature whereas all other variables have 5 points associated with highest CC. The selected variables with highest CCs are mapped and shown in Figure 5.1. The data of these specified coordinates are further subjected to data mining.

In Step 3, 4, 5, and 6 - frequent itemset mining and association rule mining are used. This process is described in section 3.3. Association rule mining is applied on 23 variables and subsequently only 6 variables viz Mslp5, NINO3.4, Min1, DSLP, Wind2, Wind5 are chosen as predictors.

Step 7, 8, 9 and 10 - Specify the use of fuzzy logic for quantifying prediction of rainfall. Fuzzy logic mimics the human ability to interpret and imprecise an incomplete sensory (Orozco-del Castillo et al., 2011). Fuzzy set theory provides a systematic calculus to deal with such imprecise information by using linguistic labels stipulated by membership functions (Zadeh, 1965). A membership function shows the membership degrees of a variable to a certain set (Peters, 2008). For example, a temperature t=30°C belongs to the set of hot temperature with some membership degree between 0 and 1, in this case the membership degree of t=40°C is higher than t=30°C, because 40°C is hotter than 30°C. These membership degrees are graphically represented by mapping them on to curves like triangular, gaussian etc., depending on the problem being focused on. Fuzzy inference system provides the structural knowledge, represented in the form of fuzzy "If-then" rules. Fuzzy "If-then" or fuzzy conditional statements are expressions of the form "if x is A then y is B"; where A and B are linguistic values (example good, medium, bad etc.) defined by fuzzy sets on the ranges (universe of discourse) of X and Y, respectively. The "If"-part of the rule "x is A" is called the antecedent or premises, while the "then"-part of the rule "y is B" is called the consequent or conclusion. An example of such a rule might be "If service is good then tip is average". Fuzzy "If-then" rules are employed to capture the imprecise modes of reasoning that play an essential role in the human ability to make decision in an environment of uncertainty and imprecision (Mohammadian, 2009).

Fuzzy logic is capable of using human like reasoning and does not have inbuilt learning mechanism, whereas in the real world scenario, due to the changes imposed from the external environment, there is always a need for self learning mechanism. Since, fuzzy logic does not have a mechanism to learn, it lacks the ability to deal with changes introduced by external environment. Hence in this work, Neural Network learning is combined with fuzzy inference system. This helps in constructing fuzzy "If-then" rules by using fine tuned membership functions to specify input output pairs (Jang, 1993). This combination is designed by embedding fuzzy interface into ANN. Fuzzy rules are based on human ways of reasoning in uncertain conditions and Fuzzy interfaces formulate "If-then" rules, transforming human knowledge into the rule base. Adaptive Neural Network, is a superset of all kinds of feed forward neural networks with supervised learning capability. These networks are called adaptive because their nodes output can change as and when the parameter values change.

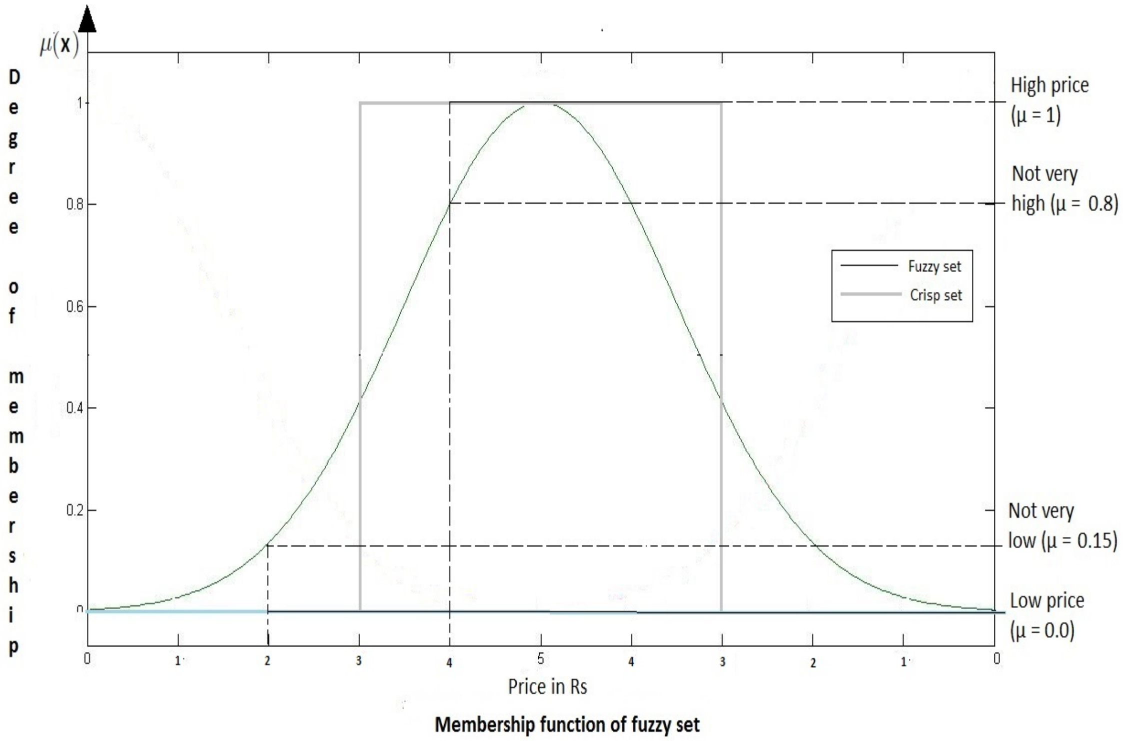
Membership functions are created for fuzzifying the data. For creating membership functions, input variables must be divided into groups based on some criteria. The number of rules is decided by human experts by visualizing the data for similarity, if the number of input variables is more, it becomes difficult to visualize and decide the similarity of data. In the current work, since 6 variables are used, it would be difficult for a human expert to decide on the similarity of data. In such cases, clustering is the best alternative. Each of the variables is subjected to Subtractive Clustering (SC). SC groups similar values of each predictor into groups where the values of predictors may be present in more than one group. Each predictor's clusters are used to represent the membership functions. Since the grouping (clusters) is made based on data points of individual predictor and not the instance of a particular year, the computations in this technique are proportional to the number of data points and not on the dimension. Hence, it seems appropriate for high dimensional data.

A Membership Function (MF) is a curve that defines how each point in the input space is mapped to a membership value (or degree of membership) between 0 and 1. A general example of a membership function is shown in Figure 5.2. Membership function of X represents fuzzy subset of X. $\mu(X)$ represents a membership function with fuzzy set X. If x is an element in X, then $\mu x(X)$ is called the membership degree of x in fuzzy set X. $\mu x(X)$ specifies the value of membership of element x in the fuzzy set X. If the value is 0 then x is not a member of the fuzzy set X; and if the value is 1 then x is fully a member of fuzzy set X. Those values between 0 and 1 represent partial belongingness of x to fuzzy set X. The membership functions are used to create "If-then" rules. The adaptive neural network takes the inputs in the form of input membership functions and the output is given as output membership function. The parameters associated with the membership function (the size and shape of the curve mapped to), change in the learning process. The fuzzy inference system models the input and output for a given set of parameters, the error measure is calculated, optimization techniques can be applied in order to reduce error. Either backward propagation or a combination of least squares estimation and back propagation (hybrid) is used for membership function parameter estimation (Jang, 1993). The error is propagated backward and the hybrid technique is used to correct the error and fine tune the membership function.

If the X-axis of Figure 5.2 represents the price of a commodity, then in the crisp graph, price Rs 2 represents low price, whereas, in fuzzy system, Rs 2 represents not very low price with a membership degree (degree of truth) of 0.15. Similarly when the price is Rs 4, the crisp graph represents high price, whereas, in fuzzy system, Rs 4 represents not very high price with membership degree (degree of truth) of 0.8. Hence, it can be seen that in crisp terms the price may be either low or high, whereas in fuzzy system intermediate pricing can be made similar to human perception.

The adaptive neural network is used to fine tune the input membership functions by learning from the input data. Adaptive neural network is a superset of all kinds of feed forward Neural Networks. This is a network consisting of nodes and directional links. In a network, part or all nodes are adaptive; which means that, the output of these nodes is dependent on the respective parameters and incoming signals, while the other nodes are fixed; that means the outputs of these nodes do

Figure 5.2: Pictorial representation of a membership function example



not change. The learning mechanism specifies the way these parameters change.

5.3.2 North Interior Karnataka Summer Monsoon Rainfall Prediction

The scope of this work is to predict quantitative value of a particular year's rainfall based on other related climate attributes. The total number of predictors used in the work are 23; historic data of 37 years (1969-2005) are considered for experimentation. Table 5.2. shows the list of predictors and their notations (labels). 23 variables have been given as input to association rule mining algorithm, the minimum support has been set in accordance to section 3.3, yielding Mslp5, Min1, NINO3.4, Wnd2, DSLP and Wnd5 as the 6 most influential predictors.

In order to identify related groups of data, inorder to create membership functions, the predictor data along with data of rainfall are further subjected to subtractive clustering. Unlike other clustering methods where computations are directly proportional to the dimension; the computations in subtractive clustering are simply proportional to the data points and are independent of the dimension

Table 5.2: Variables considered for predicting NKSMR and their short representation.

Sl.no	**Predictors**	**Representation**
1	Darwin Sea Level Pressure in millibar	DSLP
2	NINO 3.4 index	NINO3.4
3	NINO 4 index	NINO4
4	Southern oscillation index	SOI
5	Maximum temperature (12.5N, 76.5E) in °C	Max1
6	Maximum temperature (9.5N, 77.5E) in °C	Max2
7	Maximum temperature (8.5N, 76.5E) in °C	Max3
8	Maximum temperature (8.5N, 77.5E) in °C	Max4
9	Minimum temperature (19.5N, 82.5E) in °C	Min1
10	Minimum temperature (19.5N, 76.5E) in °C	Min2
11	Minimum temperature (9.5N, 77.5E) in °C	Min3
12	Minimum temperature (8.5N, 76.5E) in °C	Min4
13	Minimum temperature (8.5N, 78.5E) in °C	Min5
14	Mean sea level pressure (15N, 67.5E) in millibar	Mslp1
15	Mean sea level pressure (15N, 72.5E) in millibar	Mslp2
16	Mean sea level pressure (17.5N, 70E) in millibar	Mslp3
17	Mean sea level pressure (17.5N, 72.5E) in millibar	Mslp4
18	Mean sea level pressure (17.5N, 80E) in millibar	Mslp5
19	Average wind speed (7.5N, 60E) in meters per second	Wnd1
20	Average wind speed (7.5N, 92.5E) in meters per second	Wnd2
21	Average wind speed (7.5N, 95E) in meters per second	Wnd3
22	Average wind speed (10N, 92.5E) in meters per second	Wnd4
23	Average wind speed (17.5N, 77.5E) in meters per second	Wnd5

(Chiu, 1994). Hence, applying subtractive clustering to high dimensional data does not lead to additional CPU time consumption.

Subtractive clustering, works as follows:

If there are n data points $x_1, ..., x_n$ with m dimensions, and the parameters associated with subtractive clustering are Radius, Squash factor, Accept ratio and Reject ratio (Michal, 2013).

Radius - represents a cluster radius in which cluster center will be searched.

Squash factor - multiplied by radius, is used to discourage selection of other cluster centers near the actual one.

AcceptRatio ($\overline{\sum}$) - it is a fraction of the potential of the first center, above which, another point will be accepted as another cluster center.

RejectRatio ($\underline{\sum}$) - it is a fraction of the potential of the first center, below which, another point will be rejected as another cluster center.

Since each data point is a candidate for cluster center, a density measure at data point x_i can be defined as given in equation 5.1.

$$D_i = \sum_{j=1}^{n} exp \frac{(||x_i - x_j||)^2}{(\frac{r_a}{2})^2} \tag{5.1}$$

where, r_a is a positive constant, that defines a neighborhood. According to the equation, 5.2 a data point will have a high density value, if it has many neighboring data points. After the density measure of each data point has been calculated, the data point with the highest density measure is selected as the first cluster center D_{c1}. Further density measure of each data point x_i is revised in accordance to the equation 5.2.

$$D_i = D_i - D_{c1} exp \frac{(||x_i - x_j||)^2}{(\frac{r_b}{2})^2} \tag{5.2}$$

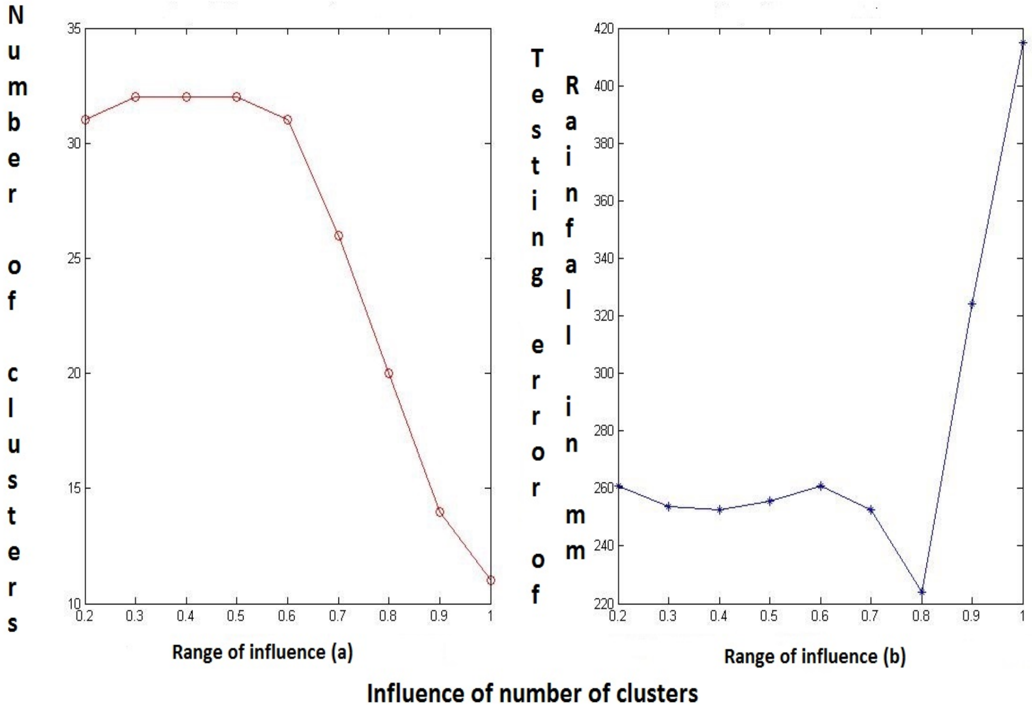
where, r_b is a positive constant, generally r_b is $1.5r_a$ (Chiu, 1994). The constant r_b is normally larger than r_a and prevents evolving closely spaced data points as cluster centers. Therefore, the data points near the first cluster center x_{c1} will have significantly reduced density measure, making the points unlikely to be selected as the next cluster center. After the revised density measure is calculated, the next cluster center is selected, similarly density measures of all data points are revised. This process continues until the stopping criteria is reached (Chiu, 1994). The stopping criteria is given in the code listing below. While selecting the k^{th} cluster center, if x_k is the k^{th} cluster center and D_k is its density value then

Algorithm 3: Stopping criteria, For deciding the number of clusters (Chiu, 1994)

```
if $D_k > \overline{\sum} D_1$
accept $x_k$ as a cluster center and continue
else if $D_k <= \overline{\sum} D_1$(Stopping criteria)
reject $x_k$ and end the clustering process.
else
let $d_{min}$ = shortest of the distances between $x_k$ and all
previously found cluster centers.
if $\frac{d_{min}}{r_b} + \frac{D_k}{D_1} >= 1$
accept $x_k$ as a cluster center
else reject $x_k$ as a cluster center.
```

In the current experiment, variation of Squash factor, Accept ratio and Reject ratio has no effect on the number of clusters. With default values for Squash factor=1.25, Accept Ratio = 0.5, Reject Ratio = 0.15 (Oweis et al., 2015) and variation of a variable called "range of influence" results in different number of clusters. The cluster radius indicates the "range of influence" of a cluster when the data space is considered as a unit hypercube. Specifying a small cluster radius yields many small clusters in the data, (resulting in many rules). Specifying a large cluster radius will usually yield a few large clusters in the data, (resulting in fewer rules) Figure 5.3 shows the plot of "range of influence" versus number of clusters (a) and "range of influence" versus testing error (b). From the Figure 5.3, it can be observed that, when the "range of influence" is 0.8, 20 clusters are created (a) and the testing error for this combination is the least compared to any other combinations (b). The cluster membership of each data point is converted into fuzzy membership value. Empirically a combination of Sigmoid and Gaussian MFs are used, to represent the current fuzzy system. This combination has been arrived at after experimenting different combinations and types of MFs. Figure 5.4. shows the plotted average error rates of different combinations of MFs for the current work. From the figure, it can be seen that when the input MSLP5 and Wnd5 are represented by Sigmoid MF and all other inputs are represented by Gaussian MF, the average error rate is comparatively less. Other MF combinations give comparatively higher errors. According to this observation, for all Mslp5 and Wnd5 Sigmoid membership functions and for others Gaussian membership functions are considered.

Figure 5.3: Deciding number of clusters for better prediction of NKSMR. (a) Plot showing the "range of influence" for minimum testing error, (b) Deciding the number of clusters based on the optimum value of "range of influence" taken from Figure 5.3 (a)

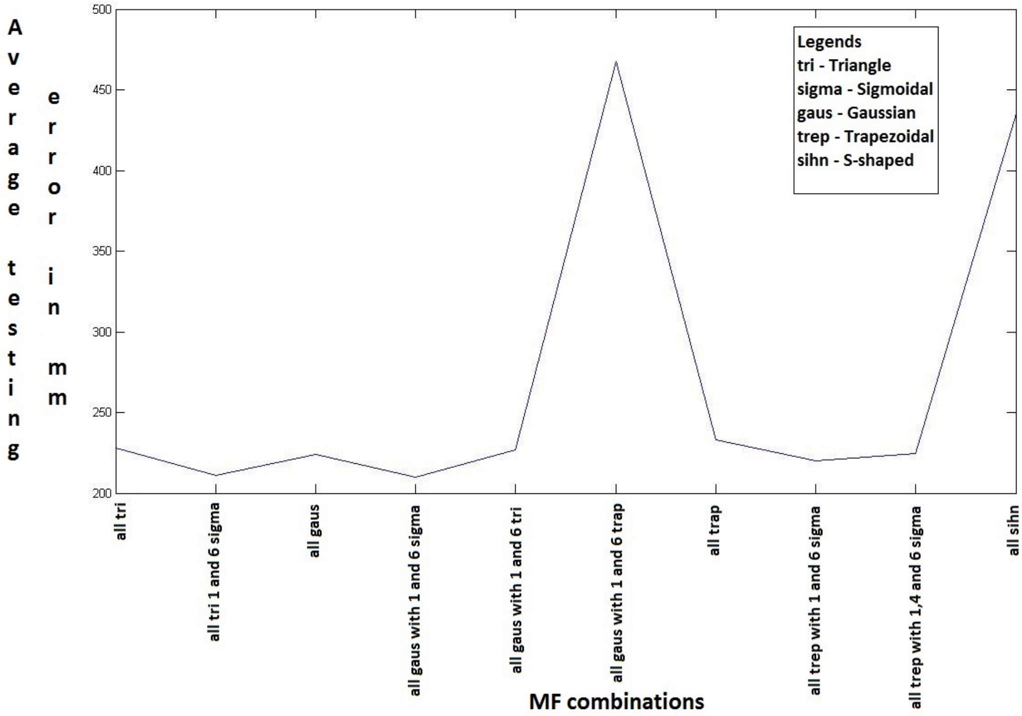


A Sigmoidal MF is defined by

$$sig(x; al, c) = \frac{1}{1 + exp[-al(x - c)]} \tag{5.3}$$

where al controls the slope at the crossover point $x = c$. The Sigmoid function given in equation 5.3 is a mapping on a vector x and depends on two parameters al

Figure 5.4: Plot of average error rate for combinations of different membership functions for NKSMR prediction.



and c. Depending on the sign of parameter al the Sigmoid function is inherently open to the right ($al < 0$) or to the left ($al > 0$) making it appropriate for representing concepts such as "very large" or "very high" etc. Figure 5.5 shows the pictorial form of Sigmoidal & Gaussian MFs.

A Gaussian MF is defined by

$$Gaussian(x; c, ro) = e^{-\frac{1}{2}(\frac{x-c}{ro})^2} \tag{5.4}$$

where c represents the center of the MFs and ro determines the spread of MFs.

Once the MFs are in place, the fuzzy "If-then" rules can be derived. Adaptive neural networks with hybrid learning method is used as a learning mechanism. The Neural network based fuzzy inference system architecture used in the current work is as shown in Figure 5.6. It has five layers. Each of the layers and its node functions are as given below. The inputs are the values of the selected variables Mslp5, Min1, NINO3.4, Wnd2, DSLP, Wnd5 . $(A_1, -A_{20}, B_1, -B_{20}, ..., F_1, -F_{20})$ are the linguistic terms (In a standard fuzzy partition, each fuzzy set corresponds to

Figure 5.5: Plot of Sigmoidal and Gaussian membership functions, (a) Sigmoidal membership function, (b) Gaussian membership function.

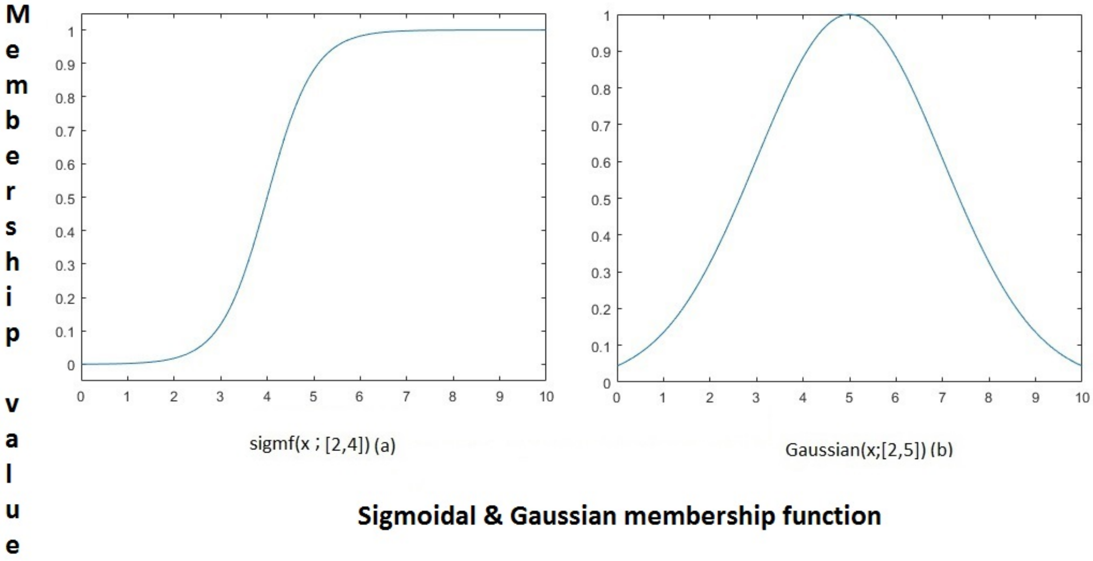
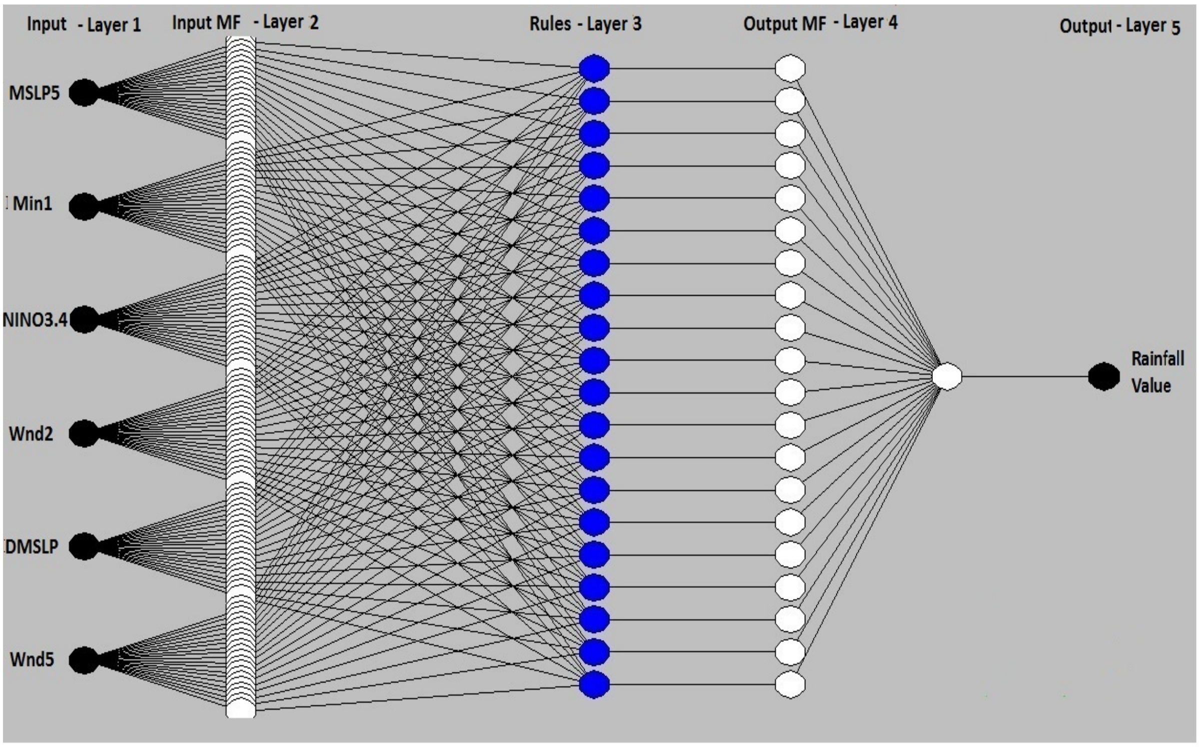


Figure 5.6: Structure of Neural Network. Used in NKSMR prediction



a linguistic concept, for instance Very low, Low, Average, High, Very High. During reasoning the variables are referred to by the linguistic terms so defined, and the fuzzy sets determine the correspondence with the numerical values). Since we have 20 clusters as a result of subtractive clustering, for simplicity, they are named $(A_1, -A_{20}, B_1, -B_{20}, C_1, -C_{20}, D_1, -D_{20}, E_1, -E_{20}, F_1, -F_{20})$ and they represent 20 clusters of six variables respectively. A fuzzy rule is written as "If situation x, then conclusion y". The situation is called rule premise or antecedent, which is defined as a combination of relations such as "$x\ is\ A$", for each component of the input vector. The conclusion part is called consequence or conclusion. Here a, b, c, d, e, g, h are considered the consequent parameters.

Suppose that the rule base contains rules like:
if Mslp5 in A_1 and Min1 in B_1 and NINO3.4 in C_1 and Wnd2 in D_1 and DSLP in E_1 and Wnd5 in F_1 then $f1 = (a_1 Mslp5 + b_1 Min1 + c_1 NINO3.4 + d_1 Wnd2 + e_1 DSLP + g_1 Wnd5 + h_1)$ and so on.

$$O_{1,i} = \mu A_i(Mslp5), i = 1, 2, .., 20 \qquad (5.5)$$

here $O_{1,i}$ is the membership function of A_i and it specifies the degree to which the given u satisfies the quantifier A_i.

Here μ represents Sigmoid or Gaussian MF.
Layer 1: Has six nodes, each node corresponds to one of the input variables Mslp5, Min1, NINO3.4, Wnd2, DSLP, Wnd5.
Layer 2: Every node in this layer is a fixed node whose output is the product of all the incoming signals

$$O_{2,i} = w_i = \mu A_i(Mslp5) \mu B_i(Min1) ...... \mu F_i(Wnd5), i = 1, 2, .., 20 \qquad (5.6)$$

each node output represents the firing strength of a rule.

Layer 3: Has fixed nodes. The i^{th} node calculates the ratio of the i^{th} rule firing strength to the sum of the firing strength of all rules.

$$O_{3,i} = \overline{w}_i = \frac{w_i}{\Sigma w_i}, i = 1, 2, .., 20 \tag{5.7}$$

Layer 4: Every node i in this layer is an adaptive node with a node function

$$O_{4,i} = \overline{w}_i f_i = \overline{w}_i (a_i Mslp5 + b_i Min1 + c_i NINO3.4 + d_i Wnd2 + e_i DSLP + g_i Wnd5 + h_i) \tag{5.8}$$

, i=1,2,..,20.

$f_i = (a_i Mslp5 + b_i Min1 + c_i NINO3.4 + d_i Wnd2 + e_i DSLP + g_i Wnd5 + h_i), i = 1, 2, .., 20$. This is the consequent output which is a linear summation made up as a product of input and consequent parameters.

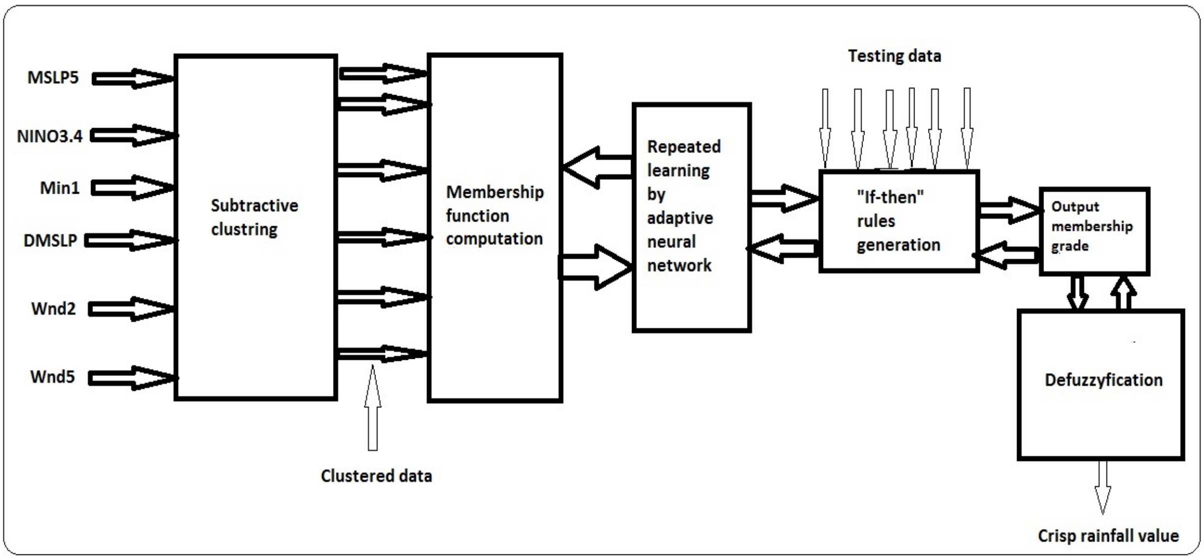
where, $\overline{w}_i$ is the normalized firing strength obtained from Layer 3 and a, b, c, d, e, g, h are the parameters of this node, these parameters are known as consequent parameters.

Layer 5: A single fixed node, which computes the overall output as the summation of all the incoming signals.

$$O_{5,i} = \sum_i \overline{w}_i f_i, i = 1, 2, .., 20 \tag{5.9}$$

The Neuro fuzzy inference system has resulted in 20 "If-then" rules. The rules obtained can be represented by a general format as given below:
If (MSLP5 is MSLP5_cluster$i$) and (Min1 is Min1_cluster$i$) and (NINO3.4 is NINO3.4_cluster$i$) and (Wnd2 is Wnd2_cluster$i$) and (DSLP is DSLP_cluster$i$) and (Wnd5 is Wnd5_cluster$i$)) then (Rainfall is Rainfall_cluster$i$) (1), where i=1,2,..20
At the end of each rule, weights are specified within brackets; this is a number between 0 and 1. Weights of those fuzzy "If-then" rules, that are likely to contribute towards successful decision, are higher. Firing strength of a rule is the product of the membership values of the premise of the rule. Every rule describes the input given and its possible output. For example; if the MSLP5, Min1, NINO3.4,

Figure 5.7: Block diagram of neuro fuzzy system, used in NKSMR prediction.

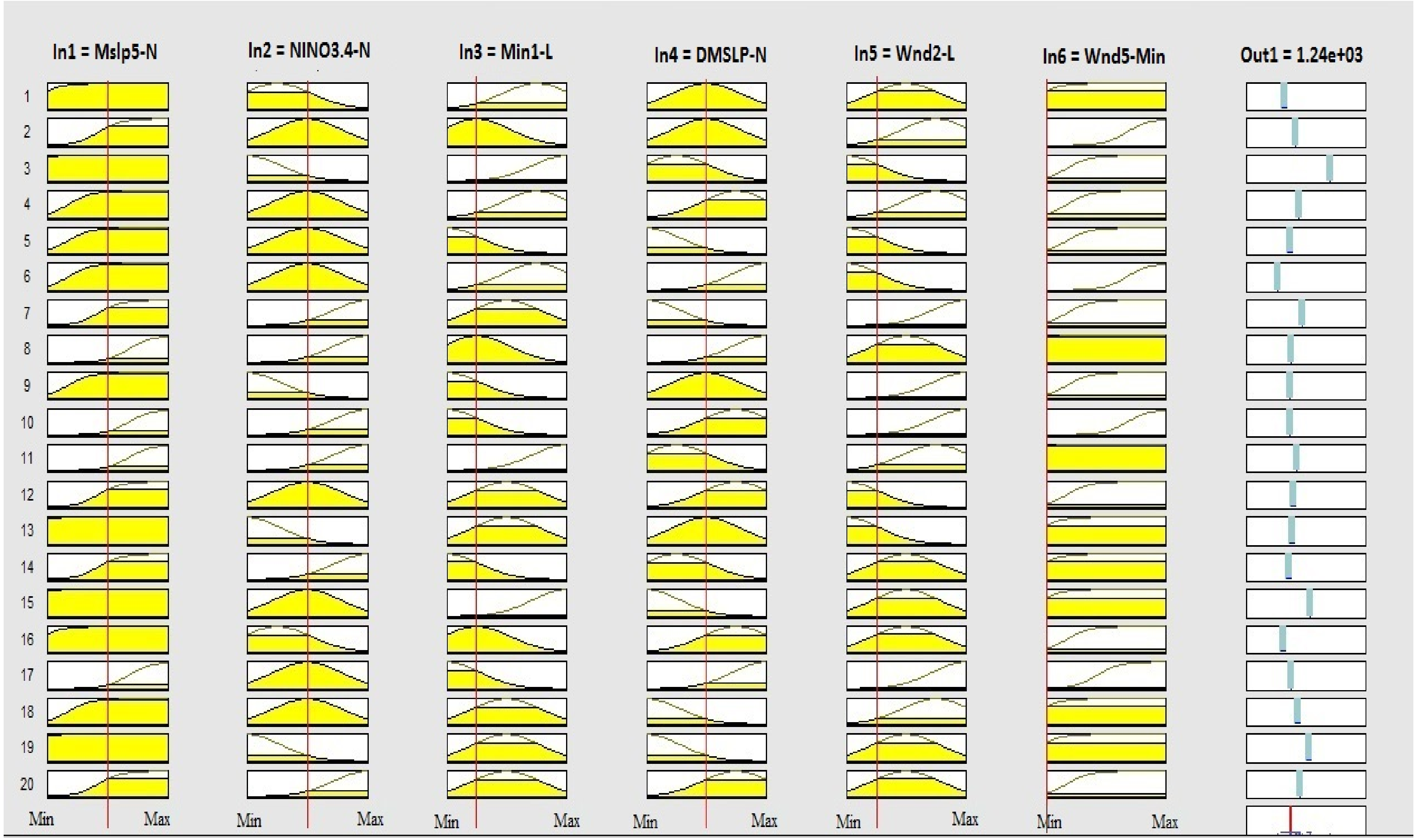


Wnd2, DSLP and Wnd5 respectively are in the first clusters of MSLP5, Min1, NINO3.4, Wnd2, DSLP and Wnd5; then the rain is expected to fall in the first cluster of rainfall variable; the weight of this rule is 1. In the current work 20 most appropriate rules are used by the model for prediction of crisp rainfall values. Further these rules are used for defuzzification. Defuzzification involves the use of consequent membership values on each of the clusters membership curve (20 curves), to find the crisp values. These crisp values are then multiplied with the firing strength of the rules respectively and summed up. The summation is then divided against the sum of the firing strengths of each rule to give the final crisp rainfall value. If the predicted rainfall in the training phase does not match with the actual rainfall value supplied through training data, then the error is propagated back to the adaptive neural network. The Neural Network makes necessary modifications in the membership functions to minimize the error. The rule generation and defuzzification cycle is repeated until the error rate is tolerable. On attaining the tolerable error rate the final set of rules is presented for testing. The block diagram in Figure 5.7. shows the sequence of steps in neuro fuzzy system, used for prediction in the current work. The training data is given as input to the model, the input values are used for computing membership functions with the help of subtractive clustering, the neural network repeatedly learns from the training data to create the rule base. The parameters associated with the membership functions will change through the learning process. The stopping criteria for the neural network learning process is the error tolerance provided by the user, in the current work, zero (aiming at minimum error) is provided as the error tolerance. Once the rule base is ready, the test data is supplied to the trained model. Depending on the test data, the most appropriate rule from the rule base gets fired, their consequents are aggregated to give weighted decision as specified in equation 5.9, thus giving a crisp rainfall value.

5.4 RESULTS AND DISCUSSION

In Figure 5.8, pictorial representation of input variables or predictors Mslp5, NINO3.4, Min1, DSLP, Wnd2 and Wnd5 that are divided into 20 clusters, forming the membership functions, may be seen. The X-axis gives the ranges of input predictors namely minimum (Min), low, normal, high and maximum (Max). The Y-axis gives the membership functions for each input predictor. Each row of plots corresponds to one rule and each column corresponds to either an input variable (the plots towards the left) or an output variable (the plot towards the right end). The first six columns of plots (the 120 yellow plots) show the membership functions inferred by the antecedent or the "If"-part of each rule. The seventh column (the 20 blue plots) shows the membership functions inferred by the consequent or the "then"-part of each rule. In the figure, over each column (except the last column), there is a line called as the index line. The index line is used for giving input to the model; the index line is positioned on the plots based on the input value of each predictor. The index line is positioned at the beginning of the columnar plots then it specifies minimum value, as in the 6^{th} column of Figure 5.8, specifying Wind5 as minimum (Wnd5-Min). If the index line is placed in the middle of the columnar plots as in case of 1^{st}, 2^{nd} and 4^{th} columns, it specifies; normal, representing

Figure 5.8: Pictorial representation of "If-then" rules showing prediction of one test case of NKSMR prediction.



normal values of Mslp5-N, NINO3.4-N and DSLP-N respectively. If the index line is positioned mid way between minimum and normal (start and the middle of the columnar plots), it specifies low value, as in the case of Wind2 (5^{th} column) of Figure 5.8; Similarly, the index line placed at the extreme end of the columnar plots, specifies maximum; while midway between normal and maximum, specifies high predictor values. The test case representation of "if Mslp5 is normal (Mslp-N), NINO3.4 is normal (NINO3.4-N), Minimum temperature is low (Min1-L), DSLP is normal(Dslp-N), Wind2 is low (Wnd2-L) and Wind5 is minimum (Wnd5-M), then the rainfall value predicted is 1.24e+03 mm", as shown in Figure 5.8. The first 20 plots of the seventh column represent the output membership function specified by the rules. In order to get the quantitative value from these rules, an aggregate weighted decision, in accordance to equation 5.9, is calculated. The twenty first box in the seventh column of plot, represent the aggregate weighted decision by the given inference system. The output crisp value is displayed as a bold vertical line on the plot of this box.

Table 5.3: Average testing error using Neuro-fuzzy system in the prediction of NKSMR, all units in mm

	85% training set	**90% training set**
case 1	209.77	233
case 2	136.5	169.3
case 3	47	51.73

Table 5.3. shows the average testing error rates with 85% and 90% training sets. In order to asses the performance of the proposed model, the results are reported for three cases.

case 1: test set containing the data of the year 2003's low rainfall condition.

case 2: training set contains one instance; a synthetic data of the year 2003, which makes the training phase a bit robust.

case 3: without the rainfall data of the year 2003 (to avoids the problems of insufficient training data).

Table 5.4: Prediction statistics using fuzzy neural network in NKSMR prediction, all units in mm

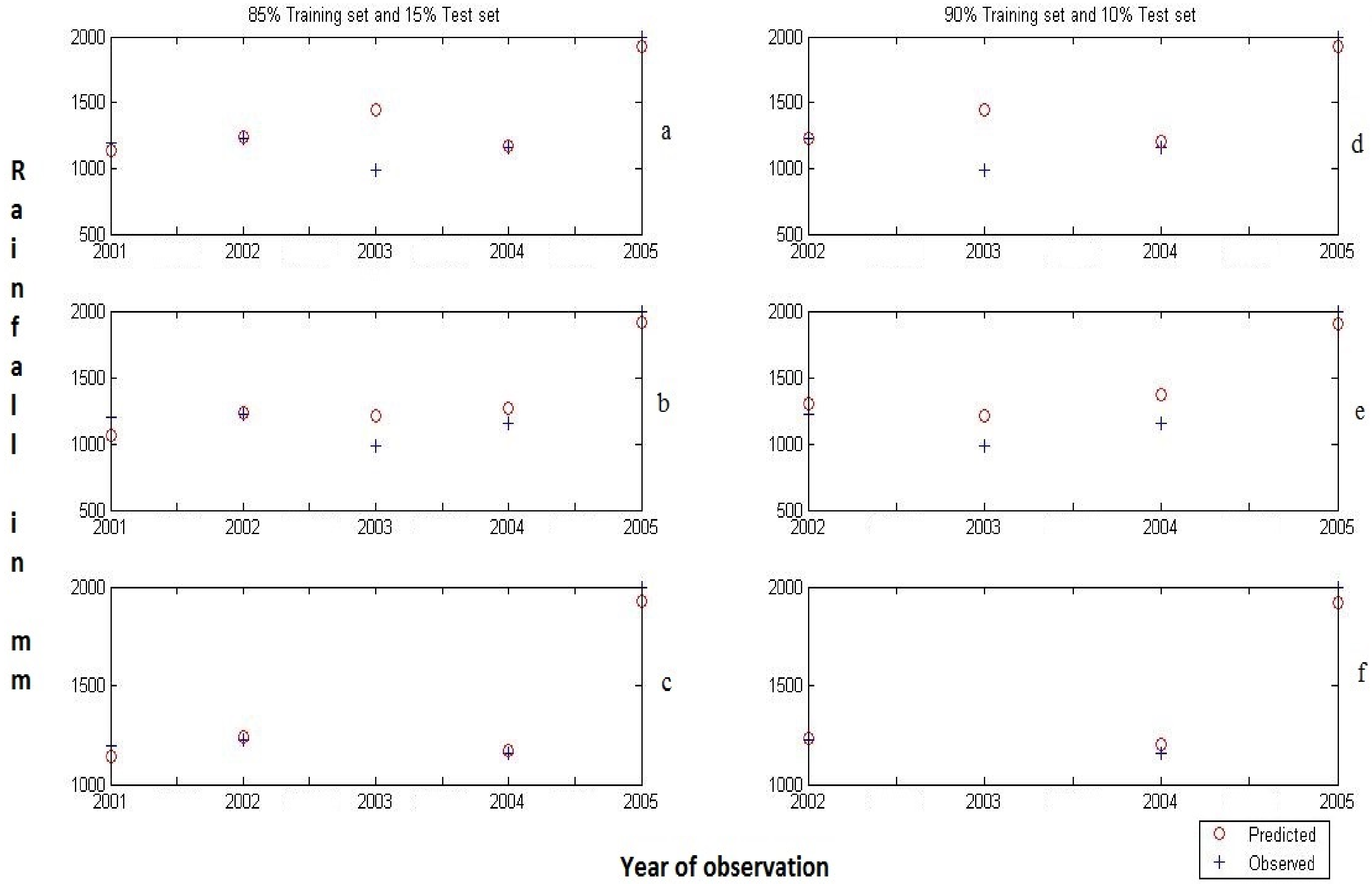
Observed data		**85% training set**						**90% training set**					
		Case 1		Case 2		Case 3		Case 1		Case 2		Case 3	
Year	Rainfall	Predicted	Error	Predicted	Error	Predicted	Error	Predicted	Error	Predicted	Error	Predicted	Error
2001	1197.75	1140	57.25	1060	137.75	1140	57.25	NA	NA	NA	NA	NA	NA
2002	1226.75	1240	-13.25	1230	-3.25	1240	-13.25	1230	-3.25	1300	-74	1230	-3.25
2003	982.25	1440	-457.75	1210	-227.75	NA	NA	1440	-457.75	1210	-227.75	NA	NA
2004	1154.25	1170	-15.75	1270	-115.75	1170	-15.75	1200	-45.75	1370	-215.75	1200	-45.75
2005	1999.5	1930	69.5	1910	89.5	1930	69.5	1920	79.5	1900	99.5	1920	79.5

These three cases are considered to demonstrate that the proposed model predicts normal and high rainfall, with good accuracy, but, due to insufficiency of training data, in the presence of low rainfall condition in test data, the model predicts, with average accuracy. However when sufficient training data of low rainfall is supplied, the proposed model can perform better. In Table 5.3 and Figure 5.9, a model with 85% training set is tested, with a total of 5 test instances in case 1 and 2 (testing with 2001, 2002, 2003, 2004 and 2005 years data), whereas; 4 test instances in case 3 (testing with 2001, 2002, 2004 and 2005 years data). A model with 90% training set is tested, with 4 test instances in case 1 and 2 (testing with 2002, 2003, 2004 and 2005 years data) whereas; 3 test instances in case 3 (testing with 2002, 2004 and 2005 years data). From the Table 5.3, it can be observed, that the model performs better in case 3 of both training conditions. This means that the model predicts normal and high rainfall conditions with good accuracy. It can also be seen in both training conditions that, case 1 has the highest error, indicating that the model performance is average, when low rainfall data is part of the test set (year 2003 low rainfall data). The reason, for this average prediction accuracy is the absence of low rainfall conditions in the training data. The model prediction accuracy, improves when synthetic data of low rainfall is added to the training set, this is observed in case 2 of both training conditions. Figure 5.9. gives the plot of the observed and predicted rainfall of all three cases noted above. The corresponding values are tabulated in Table 5.4. From the Figure 5.9. it may be observed that the normal and high rainfall conditions are very well predicted by the proposed model, whereas the prediction of low rainfall condition of 2003 is not up to the expected level of accuracy. It is also evident from Figure 5.9, that the model predicts the rainfall of the years 2001, 2002, 2004 and 2005 with appreciably good accuracy with average testing error of 47 mm with 85% training set and 51.73 mm with 90% training set respectively. This hints that, there might be some other predictors that can help capture the information related to the low rainfall conditions. In this regard inclusion of some more local weather conditions may improve the results. Another strong reason for any inferior performance by the classifier would be attributed to insufficient training data; if we observe the training data from this perspective, then the "2003 like" rainfall conditions have occurred only once, in 1980. The proposed model does not have enough number of data instances to learn the phenomenon. Figure 5.9 a, b and c show the prediction performance with 85% training and 15% testing set. Similarly 6d, 6e and 6f show the same with 90% training data and 10% data used for testing. Figure 5.9a and 5.9d show the results of case 1, 6b and 6e of case 2 and 6c and 6f of case 3. It is observed in the figures that the average prediction error of 209.77 mm and 233 mm is observed with 85% and 90% training sets for the test case containing the year 2003 (Figure 5.9a and d). Excluding the test instance of the year 2003, gives the average prediction error of 47 mm and 51.73 mm with 85% and 90% training sets respectively (Figure 5.9c and f). One can also note that the observed rainfall of 2003 of North Interior Karnataka being 982.25 mm, is predicted as 1440 mm. Interestingly the observed rainfall (North Interior Karnataka) of 1980 which contains the "2003 like" parameter values is 1441.5 mm. This indicates that the proposed model learnt the prediction scenario from 1980 data and used it for predicting the rainfall of the year 2003. This kind of error is due to the lack of data, which can be overcome by including more training instances to match with year "2003 like" predictor conditions. In order to see the effect of adding "2003 like" condition, one instance of synthetic data of 2003 is added to the training set (case 2). The results show improvement in performance, with the average error decreasing to 136.5 mm and 169.3 mm in 85% and 90% training cases respectively (Figure 5.9b and e). From the statistics it can be inferred that, exploring new predictors and also adding few more data instances to the training set that accommodates year "2003 like" conditions, can further improve the model performance.

5.4.1 Calculations involved

The fuzzification and defuzzification related to quantitative rainfall prediction of NKSMR is presented in the current section. The 85% training and 15% testing set under case 2 is presented.

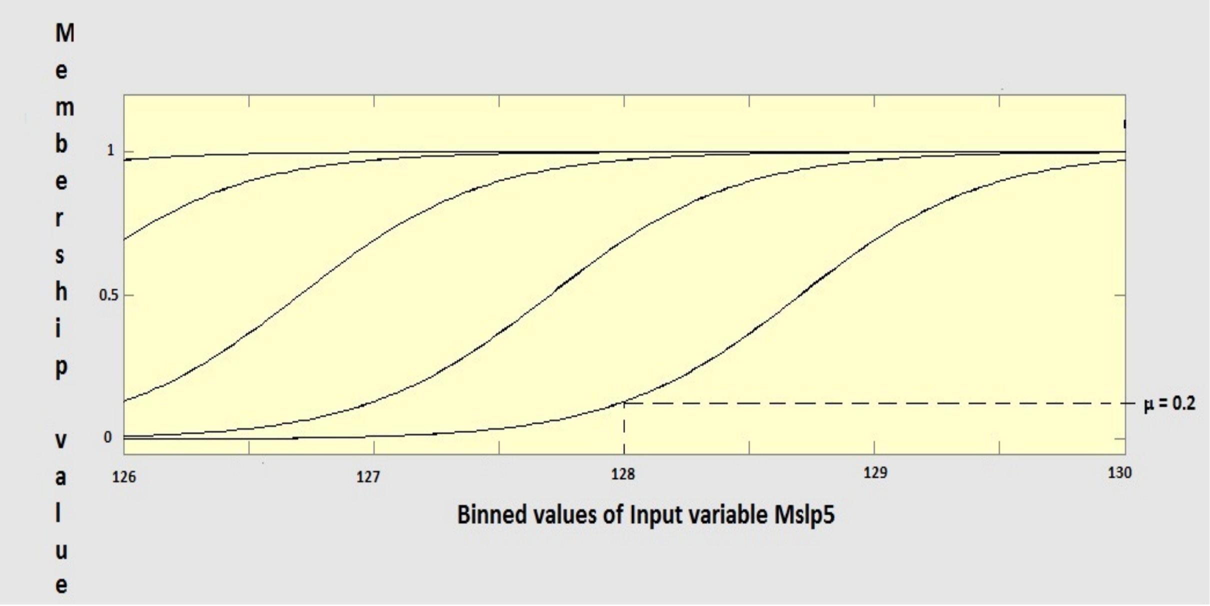
Figure 5.9: Testing cases used in NKSMR considered. a, d - Trained using data from the years 1969 - 2000 and 1969 - 2001 respectively; b, e - Trained using data from the years 1969 - 2000 and 1969 - 2001 respectively along with synthetic data for the year 2003; c, f - Trained using data from the years 1969 - 2000 and 1969 - 2001 respectively excluding year 2003 from the test set.



The observed data value of predictors Mslp5, NINO3.4, Min1, DSLP, Wnd2, Wnd5 and the corresponding rainfall value of year 2002 are 1007.02, 27.8, 23.02, 3.2, -2.96, -1.91, 1226.75 respectively. When these data are binned (only predictors) they are represented with labels as 128, 46, 27, 3, 87, 101, 1226.75. The observed rainfall value of year 2002 is 1226.75 mm.

The prediction process of rainfall for the year 2002 is discussed below. In accordance to our algorithm, the model has generated 20 rules as discussed in section 5.3.2 and 5.4. The membership functions are shown in Figure 5.8. The membership values of given data, of each predictor, in each cluster (derived from membership function), along with, the firing strength of each rule, is given in Table 5.5. The firing strength of a rule is the product of membership value of each predictor in the premise of a rule.

Figure 5.10: Membership plots of Mslp5 in NKSMR prediction.



Finding the membership value of one predictor in 20 different cluster, is an important task. In Figure 5.10, the sigmoidal curve representing membership functions of Mslp5 is shown. The 20 clusters of Mslp5 are plotted, only 5 curves are seen, other 15 curves are hidden by the 5 curves, due to overlapping. Figure 5.10, shows the membership value of the third cluster and the other clusters hidden by it. The membership value as seen from the figure is 0.2. The same is seen in Table 5.5;

in the first column third row under Mslp5. Similarly the other membership values are extracted and presented in Table 5.5. The consequent values are produced by the 20 clusters of rainfall (output) as seen in the last column of Table 5.5. These are calculated from the consequent of each rule, after learning from the training data.

Table 5.5: Membership values of all predictors in accordance to the rules generated using neuro-fuzzy system for NKSMR prediction

Mslp5	**NINO3.4**	**Min1**	**DSLP**	**wnd2**	**wnd5**	**Firing strength**	**Consequent values**
0.6766	0.6766	0.2096	1.0000	0.6766	0.6766	0.0439	1107.6
0.6766	1.0000	1.0000	1.0000	0.2096	0.0019	0.0003	1570.9
0.2096	0.2096	0.0297	0.6766	0.6766	0.2096	0.0001	3791.7
1.0000	1.0000	0.2096	0.6766	0.2096	0.2096	0.0062	1681.7
1.0000	1.0000	0.6766	0.2096	0.6766	0.2096	0.0201	1169.0
1.0000	1.0000	0.2096	0.2096	0.6766	0.0019	0.0001	0417.6
0.6766	0.2096	0.6766	0.2096	0.0297	0.2096	0.0001	2050.1
0.2096	0.2096	1.0000	0.2096	0.6766	1.0000	0.0062	1269.1
1.0000	0.2096	0.6766	1.0000	0.0297	0.2096	0.0009	1029.1
0.2096	0.2096	0.6766	0.6766	0.0297	0.0019	0.0000	1199.8
0.2096	0.2096	0.0297	0.6766	0.2096	1.0000	0.0002	1691.2
0.6766	1.0000	0.6766	0.6766	0.6766	0.2096	0.0439	1241.6
0.2096	0.2096	0.6766	1.0000	0.6766	0.6766	0.0136	1760.6
0.6766	0.2096	0.6766	0.6766	0.6766	0.6766	0.0297	1119.0
0.2096	1.0000	0.0297	0.2096	0.6766	0.6766	0.0006	2628.6
0.6766	0.6766	1.0000	0.6766	0.6766	0.2096	0.0439	0817.8
0.2096	1.0000	0.6766	0.2096	0.0297	0.0297	0.0000	1322.1
1.0000	1.0000	0.6766	0.2096	0.2096	0.6766	0.0201	1601.2
0.2096	0.2096	0.6766	0.2096	0.6766	0.6766	0.0029	2538.9
0.6766	0.2096	0.6766	0.6766	0.6766	0.2096	0.0092	1841.5

Defuzzification, for finding the crisp value, has two requirements; the firing strength of the rules and the consequent presented by the rules. The predicted crisp rainfall value is given by the following equation with reference to Equation 5.9;

$$Crisp\,value = \frac{\sum(Firing\ strength\ of\ each\ rule\ *\ consequent\ value\ of\ each\ rule)}{\sum(firing\ strength\ of\ all\ the\ rules)} \tag{5.10}$$

Let nom = 1107.6 * 0.0439 + 1570.9 * 0.0003+ 3791.7 * 0.0001 + 1681.7 * 0.0062 + 1169.0 * 0.0201 + 0417.6 * 0.0001 + 2050.1 * 0.0001 + 1269.1 * 0.0062 + 1029.1 * 0.0009 + 1199.8 * 0.0000 + 1691.2 * 0.0002 + 1241.6 * 0.0439 + 1760.6 * 0.0136 + 1119.0 * 0.0297 + 2628.6 * 0.0006 + 0817.8 * 0.0439 + 1322.1 * 0.0000 + 1601.2 * 0.0201 + 2538.9 * 0.0029 + 1841.5 * 0.0092

Let denom = 0.0439 + 0.0003 + 0.0001 + 0.0062 + 0.0201 + 0.0001 + 0.0001 + 0.0062 + 0.0009 + 0.0000 + 0.0002 + 0.0439 + 0.0136 + 0.0297 + 0.0006 + 0.0439 + 0.0000 + 0.0201 + 0.0029 + 0.0092

$$Predicted\ Rainfall\ value = \frac{nom}{denom} = 1.23e+03 \tag{5.11}$$

The crisp NKSMR value for the year 2002 obtained from the calculation is 1230.0 mm and the observed NKSMR value is 1226.75 mm. The difference in observed and predicted value of NKSMR for the year 2002 is very less, the predicted value is in concordance with the observed value, thus proving the accuracy of the presented quantitative rainfall prediction model.

5.5 SUMMARY

A model for quantitative rainfall prediction is presented in this chapter. The model is applied to North Interior Karnataka Rainfall prediction. The NKSMR predictors are explored, based on the combination of Correlation Coefficient and association rule mining techniques. Number of membership functions are decided based on the result of subtractive clustering. The learning and fine tuning of membership functions are carried out using Adaptive Neural Networks. Fuzzy logic is employed for effective decision making. Totally 23 predictors are considered in the beginning, of which, 6 variables have resulted as selected features, after applying association rule mining. Subtractive clustering has given 20 clusters, each for the 6 selected features, that are converted into membership functions for deriving fuzzy "If-then" rules. The neuro fuzzy inference system has predicted the values of rainfall; the algorithm shows good prediction skill for normal and high rainfall. The model's performance drops to average prediction accuracy, when low rainfall condition occurs in testing data. This is due to insufficient training data in case of low rainfall. Improvement is seen in model prediction accuracy, when synthetic data of low rainfall condition is included in the training set.

Using a combination of neuro fuzzy system, for quantitative rainfall prediction of NKSMR, has given promising results with minimal errors. The local climate variables have proved to be having better relationship with NKSMR. Further, new domains for exploring climate variables can be included in order to find reliable predictors that can further boost accuracy of quantitative rainfall prediction. Inclusion of a data of longer time duration may also be experimented for better accuracy.

CHAPTER 6

SUMMARY AND CONCLUSION

The thesis is organized into 6 chapters. The first chapter introduces Indian Summer Monsoon Rainfall prediction as an important research area. The second chapter critically reviews the research work done in the area of Indian Summer Monsoon Rainfall prediction, with respect to different predictors and models for forecasting ISMR. At the end of this chapter, motivation and scope for the present work is derived from the literature review. In the third chapter, an algorithm, that uses various predictors, is proposed for prediction of Indian Summer Monsoon Rainfall, using data mining and statistical approaches. In chapter 4, new local predictors, are suggested along with the use of known global conditions. A new model is proposed for prediction of Peninsular Indian Summer Monsoon Rainfall using data mining and statistical approaches. Chapter 5 proposes a model for forecasting of quantitative values of rainfall using, data mining and soft computing approaches. The model performance is verified, based on North Interior Karnataka Rainfall data. New local condition climate variables are suggested as predictors while retaining the known global condition predictors for North Interior Karnataka Rainfall prediction. Chapter 6 concludes the present work and elucidates on the directions for further research.

6.1 SUMMARY OF THE PRESENT WORK

Owing to the importance of forecasting a complex phenomenon like ISMR, different approaches are proposed in the literature for developing forecasting models. The role of predictors and techniques used in models are ever improving over the time. Majority of the models seen so far, mainly concentrate on ENSO indicators and do not consider mix of variables from different domains. The models developed for prediction of homogeneous rainfall regions are yet to improve. Since prediction of rainfall, in ranges, is not sufficient for many applications, quantitative rainfall prediction is most desirable for improving accuracy of prediction and also improving the quality of applications. Weather variables, that can be considered as potential predictors, are increasing with recent inventions adding a new domain to the climate scenario, such as; research on volcanoes, pollution, tsunamis, glaciers, global weather etc. With availability of so many variables it becomes essential to find the best variables that can become potential predictors with robust performance. From the literature it is seen that the predictors are selected based on some studies and are directly used in models. It is also observed that there are no research outcomes reported so far, where almost all know variables are used as predictors for the rainfall prediction. In this research, three prediction models are developed for ISMR, PISMR and NKSMR. In order to select most appropriate variables for each scenario from the abundantly available variables, a data mining approach known as association rule mining is employed. With large number of predictors the dimensionality of features increases. Thus, another data mining technique called "clustering", is used for overcoming the adverse effects of high dimensionality. Finally classifiers are used for prediction. Database consisting of global variables, along with local variables which include atmospheric, land and oceanic domains, are included in the model development. The main data sources are NECP-NCAR reanalysis data and observed data hosted by IITM, India. The list of techniques used in the work are: closed itemset mining, Expectation Maximization, simple logistic regression, Correlation Coefficient analysis, K-means clustering, Neural Networks, subtractive clustering and fuzzy logic. Other tools used for carrying out the work are Weka, Netbeans, Sequential Pattern Mining Framework (SPMF), NetCDF and Matlab.

The first proposed model is a hybrid model, to better the prediction of the Indian Summer Monsoon Rainfall. The proposed three-step algorithm, comprises the closed itemset generation based association rule mining for feature selection, cluster membership for dimensionality reduction and simple logistic regression function for prediction. The application predicts rainfall into Flood, Excess, Normal, Deficit and Drought, based on 36 predictors consisting of land and ocean variables. Results show 97% accuracy in cross validation schemes and 91% accuracy in training and testing scheme. Over all, the model presents good accuracy in predicting ISMR.

The second model proposed is for the prediction of Peninsular Indian Summer Monsoon Rainfall that combines data mining and statistical techniques. We select likely predictors based on association rule mining, new most influential predictors are derived from local conditions in southern India, including Mean Sea Level Pressure, wind speed, and maximum and minimum temperatures. The global condition variables include southern oscillation and Indian Ocean dipole conditions. The model predicts rainfall in five categories: Flood, Excess, Normal, Deficit and Drought. Along with closed itemset mining, cluster membership calculations, based on K-means clustering and multilayer perception function, are used to predict monsoon rainfall in Peninsular India. Using Indian Institute of Tropical Meteorology data, results show 94.5% accuracy in cross validation schemes and 90% accuracy in training and testing scheme. The model performance is approximately good in predicting PISMR.

The third model predicts the quantitative value of rainfall measured in the region of North Interior Karnataka, a region of a state in India. The proposed algorithm uses a combination of data mining and neuro fuzzy inference system for prediction. The predictors are derived from local and global climate conditions. The local condition variables are derived from the Mean Sea Level Pressure, temperature and wind speed in South India. The global variables affecting the North

Interior Karnataka Rainfall include, Darwin Sea Level Pressure, the ENSO indices and southern oscillation. The data mining technique, association rule mining, is used to study the associations among the predictors; subtractive clustering is used for predictor selection as well as membership function creation for fuzzyfication. Neuro fuzzy inference system is further used for fine tuning the "If-then" rules and crisp value prediction of the rainfall. The prediction accuracy is observed to be good considering Tropical Meteorological Department data. The rainfall of 2001 to 2005 except 2003 are quantitatively predicted with considerably good accuracy. Whereas 2003 rainfall prediction has shown improved results when synthetic data matching the same scenario was included. All studies presented in this work use data of the study period of 37 years (1969-2005). Thus, three models that give good results are developed for predicting, All India, homogeneous rainfall regions and quantitative rainfall predicting.

6.2 CONTRIBUTIONS OF THE PRESENT WORK

History of rainfall prediction in India has seen two types of models namely, the statistical models and the General Circulation Models (GCMs). It is observed that the statistical models, due to the ease of implementation, are widely used compared to the GCMs that are still not full fledged. This is due to the short falls in implementing fluid dynamic equations, to represent different components of climate. The drawback of GCM can be attributed mainly to limited understanding of the entire mechanism of monsoons by humans. Statistical models have shown better accuracy and they can further be improved; taking help from the recently developed advanced technologies. The following are the contributions made in the current work, based on the motivation derived from various shortfalls presented in the history of Indian monsoon prediction.

- Various semi permanent systems setup, during summer monsoon over India, are explored in relation to different climate variables in order to find local and global climate variables, so as to improve prediction capabilities of the statistical models.

- From literature it is seen that statistical techniques can be improved for better prediction results. In this regard the ensemble of techniques (data mining, statistical and soft computing) are explored for development of forecasting models.

- Forecasting for the scaled down geographical units does not give good results with same predictors as that of all India region; new predictors were derived form local conditions pertaining to scaled down units, based on Correlation Coefficient calculations.

- Association rule mining as a predictor selection technique is explored in order to select best predictors for better prediction accuracy.

- Different clustering techniques are explored for different models, to overcome the adverse effects of large dimensionality.

- Probabilistic classification techniques like regression and neural networks are used for rainfall range prediction.

- Quantitative value prediction is achieved by the use of Neuro-Fuzzy system; by devising a mechanism to obtain effective membership functions for fuzzy If-then inference rules.

6.3 CONCLUSIONS FROM THE PRESENT WORK

- Using a few predictors may not be sufficient for the model to predict different ranges of rainfall accurately, whereas, use of large number of known predictors for forecasting can help achieve good accuracy. Different groups of variables are seen to be supporting different ranges of rainfall, this is evident from the association rules derived.

- Many of the researches so far have used ENSO based indicators alone to predict rainfall of scaled down geographical units with average results. Deriving new variables for scaled down geographical units gives better prediction compared to using the same predictors for all the geographical units.

- Statistical models give average results when used with apt predictors and thus need post and preprocessing techniques to improve their results. Ensemble techniques complement each other and help in improving each others results, thus showing overall improvements in the forecasting performance. Thus, the combination of association rule mining, clustering and classification with learning techniques, has given good results.

- It can be concluded from the model performances that closed itemset mining, to derive frequent itemsets, can be used on huge data for extracting effective association rules, which can be used as a good technique for predictor selection.

- Large number of predictors leads to dimensionality related constraints. Clustering technique in combination with closed itemset mining can reduce feature dimensions thus helps in easy management and processing of large data.

- Definitive crisp value prediction of rainfall being a requirement, the learning mechanism of neural network combined with the fuzzy logic can successfully be used for crisp value prediction of rainfall.

- Three models presented in the current work have shown considerably good accuracy in predicting ISMR, PISMR and NKSMR. In order to improve the performance of the models, more local condition variables as predictors and data of more number of years can be used.

6.4 SCOPE FOR FUTURE WORK

- The success of the model depends on the predictors and prediction techniques. In the current research, variables related to domains such as Ocean, atmosphere, land and sea-ice are used. Variables from the effects of pollution, deforestation, volcanic eruptions, earthquakes, urbanization etc. can be included so as to incorporate global weather changes which could improve the results of forecasting.

- A real time model that takes the online streaming data of all possible variables on Earth, at any particular instant of time, to extract the relationships among data can be explored, ensuring to capture new climatic changes for prediction. This experiment may be helpful in developing real time forecasters.

- The proposed algorithms may also be applied on other phenomenon such as earthquake and tsunami forecasting in geosciences.

- General Circulation Models (GCMs) seem to be more efficient, when all components of the environment are included. But, in reality, their results are not up to the mark. Post processing the output of GCM by applying statistical techniques can help improve the results of GCM.

- Cascading GCMs with statistical models (where ever a component of the environment is difficult to implement in GCM) can possibly provide good forecasting accuracy.

- Currently the GCMs that have been developed by other countries, are used by customizing them to the Indian scenario. If a GCM is indigenously developed to suit Indian scenario, it will reduce the compromises in implementation that have to be made while including Indian conditions on other GCMs. Thus, they can be expected to give better results.

- Extensive research is required to be carried out to find predictors of scaled down geographical units. In the current study only the local conditions of South India, in terms of MSLP, temperatures and wind speed, are explored. There are many data available on variables of climate that can be processed to find possible good predictors of such small geographical areas.

- Interesting relationship is found between the new predictors, based on MSLP at 5°N latitude between 77.5°E, 92.5°E longitude and PISMR which is an interesting phenomenon. This relationship can further be investigated with the low pressure gradient formed in North India during the pre monsoon and monsoon months.

- The models presented in the current thesis can be used with other countries and other homogeneous rainfall regions of India, with suitable predictors to check their applicability.

- The techniques used to develop prediction models can be experimented with the verity available either by changing the classification / clustering techniques or by increasing the amount of data (currently 1969-2005 data is used due to limited availability) for possible improvement in the results.

- Comparative study of the current work including different predictors derived from satellite imagery.

BIBLIOGRAPHY

Abe, M., Hori, M., Yasunari, T. and Kitoh, A. (2013). "Effects of the Tibetan Plateau on the onsetof the summer monsoon in South Asia: The role of the air-sea interaction." *Journal of Geophysical Research: Atmospheres*, 118(4), 1760–1776.

Acharya, N., Chattopadhyay, S., Mohanty, U., Dash, S. and Sahoo, L. (2013). "On the bias correction of general circulation model output for Indian summer monsoon." *Meteorological Applications*, 20(3), 349–356.

Acharya, N., Chattopadhyay, S., Mohanty, U. C. and Ghosh, K. (2014). "Prediction of Indian summer monsoon rainfall: a weighted multi-model ensemble to enhance probabilistic forecast skills." *Meteorological Applications*, 21(3), 724–732.

Acharya, N., Kar, S. C., Mohanty, U. C., Kulkarni, M. A. and Dash, S. K. (2011). "Performance of GCMs for seasonal prediction over India—a case study for 2009 monsoon." *Theoretical and Applied Climatology*, 105(3), 505–520.

Aggarwal, C. C. (2015). *Data mining: the textbook.* Springer.

Aggarwal, C. C., Bhuiyan, M. A. and Hasan, M. A. (2014). "Frequent Pattern Mining Algorithms: A Survey." C. C. Aggarwal and J. Han (eds.), *Frequent Pattern Mining.* Springer International Publishing, 19–64.

Agrawal, R., Imieliaski, T. and Swami, A. (1993). "Mining Association Rules Between Sets of Items in Large Databases." *Proceeding SIGMOD 93, International Conference on Management of Data.* ACM New York, 207–216.

Ananthakrishnan, R., Acharya, U. R. and Krishnan, A. R. R. (1967). "On the criteria for declaring the onset of southwest monsoon over Kerala." Forecasting manual unit report no.iv-18.1, India Meteorological Department, Pune, India.

Angell, J., K (1981). "Comparison of variations in atmospheric quantities with sea surface temperature variations in the equatorial Pacific." *Monthly Weather Review*, 109, 230–243.

Anjaneylu, T. S. S. (1869). "On the estimates of heat and moisture budgets over the Indian monsoon trough zone." *Tellus*, 21, 64 – 74.

Armstrong, R. (2001). "Historical Soviet daily snow depth version 2 (HS-DSD). Boulder, CO: National Snow and Ice Data Center, CD-ROM. http://dx.doi.org/10.7265/N5JW8BS3."

Ashok, K., Guan, Z., Saji, N. and Yamagata, T. (2004). "Individual and combined influences of ENSO and the Indian Ocean dipole on the Indian summer monsoon." *Journal of Climate*, 17(16), 3141–3155.

Bala Subrahamanyam, D., Gupta, K. S., Ravindran, S. and Krishnan, P. (2001). "Study of Sea Breeze and Land Breeze Along the West Coast of Indian Sub-Continent Over the Latitude Range 15°N to 8°N During INDOEX." *Current Science*, 80.

Bamzai, A. S. and Shukla, J. (1999). "Relation between Eurasian snow cover, snow depth, and the Indian summer monsoon: An observational study." *Journal of Climate*, 12(10), 3117–3132.

Bansod, S., Fadnavis, S. and Ghanekar, S. (2015). "Association of the pre-monsoon thermal field over north India and the western Tibetan Plateau with summer monsoon rainfall over India." *Annales Geophysicae*, 33, 1051–1058.

Bawiskar, S., Chipade, M. and Puranik, P. (2009). "Energetics of lower tropospheric ultra-long waves: A key to intra-seasonal variability of Indian monsoon." *Journal of earth system science*, 118(2), 115–121.

Bhatla, R., Ghosh, S., Mandal, B., Mall, R. and Sharma, K. (2016). "Simulation of Indian summer monsoon onset with different parameterization convection schemes of RegCM-4.3." *Atmospheric Research*, 176, 10–18.

Bishop, R. (2002). *Design optimization of mechatronic systems, The Mechatronics Handbook, Second Edition - 2 Volume Set.* Mechatronics Handbook 2e, CRC Press.

Blanford, H. (1884). "On the connection of Himalayan snowfall and seasons of drought in India." *Proceedings of the Royal Society, London*, 37, 3 – 22.

Bosilovich, M. G., Kennedy, J., Dee, D., Allan, R. and O'Neill, A. (2013). "On the reprocessing and reanalysis of observations for climate." *Climate Science for Serving Society.* Springer, 51 – 71.

Capaldo, P., Crespi, M., Fratarcangeli, F., Nascetti, A., Francesca, P., Agugiaro, G., Poli, D. and Remondino, F. (2012). "DSM generation from optical and SAR high resolution satellite imagery: Methodology, problems and potentialities." *Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International.* IEEE, 6936–6939.

Chakravarty, S., C (2011). *Sea Surface Temperature (SST) and the Indian Summer Monsoon, Scientific and Engineering Applications Using MATLAB.* InTech.

Chakravorty, S., Chowdary, J. and Gnanaseelan, C. (2013). "Spring asymmetric mode in the tropical Indian Ocean: role of El Niño and IOD." *Climate dynamics*, 40(5-6), 1467–1481.

Charlotte, B. V., Dhanya and Mathew, B. (2012). "EQUINOO: The entity and validity of this oscillation to Indian monsoon." *International Journal of Engineering and Science*, 1, 45–54.

Chaudhari, H., Shinde, M. and Oh, J. (2010). "Understanding of anomalous Indian summer monsoon rainfall of 2002 and 1994." *Quaternary International*, 213(1), 20–32.

Chaudhuri, S. and Pal, J. (2014). "The influence of El Niño on the Indian summer monsoon rainfall anomaly: a diagnostic study of the 1982/83 and 1997/98 events." *Meteorology and Atmospheric Physics*, 124(3), 183–194.

Chiu, S. (1994). "Fuzzy Model Identification Based on Cluster Estimation." *Journal of Intelligent and Fuzzy Systems*, 2, 267–278.

Chowdary, J., Gnanaseelan, C., Vaid, B. and Salvekar, P. (2006). "Changing trends in the tropical Indian Ocean SST during La Nina years." *Geophysical research letters*, 33(18).

Clemens, S. C. and Prell, W. L. (1990). "Late Pleistocene variability of Arabian Sea summer monsoon winds and continental aridity: Eolian records from the lithogenic component of deep-sea sediments." *Paleoceanography*, 5(2), 109–145.

Csáji, B. C. (2001). "Approximation with artificial neural networks." *Faculty of Sciences, Etvs Lornd University, Hungary*, 24, 48.

Cui, X., Gao, Y., Sun, J., Guo, D., Li, S. and Johannessen, O. M. (2014). "Role of natural external forcing factors in modulating the Indian summer monsoon rainfall, the winter North Atlantic Oscillation and their relationship on interdecadal timescale." *Climate dynamics*, 43(7-8), 2283–2295.

Daellenbach, H., McNickle, D. and Dye, S. (2012). *Management Science: Decision-making Through Systems Thinking*, chapter Uncertainty. Palgrave Macmillan, 428.

Das, S. K., Deb, S. K., Kishtawal, C. and Pal, P. (2012). "Assessment of Indian summer monsoon simulation by Community Atmosphere Model (CAM3)." *Theoretical and Applied Climatology*, 109(1-2), 81–94.

Das, S. K., Deb, S. K., Kishtawal, C. and Pal, P. K. (2015). "Validation of seasonal forecast of Indian summer monsoon rainfall." *Pure and Applied Geophysics*, 172(6), 1699–1716.

Dash, S., Shekhar, M. and Singh, G. (2006). "Simulation of Indian summer monsoon circulation and rainfall using RegCM3." *Theoretical and applied climatology*, 86(1-4), 161–172.

Datta, R. (1993). "Advances in Tropical Meteorology: Meteorology and National Development." *Proceedings of the National Symposium TROPMET-93 Organised by the Indian Meteorological Society at New Delhi from March 17-19, 1993.*

David, R. E. (2008). "Issues Related to Uncertainty in the Observed Climate Record." Technical report, NOAA/National Climatic Data Center.

Dee, D., Uppala, S., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, P. et al. (2011). "The ERA-Interim reanalysis: Configuration and performance of the data assimilation system." *Quarterly Journal of the royal meteorological society*, 137, 553–597.

Deepa, R., Gnanaseelan, C., Seetaramayya, P. and Nagar, S. (2010). "On the relationship between Arabian Sea warm pool and formation of onset vortex over east-central Arabian Sea." *Meteorology and atmospheric physics*, 108, 113–125.

DelSole, T. and Shukla, J. (2002). "Linear prediction of Indian monsoon rainfall." *Journal of Climate*, 15(24), 3645–3658.

Delworth, T. L., Broccoli, A. J., Rosati, A., Stouffer, R. J., Balaji, V., Beesley, J. A., Cooke, W. F., Dixon, K. W., Dunne, J., Dunne, K. et al. (2006). "GFDL's CM2 global coupled climate models. Part I: Formulation and simulation characteristics." *Journal of Climate*, 19(5), 643–674.

Dempster, A., P, Laird, N., M and Rubin, D., B (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society*, 39, 1–38.

Dey, B. and Kumar, O. (1983). "Himalayan winter snow cover area and summer monsoon rainfall over India." *Journal of Geophysical Research: Oceans*, 88(C9), 5471–5474.

Dhanya, C. T. and Nagesh Kumar, D. (2009). "Data mining for evolution of association rules for droughts and floods in India using climate inputs." *Journal of Geophysical Research: Atmospheres*, 114(D2).

Dhar, O. and Nandargi, S. (2006). "Cherrapunji breaks the world precipitation record for one-day duration." *International Journal of Meteorology*, 31, 146–147.

Dhar, O. and Nandargi, S. (2007). "Monsoon precipitation over northeast India and Tibetan region a comparative study." *International Journal of Meteorology*, 32, 47–51.

Dickson, R. R. (1984). "Eurasian snow cover versus Indian monsoon rainfall - An extension of the Hahn-Shukla results." *Journal of Climate and Applied Meteorology*, 23(1), 171–173.

D'souza, G., Barrett, E. and Power, C. (1990). "Satellite rainfall estimation techniques using visible and infrared imagery." *Remote Sensing Reviews*, 4.

Dugam, S. (2008). "Use of interactions between NAO and MJO for the prediction of dry and wet spell in monsoon season." *e-Journal Earth Science India*, 111, 219–228.

Dwivedi, S. and Pandey, A. C. (2011). "Forecasting the Indian summer monsoon intraseasonal oscillations using genetic algorithm and neural network." *Geophysical Research Letters*, 38(15).

Ellison, J. and Milstein, J. (1995). "Improved reduced-resolution satellite imagery."

Fasullo, J. (2004). "A stratified diagnosis of the Indian monsoon-Eurasian snow cover relationship." *Journal of climate*, 17(5), 1110–1122.

Findlatter, J. (1969). "A major air current near the west Indian Ocean during the northern summer." *Quarterly Journal of the Royal Meteorological Society*, 95, 1251–1262.

Flohn, H. (1960). “Recent investiga. tions on the mechanism of the summer monsoon over southern and eastern Asia.” *In Monsoons of the World, Indian Meteorological Department*, 75–88.

Francis, P., A and Gadgil, S. (2013). “A note on new indices for the equatorial Indian Ocean oscillation.” *Journal of Earth System Science*, 122(4), 1005–1011.

Furevik, T., Bentsen, M., Drange, H., Kindem, I., Kvamstø, N. G. and Sorteberg, A. (2003). “Description and evaluation of the Bergen climate model: ARPEGE coupled with MICOM.” *Climate Dynamics*, 21(1), 27–51.

Gadgil, S. (2003). “The Indian Monsoon and its Variability.” *Annual Review Earth Planet Science*, 31, 429–467.

Gadgil, S., Rajeevan, M. and Nanjundiah, R. (2005). “Monsoon prediction - Why yet another failure.” *Current science*, 88, 1389–1401.

Gadgil, S., Vinayachandran, P., Francis, P. and Gadgil, S. (2004). “Extremes of the Indian summer monsoon rainfall, ENSO and equatorial Indian Ocean oscillation.” *Geophysical Research Letters*, 31(12).

Gago, P. and Bento, C. (1998). *Principles of Data Mining and Knowledge Discovery: Second European Symposium, PKDD’98 Nantes*. Springer Berlin Heidelberg, 19–27.

GarcÃŋa-Morales, M. B. and Dubus, L. (2007). “Forecasting precipitation for hydroelectric power management: how to exploit GCM’s seasonal ensemble forecasts.” *International Journal of Climatology*, 27(12), 1691–1705.

Goswami, P. and Srividya, P. (1996). “A Neural Network design for long range prediction of rainfall pattern.” *Current science*, 70, 447–457.

Gowariker, V., Thapliyal, V., Kulshrestha, S., Mandal, G., Senroy, R. and Sikka, D. (1991). “Power regression model for long range forecast of southwest monsoon rainfall over India.” *Mausam*, 42, 125–130.

Gowariker, V., Thapliyal, V., Sarker, R., Mandal, G. and Sikka, D. (1989). “Parametric and power regression models: New approach to long range forecasting of monsoon rainfall in India.” *Mausam*, 40, 115–122.

Grimm, A. M., Sahai, A. K. and Ropelewski, C. F. (2006). “Interdecadal variations in AGCM simulation skills.” *Journal of Climate*, 19(14), 3406–3419.

Guhathakurta, P., Rajeevan, M. and Thapliyal, V. (1999). “Long range forecasting Indian summer monsoon rainfallby a hybrid principal component neural network model.” *Meteorology and atmospheric physics*, 71(3), 255–266.

Hagedorn, R., Doblas-reyes, J., Francisco and Palmer, T. (2005). “The rationale behind the success of multi-model ensembles in seasonal forecasting–I. Basic concept.” *Tellus A*, 57(3), 219–233.

Hahn, D. G. and Shukla, J. (1976). “An apparent relationship between Eurasian snow cover and Indian monsoon rainfall.” *Journal of the Atmospheric Sciences*, 33(12), 2461–2462.

Halley, E. (1686). “An historical account of the trade winds and monsoons observable in the seas between and near the tropicks, with an attempt to assign the cause of the said winds.” *Philosophical Transactions of the Royal Society*, 16, 153–168.

Hastenrath, S. (2012). *Climate dynamics of the tropics*. Springer Science & Business Media.

Hastenrath, S. and Greischar, L. (1993). “Changing predictability of Indian monsoon rainfall anomalies?” *Proceedings of the Indian Academy of Sciences - Earth and Planetary Sciences*, 102(1), 35–47.

Hildebrandsson, H. (1897). *Quelques recherches sur les centres d’action de l’atmosphère*. Number v. 2 in Kongl. Svenska vetenskaps-akademiens handlingar, P.A. Norstedt & söner.

Holcombe, M. and Paton, R. (1997). "Information Processing in Cells and Tissues." *Proceedings of the International Workshop on Information Processing in Cells and Tissues, 1 - 4 September, Sheffield, UK.*

Hudson, D., Alves, O., Hendon, H. H. and Wang, G. (2011). "The impact of atmospheric initialisation on seasonal prediction of tropical Pacific SST." *Climate dynamics*, 36(5-6), 1155–1171.

Ihara, C., Kushnir, Y., Cane, A., Mark and Victor, H. (2006). "Indian summer monsoon rainfall and its link with ENSO and Indian Ocean climate indices." *International journal of climatology*, 27, 179–187.

Jadhav, S. K. and Munot, A. A. (2009). "Warming SST of Bay of Bengal and decrease in formation of cyclonic disturbances over the Indian region during southwest monsoon season." *Theoretical and Applied Climatology*, 96(3), 327–336.

Janakiraman, S., Ved, M., Laveti, R. N., Yadav, P. and Gadgil, S. (2011). "Prediction of the Indian summer monsoon rainfall using a state-of-the-art coupled ocean-atmosphere model." *Current Science (Bangalore)*, 100(3), 354–362.

Jang, J. S. R. (1993). "ANFIS: adaptive-network-based fuzzy inference system." *IEEE Transactions on Systems, Man, and Cybernetics*, 23(3), 665–685.

Jin, X. and Han, J. (2011). "Expectation maximization clustering." *Encyclopedia of Machine Learning*. Springer, 382–383.

Jones, J. W., Hansen, J. W., Royce, F. S. and Messina, C. D. (2000). "Potential benefits of climate forecasting to agriculture." *Agriculture, ecosystems & environment*, 82(1), 169–184.

Joseph, P., Sooraj, K. and Rajan, C. (2006). "The summer monsoon onset process over South Asia and an objective method for the date of monsoon onset over Kerala." *International journal of Climatology*, 26(13), 1871–1893.

Joseph, P. V. and Raman, P. L. (1996). “Existence of low level westerly jet stream over Peninsular India during July.” *Indian Journal of Meteorology, Hydrology and Geophysics*, 17, 407–410.

Joseph, P. V. and Sijikumar, S. (2004). “Intra-seasonal variability of the low-level jet stream of the Asian summer monsoon.” *Current science*, 17, 1449–1458.

Joseph, S., Sahai, A. and Goswami, B. (2010). “Boreal summer intraseasonal oscillations and seasonal Indian monsoon prediction in DEMETER coupled models.” *Climate dynamics*, 35(4), 651–667.

Kakade, S. and Dugam, S. (2006). “Spatial monsoon variability with respect to NAO and SO.” *Journal of earth system science*, 115(5), 601–606.

Kakade, S. and Dugam, S. (2008). “Impact of cross-equatorial flow on intraseasonal variability of Indian summer monsoon rainfall.” *Geophysical Research Letters*, 35(12).

Kakade, S. B. and Kulkarni, A. (2014). “Convective activity over heat-low region and Indian summer monsoon rainfall during contrasting phases of ESI tendency.” *Theoretical and Applied Climatology*, 115(3), 591–597.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J. et al. (1996). “The NCEP/NCAR 40-year reanalysis project.” *Bulletin of the American meteorological Society*, 437–471.

Kanamitsu, M., Ebisuzaki, W., Woollen, J., Yang, S.-K., Hnilo, J., Fiorino, M. and Potter, G. (2002). “NCEP–DOE AMIP-II reanalysis (r-2).” *Bulletin of the American Meteorological Society*, 1631–1643.

Kar, S. C., Acharya, N., Mohanty, U. and Kulkarni, M. A. (2012). “Skill of monthly rainfall forecasts over India using multi-model ensemble schemes.” *International Journal of Climatology*, 32(8), 1271–1286.

Kashid, S. S. and Maity, R. (2012). "Prediction of monthly rainfall on homogeneous monsoon regions of India based on large scale circulation patterns using Genetic Programming." *Journal of Hydrology*, 454, 26–41.

Keshavamurthy, R. and Rao, M. S. (1992). *The physics of monsoons.* Allied Publishers.

Keshavamurty, R. N. (1968). "On the maintenance of the mean zonal motion in the Indian summer monsoon." *Monthly weather review*, 96, 23 –31.

Keshvamurty, R. N. and Awade, S. T. (1970). "On the maintenance of the mean monsoon trough over north India." 98, 315 – 320.

Khandekar, M., L and Neralla, V., R (1984). "On the relationship between the sea surface temperatures in the equatorial Pacific and the Indian monsoon rainfall." *Geophysical Research Letters*, 11, 1137–1140.

Kishtawal, C., Basu, S., Patadia, F. and Thapliyal, P. (2003). "Forecasting summer rainfall over India using genetic algorithm." *Geophysical Research Letters*, 30(23).

Klimarechenzentrum, D. (1992). "The ECHAM3 atmospheric general circulation model." *Modellbetreuungsgruppe Techn. Rep*, 6.

Koteswaram, P. (1958a). "The Asian summer monsoon and the general circulation over the tropics." *Monsoons of the World*, 105–110.

Koteswaram, P. (1958b). "The easterly jet stream in the tropics." *Tellus*, 10, 43–57.

Krishnamurthy, V. and Kinter, I., JamesL. (2003). "The Indian Monsoon and its Relation to Global Climate Variability." *Global Climate*. Springer Berlin Heidelberg, 186–236.

Krishnamurti, T. and Ramanathan, Y. (1982). "Sensitivity of the monsoon onset to differential heating." *Monthly Weather Review*, 39, 1290–1306.

Krishnamurti, T., N and Kishtawal, C., M (2000). “A Pronounced Continental-Scale Diurnal Mode of the Asian Summer Monsoon.” *Monthly Weather Review*, 462–473.

Krishnamurti, T. N. (1971). “Tropical east-west circulation of the tropical upper troposphere motion field during the northern hemisphere summer.” *Journal of Atmospheric Science*, 28, 1342–1347.

Krishnamurti, T. N. and Bhalme, H. N. (1976). “Oscillations of a Monsoon System. Part I. Observational Aspects.” *Journal of the Atmospheric Sciences*, 33(10), 1937–1954.

Krishnan, R., Ayantika, D., Kumar, V. and Pokhrel, S. (2011). “The long-lived monsoon depressions of 2006 and their linkage with the Indian Ocean Dipole.” *International Journal of Climatology*, 31(9), 1334–1352.

Kumar, A., Hoerling, M., Ji, M., Leetmaa, A. and Sardeshmukh, P. (1996). “Assessing a GCM’s suitability for making seasonal predictions.” *Journal of climate*, 9(1), 115–129.

Kumar, K. K., Rajagopalan, B., Hoerling, M., Bates, G. and Cane, M. (2006). “Unraveling the mystery of Indian monsoon failure during El Niño.” *Science*, 314(5796), 115–119.

Kumar, K. K., Soman, M. and Kumar, K. R. (1995). “Seasonal forecasting of Indian summer monsoon rainfall: a review.” *Weather*, 50(12), 449–467.

Kumar, K. N., Rajeevan, M., Pai, D., Srivastava, A. and Preethi, B. (2013). “On the observed variability of monsoon droughts over India.” *Weather and Climate Extremes*, 1, 42–50.

Kumar, O. B. (1988). “Eurasian snow cover and seasonal forecast of Indian summer monsoon rainfall.” *Hydrological sciences journal*, 33(5), 515–525.

Kumar, P., Kumar, K. R., Rajeevan, M. and Sahai, A. (2007). "On the recent strengthening of the relationship between ENSO and northeast monsoon rainfall over South Asia." *Climate Dynamics*, 28(6), 649–660.

Kumar, P., Singh, R., Joshi, P. C. and Pal, P. K. (2011). "Impact of additional surface observation network on short range weather forecast during summer monsoon 2008 over Indian subcontinent." *Journal of Earth System Science*, 53 –64.

Landwehr, N., Hall, M. and Frank, E. (2005). "Logistic Model Trees." *Mach. Learn.*, 59(1-2), 161–205.

Levine, R. C., Turner, A. G., Marathayil, D. and Martin, G. M. (2013). "The role of northern Arabian Sea surface temperature biases in CMIP5 model simulations and future projections of Indian summer monsoon rainfall." *Climate dynamics*, 41(1), 155–172.

Li, J., Swinbank, R., Grotjahn, R. and Volkert, H. (2016). *Dynamics and Predictability of Large-Scale, High-Impact Weather and Climate Events*, volume 2. Cambridge University Press.

Liu, B., Wu, G. and Ren, R. (2015). "Influences of ENSO on the vertical coupling of atmospheric circulation during the onset of South Asian summer monsoon." *Climate Dynamics*, 45(7-8), 1859–1875.

Liu, X. and Yanai, M. (2002). "Influence of Eurasian spring snow cover on Asian summer rainfall." *International Journal of Climatology*, 22(9), 1075–1089.

Lockyer, N. and Lockyer, W. J. S. (1904). "The Behaviour of the Short-Period Atmospheric Pressure Variation over the Earth's Surface." *Proceedings of the Royal Society of London*, 73, 457–470.

Lu, Q. and Getoor, L. (2003). "Link-based classification." *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*. 496–503.

Luo, J.-J., Masson, S., Behera, S., Shingu, S. and Yamagata, T. (2005). “Seasonal climate predictability in a coupled OAGCM using a different approach for ensemble forecasts.” *Journal of climate*, 18(21), 4474–4497.

Makela, J. J., Kelley, M. C. and Beaujardiére, O. (2006). “Convective ionospheric storms: A major space weather problem.” *Space Weather*, 4(2).

Mamgain, A., Dash, S. K. and Sarthi, P. P. (2010). “Characteristics of Eurasian snow depth with respect to Indian summer monsoon rainfall.” *Meteorology and Atmospheric Physics*, 110, 71–83.

Meehl, G. A., Stocker, T. F., Collins, W. D., Friedlingstein, A., Gaye, A. T., Gregory, J. M., Kitoh, A., Knutti, R., Murphy, J. M., Noda, A. et al. (2007). “Global climate projections.”

Michal, W. (2013). *Hybrid classifiers: methods of data, knowledge, and classifier combination*, volume 519. Springer.

Mitra, A. K., Stefanova, L., Kumar, T. S. V. V. and Krishnamurti, T. N. (2005). “Seasonal Prediction for the Indian Monsoon Region with FSU Ocean-atmosphere Coupled Model: Model Mean and 2002 Anomalous Drought.” *pure and applied geophysics*, 162(8), 1431–1454.

Mohammadian, M. (2009). *Encyclopedia of Artificial Intelligence*, chapter Designing Unsupervised Hierarchical Fuzzy Logic Systems. 456–463.

Mooley, D., Parthasarathy, B., Sontakke, N. and Munot, A. (1981). “Annual rain-water over India, its variability and impact on the economy.” *Journal of climatology*, (1), 167–186.

Munot, A. and Kumar, K. K. (2007). “Long range prediction of Indian summer monsoon rainfall.” *Journal of earth system science*, (1), 73 – 79.

Munot, A., Patil, S., Preethi, B. and Singh, N. (2011). “Seasonal behaviour of NCEP-NCAR longwave cloud radiative forcing and its relationship with all-

India summer monsoon rainfall." *International journal of remote sensing*, 32(5), 1421–1430.

Murakami, T. (1987). *Effects of the Tibetan Pleateau, In, Monsoon Meteorology.* Oxford University Press, Inc.

Nanjundiah, R. S., Francis, P., Ved, M. and Gadgil, S. (2013). "Predicting the extremes of Indian summer monsoon rainfall with coupled ocean-atmosphere models." *Current Science*, 104(10), 1380–1393.

Navone, H. and Ceccatto, H. (1994). "Predicting Indian monsoon rainfall: a neural network approach." *Climate Dynamics*, 10(6-7), 305–312.

Neena, J., Suhas, E. and Goswami, B. (2011). "Leading role of internal dynamics in the 2009 Indian summer monsoon drought." *Journal of Geophysical Research: Atmospheres*, 116(D13).

Niels, L., Mark, H. and Eibe, F. (2005). "Logistic Model Trees." *Machine Learning*, 59(1), 161–205.

Nychka, D., Restrepo, J. M. and Tebaldi, C. (2017). "Uncertainty in Climate Predictions." `http://www.mathaware.org/mam/09/essays/UncertainClimate.pdf`.

Office, G. A. (1995). "Global warming : limitations of general circulation models and costs of modeling efforts : report to the Ranking Minority Member, Committee on Commerce, House of Representatives, United States General Accounting Office, (microform)." Technical report.

Okoola, R. and Asnani, G. (1981). "Pressure surge in southwest Indian Ocean in relation to onset and activity of summer monsoon in South India." International Conference on Scientific Results of Monsoon Experiments, Bali, Indonesia, 1.13–1.16.

Ordonez, C. and Zhao, K. (2011). "Evaluating association rules and decision trees to predict multiple target attributes." *Intelligent Data Analysis*, 15(2), 173–192.

Orozco-del Castillo, M., Ortiz-Alemán, C., Urrutia-Fucugauchi, J. and Rodríguez-Castellanos, A. (2011). “Fuzzy logic and image processing techniques for the interpretation of seismic data.” *Journal of Geophysics and Engineering*, 8(2), 185.

Oweis, R. J., Abdulhay, E. W., Khayal, A., Awad, A. et al. (2015). “An alternative respiratory sounds classification system utilizing artificial neural networks.” *Biomedical journal*, 38, 153.

Pai, D. and Bhan, S. (2015). “Monsoon 2014, a report.” Technical report, National Climate Centre, India Meteorological Department, Pune, India.

Pai, D. and Rajeevan, M. (2007). *Indian summer monsoon onset: variability and prediction.* National Climate Centre, Indian Meteorological Department.

Palmer, T., Doblas-Reyes, F., Hagedorn, R., Alessandri, A., Gualdi, S., Andersen, U., Feddersen, H., Cantelaube, P., Terres, J., Davey, M. et al. (2004). “Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER).” *Bulletin of the American Meteorological Society*, 85(6), 853–872.

Pang, H., He, Y., Lu, A., Zhao, J., Song, B., Ning, B. and Yuan, L. (2005). “Influence of Eurasian snow cover in spring on the Indian Ocean Dipole.” *Climate Research*, 30(1), 13–19.

Pant, G. and Rupa, K. K. (1997). *Climates of South Asia.* John Wiley and Sons, Chichester.

Pappenberger, F., Bartholmes, J., Thielen, J., Cloke, H. L., Buizza, R. and de Roo, A. (2008). “New dimensions in early flood warning across the globe using grand-ensemble weather predictions.” *Geophysical Research Letters*, 35(10).

Parthasarathy, B., Kumar, K. R. and Munot, A. (1991). “Evidence of secular variations in Indian monsoon rainfall–circulation relationships.” *Journal of Climate*, 4(9), 927–938.

Parthasarathy, B., Kumar, K. R. and Munot, A. (1992). "Surface pressure and summer monsoon rainfall over India." *Advances in atmospheric sciences*, 9(3), 359–366.

Parthasarathy, B., Munot, A. and Kothawale, D. (1995). "All India monthly and seasonal rainfall series : 1871-1993." *Journal of theoritical and applied climatology*, 49, 217–224.

Parthasarathy, B., Rupa, K. K. and Munot, A. (1993). "Homogeneous Indian Monsoon Rainfall : variability and prediction." *Proceedings of Indian acadamy of science*, 121–155.

Parthasarathy, B., Sontakke, N., Munot, A. and Kothawale, D. (1987). "Droughts/floods in the summer monsoon rainfall season over different meteorological subdivisions of India for the period 1871-1984." *Journal of climatology*, (7), 57–70.

Pattnayak, K., Panda, S., Saraswat, V. and Dash, S. (2016). "Relationship between tropospheric temperature and Indian summer monsoon rainfall as simulated by RegCM3." *Climate dynamics*, 46(9-10), 3149–3162.

Peings, Y. and Douville, H. (2010). "Influence of the Eurasian snow cover on the Indian summer monsoon variability in observed climatologies and CMIP3 simulations." *Climate dynamics*, 34(5), 643–660.

Peng, P., Kumar, A., van den Dool, H. and Barnston, A. G. (2002). "An analysis of multimodel ensemble predictions for seasonal climate anomalies." *Journal of Geophysical Research: Atmospheres*, 107(D23).

Peters, G. (2008). *Encyclopedia of Decision Making and Decision Support Technologies*, chapter Uncertainty and Vagueness Concepts in Decision Making. IGI Global, 901–909.

Pillai, P. A. and Mohankumar, K. (2010). "Individual and combined influence of El Niño-Southern Oscillation and Indian Ocean Dipole on the Tropospheric

Biennial Oscillation." *Quarterly Journal of the Royal Meteorological Society*, 136(647), 297–304.

Plummer, N., Allsopp, T. and lopez, J. A. (2003). "Guidelines on Climate Observation Networks and Systems." Wmo/td no. 1185, World Meteorological Organization.

Pokhrel, S., Saha, S. K., Dhakate, A., Rahman, H., Chaudhari, H. S., Salunke, K., Hazra, A., Sujith, K. and Sikka, D. (2016). "Seasonal prediction of Indian summer monsoon rainfall in NCEP CFSv2: forecast and predictability error." *Climate Dynamics*, 46(7-8), 2305–2326.

Prabhu, A., Mahajan, P., Khaladkar, R. and Bawiskar, S. (2009). "Connection between Antarctic sea-ice extent and Indian summer monsoon rainfall." *International Journal of Remote Sensing*, 30(13), 3485–3494.

Prabhu, A., Mahajan, P., Khaladkar, R. and Chipade, M. (2010). "Role of Antarctic circumpolar wave in modulating the extremes of Indian summer monsoon rainfall." *Geophysical Research Letters*, 37(14).

Pradhan, P. K., Prasanna, V., Lee, D. Y. and Lee, M.-I. (2016). "El Niño and Indian summer monsoon rainfall relationship in retrospective seasonal prediction runs: experiments with coupled global climate models and MMEs." *Meteorology and Atmospheric Physics*, 128(1), 97–115.

Prasad, K. and Singh, S. (1996). "Forecasting the spatial variability of the Indian monsoon rainfall using canonical correlation." *International journal of climatology*, 16(12), 1379–1390.

Preethi, B. and Revadekar, J. (2009). "Impact of Summer Monsoon Precipitation on Winter Crop Yields Across India." *IUP Journal of Soil & Water Sciences.*

Preethi, B., Revadekar, J. and Kripalani, R. (2011). "Anomalous behaviour of the Indian summer monsoon 2009." *Journal of earth system science*, 120(5), 783–794.

Puranik, S., Ray, K., Sen, P. and Kumar, P. P. (2013). "An index for predicting the onset of monsoon over Kerala." *Curr Sci*, 105(7), 954–961.

Puranik, S. S., Ray, K., Sen, P. and Kumar, P. P. (2014). "Impact of cross-equatorial meridional transport on the performance of the southwest monsoon over India." *Current Science*, 107(6).

Rai, A., Saha, S. K., Pokhrel, S., Sujith, K. and Halder, S. (2015). "Influence of preonset land atmospheric conditions on the Indian summer monsoon rainfall variability." *Journal of Geophysical Research: Atmospheres*, 120(10), 4551–4563.

Rajeevan, M. (2001). "Prediction of Indian summer monsoon: Status, problems and prospects." *Current Science*, 81(11), 1451–1457.

Rajeevan, M., Bhate, J., Kale, J. and Lal, B. (2006). "High resolution daily gridded rainfall data for the Indian region: analysis of break and active monsoon spells." *Current science*, 296 –306.

Rajeevan, M., Guhathakurta, P. and Thapliyal, V. (2000). "New models for long range forecasts of summer monsoon rainfall over North West and Peninsular India." *Meteorology and Atmospheric Physics*, 73(3), 211–225.

Rajeevan, M. and Nanjundiah, R. S. (2009). "Coupled model simulations of twentieth century climate of the Indian summer monsoon." *Platinum Jubilee Special Volume of the Indian Academy of Sciences*, 537–568.

Rajeevan, M., Pai, D., Dikshit, S. and Kelkar, R. (2004). "IMD's new operational models for long-range forecast of southwest monsoon rainfall over India and their verification for 2003." *Current Science*, 86(3), 422–431.

Rajeevan, M., Pai, D., Kumar, R. A. and Lal, B. (2007). "New statistical models for long-range forecasting of southwest monsoon rainfall over India." *Climate Dynamics*, 28(7-8), 813–828.

Rajeevan, M., Thapliyal, V., Patil, S. and De, U. (1999). "Canonical Correlation Analysis (CCA) models for long-range forecasts of sub-divisional monsoon rainfall over India." *Mausam*, 50, 145–152.

Rajeevan, M., Unnikrishnan, C., Bhate, J., Niranjan Kumar, K. and Sreekala, P. (2012). "Northeast monsoon over India: variability and prediction." *Meteorological Applications*, 19(2), 226–236.

Raju, P., Mohanty, U. and Bhatla, R. (2005). "Onset characteristics of the southwest monsoon over India." *International journal of climatology*, 25(2), 167–182.

Ramaswamy, C. (1962). "Breaks in the Indian summer monsoon as a phenomenon of interaction between the easterly and the sub-tropical westerly jet streams." *Tellus*, 14(3), 337–349.

Ramesh Kumar, M., Sankar, S. and Reason, C. (2009). "An investigation into the conditions leading to monsoon onset over Kerala." *Theoretical and Applied Climatology*, 95(1), 69–82.

Rasmusson, Eugene, M. and Thomas Carpenter, H. (1982). "Variations in Tropical Sea Surface Temperature and Surface Wind Fields Associated with the Southern Oscillation/El Niño." *Monthly Weather Review*, 110, 354–384.

Rayner, N., A, Parker, D., E, Horton, E., B, Folland, C., K, Alexander, L., V, Rowell, D., P, Kent, E., C and Kaplan, A. (2003). "Global analyses of SST, sea ice and night marine air temperature since the late nineteenth century." *Journal Geophysical Research*, 108, 4407.

Revadekar, J. and Kulkarni, A. (2008). "The El Nino-Southern Oscillation and winter precipitation extremes over India." *International Journal of Climatology*, 28(11), 1445–1452.

Reynolds, R. W., Smith, T. m., Liu, C., Chelton, D. B., Casey, K. S. and Schlax, M. G. (2007). "Daily High-Resolution-Blended Analyses for Sea Surface Temperature." *Journal of Climate*, 20(22), 5473–5496.

Rokach, L. (2010). “Ensemble-based classifiers.” *Artificial Intelligence Review*, 33(1-2), 1–39.

Russell, D. and Robert, J. M. (1999). *Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks.* Cambridge, MIT Press.

Sabeerali, C., Rao, S. A., Ajayamohan, R. and Murtugudde, R. (2012). “On the relationship between Indian summer monsoon withdrawal and Indo-Pacific SST anomalies before and after 1976/1977 climate shift.” *Climate dynamics*, 39(3-4), 841–859.

Sadhuram, Y. and Ramana Murthy, T. (2008). “Simple multiple regression model for long range forecasting of Indian summer monsoon rainfall.” *Meteorology and Atmospheric Physics*, 99(1), 17–24.

Saha, K. (1974). “Some aspects of the Arabian Sea summer monsoon.” *Tellus*, 26, 464–476.

Saha, S., Nadiga, S., Thiaw, C., Wang, J., Wang, W., Zhang, Q., Van den Dool, H., Pan, H.-L., Moorthi, S., Behringer, D. et al. (2006). “The NCEP climate forecast system.” *Journal of Climate*, 19(15), 3483–3517.

Saha, S. K., Pokhrel, S. and Chaudhari, H. S. (2013). “Influence of Eurasian snow on Indian summer monsoon in NCEP CFSv2 freerun.” *Climate dynamics*, 41(7-8), 1801–1815.

Saha, S. K., Pokhrel, S., Chaudhari, H. S., Dhakate, A., Shewale, S., Sabeerali, C., Salunke, K., Hazra, A., Mahapatra, S. and Rao, A. S. (2014). “Improved simulation of Indian summer monsoon in latest NCEP climate forecast system free run.” *International Journal of Climatology*, 34(5), 1628–1641.

Sahai, A., Soman, M. and Satyan, V. (2000). “All India summer monsoon rainfall prediction using an artificial neural network.” *Climate dynamics*, 16(4), 291–302.

Sahana, A., Ghosh, S., Ganguly, A. and Murtugudde, R. (2015). "Shift in Indian summer monsoon onset during 1976/1977." *Environmental Research Letters*, 10(5), 054006.

Sajani, S., Sulochana, G., Francis, P. and Rajeevan, M. (2015). "Prediction of Indian rainfall during the summer monsoon season on the basis of links with equatorial Pacific and Indian Ocean climate indices." *Environmental Research Letters*.

Sankar, S., Kumar, M. R. and Reason, C. (2011). "On the relative roles of El Nino and Indian Ocean Dipole events on the Monsoon Onset over Kerala." *Theoretical and applied climatology*, 103(3-4), 359–374.

Sathiyamoorthy, V., Pal, P. K. and Joshi, P. C. (2007). "Intraseasonal variability of the Tropical Easterly Jet, Meteorology and Atmospheric Physics." *Meteorology and Atmospheric Physics*, 96, 305–316.

Saunders, R. W. and Kriebel, K. T. (1988). "An improved method for detecting clear sky and cloudy radiances from AVHRR data." *International Journal of Remote Sensing*, (1), 123 – 150.

Schiffer, R. (1992). "Greenhouse Effect Detection Experiment (GEDEX): Global Climate Change Datasets. National Aeronautics and Space Administration."

Senan, R., Orsolini, Y. J., Weisheimer, A., Vitart, F., Balsamo, G., Stockdale, T. N., Dutra, E., Doblas-Reyes, F. J. and Basang, D. (2016). "Impact of springtime Himalayan–Tibetan Plateau snowpack on the onset of the Indian summer monsoon in coupled seasonal forecasts." *Climate Dynamics*, 47(9-10), 2709–2725.

Sengupta, D., Senan, R., Goswami, B. and Vialard, J. (2007). "Intraseasonal variability of equatorial Indian Ocean zonal currents." *Journal of Climate*, 20(13), 3036–3055.

Shackley, S., Young, P., Parkinson, S. and Wynne, B. (1998). "Uncertainty, Complexity and Concepts of Good Science in Climate Change Modelling: Are GCMs the Best Tools?" *Climatic Change*, (2), 159 –205.

Shaw, R. (2009). *Disaster Management: Global challenges and local solutions.* Universities Press.

Shukla, J. (1987). "Interannual variability of monsoons." *Monsoons*, 399–463.

Shukla, J. and Daniel A, P. (1983). "The Southern Oscillation and Long-Range Forecasting of the Summer Monsoon Rainfall over India." *Monthly Weather Review*, 111, 1830–1837.

Sikka, D. and Gray, W. (1981). "Cross-hemispheric actions and the onset of the summer monsoon over India." International Conference on Scientific Results on Monsoon Experiments, Bali, Indonesia, 3.74 – 3.78.

Singh, A., Kulkarni, M. A., Mohanty, U. C., Kar, S. C., Robertson, A. W. and Mishra, G. (2012). "Prediction of Indian summer monsoon rainfall (ISMR) using canonical correlation analysis of global circulation model products." *Meteorological Applications*, 19(2), 179–188.

Singh, N. and Sontakke, N. (1999). "On the variability and prediction of rainfall in the post-monsoon season over India." *International journal of climatology*, 19(3), 309–339.

Singh, O. and Pai, D. (1996). "An oceanic model for the prediction of SW monsoon Rainfall Over India." *Mausam*, 47, 91–98.

Singh, O. P., Ali Khan, T. M. and Rahman, M. S. (2002). "Impact of Southern Oscillation on the Frequency of Monsoon Depressions in the Bay of Bengal." *Natural Hazards*, 25(2), 101–115.

Singh, P. and Borah, B. (2013). "Indian summer monsoon rainfall prediction using artificial neural network." *Stochastic environmental research and risk assessment*, 27(7), 1585–1599.

Sivakumar, V., Saha, M., Mitra, P. and Banerjee, A. (2015). "Predicting Indian summer monsoon rainfall (ISMR) using a mixture of regression model."

Soman, M. and Kumar, K. K. (1993). "Space-time evolution of meteorological features associated with the onset of Indian summer monsoon." *Monthly Weather Review*, 121, 1177–1194.

Sontakke, N., Singh, H. and Singh, N. (2008). "Chief features of physiographic rainfall variations across India during instrumental period (1813-2006)." *IITM, Research Report No. RR-121.* Indian Institute of Tropical Meteorology, 128.

Sperber, K., Brankovic, C., Deque, M., Frederiksen, C., Graham, R., Kitoh, A., Kobayashi, C., Palmer, T., Puri, K., Tennant, W. et al. (2001). "Dynamical seasonal predictability of the Asian summer monsoon." *Monthly Weather Review*, 129(9), 2226–2248.

Sperber, K. and Palmer, T. (1996). "Interannual tropical rainfall variability in general circulation model simulations associated with the Atmospheric Model Intercomparison Project." *Journal of Climate*, 9(11), 2727–2750.

Sreekala, P., Rao, S. V. B. and Rajeevan, M. (2012). "Northeast monsoon rainfall variability over south peninsular India and its teleconnections." *Theoretical and Applied Climatology*, 108(1-2), 73–83.

Steiner, M. (2009). "Translation of ensemble-based weather forecasts into probabilistic air traffic capacity impact." *Digital Avionics Systems Conference, 2009. DASC'09. IEEE/AIAA 28th.* IEEE, 2–D.

Stolbova, V., Surovyatkina, E., Bookhagen, B. and Kurths, J. (2016). "Tipping elements of the Indian monsoon: Prediction of onset and withdrawal." *Geophysical Research Letters*, 43, 3982–3990.

Stone, P. and Risbey, J. (1990). "On the limitations of general circulation climate models." *Geophysical Research Letters*, 17(12), 2173–2176.

Sumner, M., Frank, E. and Hall, M. (2005). “Speeding up Logistic Model Tree Induction.” 9^{th} European Conference on Principles and Practice of Knowledge Discovery in Databases, 675–683.

Taniguchi, K., Rajan, D. and Koike, T. (2010). “Erratum to: Effect of the variation in the lower tropospheric temperature on the wind onset of the Indian summer monsoon.” *Meteorology and Atmospheric Physics*, 106(3-4), 227–227.

Thapliyal, V. (1982). “Stochastic dynamic model for long-range prediction of monsoon rainfall in peninsular India.” *Mausam*, 33, 399–404.

Thapliyal, V. (1987). *Climate of China and Global Climate.* China Ocean Press.

Thapliyal, V. (2001). “Long-range forecast of summer monsoon rainfall over India, Evolution and development of new power transfer model.” *Indian National Science Academy*, 67, 343–359.

Thapliyal, V. and Rajeevan, M. (2003). “Updated operational models for long-range forecasts of Indian summer monsoon rainfall.” *Mausam*, 54, 495âĂŞ504.

Thomson, M. C., Molesworth, A. M., Djingarey, M. H., Yameogo, K., Belanger, F. and Cuevas, L. E. (2006). “Potential of environmental models to predict meningitis epidemics in Africa.” *Tropical Medicine & International Health*, 11(6), 781–788.

Tomita, T. and Yasunari, T. (1996). “Role of the northeast winter monsoon on the biennial oscillation of the ENSO/monsoon system.” *Journal of the Meteorological Society of Japan. Ser. II*, 74(4), 399–413.

Tripathi, S. and Govindaraju, R. S. (2008). “Statistical forecasting of Indian summer monsoon rainfall: An enduring challenge.” *Soft Computing Applications in Industry.* Springer, 207–224.

Tu, J. V. (1996). “Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes.” *Journal of Clinical Epidemiology*, 49(11), 1225 – 1231.

Varikoden, H. and Babu, C. (2015). "Indian summer monsoon rainfall and its relation with SST in the equatorial Atlantic and Pacific Oceans." *International Journal of Climatology*, 35(6), 1192–1200.

Vathsala, H. and Koolagudi, S. G. (2016). "Long-range prediction of Indian summer monsoon rainfall using data mining and statistical approaches." *Theoretical and Applied Climatology*, 1–15.

Vathsala, H. and Koolagudi, S. G. (2017). "Prediction model for peninsular Indian summer monsoon rainfall using data mining and statistical approaches." *Computers & Geosciences*, 98, 55–63.

Verma, R., Subramaniam, K. and Dugam, S. (1985). "Interannual and long-term variability of the summer monsoon and its possible link with northern hemispheric surface air temperature." *Proceedings of the Indian Academy of Sciences-Earth and Planetary Sciences*, 94(3), 187–198.

Vernekar, A., Zhou, J. and Shukla, J. (1995). "The effect of Eurasian snow cover on the Indian monsoon." *Journal of Climate*, 8(2), 248–266.

Walker, G. (1918). "Correlation in seasonal variation of weather." *Quarterly Journal of the Royal Meteorological Society*, 44, 223–224.

Walker, G. (1923). "Correlation in seasonal variations of weather. VIII: A preliminary study of world weather." *Memoirs of India Meteorological Department*, 24, 75–131.

Walker, G. (1924). "Correlations in seasonal variations of weather, IX, A further study of world weather (World weather II)." *Memoirs of India Meteorological Department*, 24, 275–332.

Wang, B. (2006). *The asian monsoon.* Springer Science & Business Media.

Wang, B., Ding, Q. and Joseph, P. (2009). "Objective definition of the Indian summer monsoon onset." *Journal of Climate*, 22(12), 3303–3316.

Wang, B., Lee, J.-Y. and Xiang, B. (2015). “Asian summer monsoon rainfall predictability: a predictable mode analysis.” *Climate Dynamics*, 44(1-2), 61–74.

Wilks, D., S (1995). *Statistical Methods in Atmospheric Sciences*. Academic Press, Waltham.

Wu, G., Mao, J. and Duan, A. (2004). “Recent progress in the study on the impacts of Tibetan Plateau on Asian summer monsoon.” *Acta Meteor. Sinica*, 62, 529–540.

Wu, Z., Zhang, P., Chen, H. and Li, Y. (2016). “Can the Tibetan Plateau snow cover influence the interannual variations of Eurasian heat wave frequency?” *Climate Dynamics*, 46(11-12), 3405–3417.

Xie, J., Wu, J. and Qian, Q. (2009). “Feature Selection Algorithm Based on Association Rules Mining Method.” Eighth IEEE/ACIS International Conference on Computer and Information Science, 357–362.

Xie, P. and Arkin, P. A. (1996). “Analyses of global monthly precipitation using gauge observations, satellite estimates, and numerical model predictions.” *Journal of Climate*, 840–858.

Xie, P. and Arkin, P. A. (1997). “Global Precipitation: A 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs.” *Bulletin of the American Meteorological Society*, 2539–2558.

Xue, F., Wang, H. and He, J. (2003). “Interannual variability of Mascarene high and Australian high and their influences on summer rainfall over East Asia.” *Chinese Science Bulletin*, 48(5), 492–497.

Yadav, R. (2009). “Changes in the large-scale features associated with the Indian summer monsoon in the recent decades.” *International Journal of Climatology*, (1), 117 – 133.

Yadav, R. K. (2013). "Emerging role of Indian ocean on Indian northeast monsoon." *Climate dynamics*, 41(1), 105–116.

Yanai, M. and Song, Z. (1992). "Seasonal heating of the Tibetan Plateau and its effects on the evolution of the Asian Summer Monsoon." *Journal of Meteorological Society of Japan*, 70, 319–351.

Yasunari, T. (1985). "Zonally propagating modes of the global east-west circulation associated with the southern oscillation." *Journal of the Meteorological Society of Japan*, 63, 1010–1019.

Ye, D.-Z. and Wu, G.-X. (1998). "The role of the heat source of the Tibetan Plateau in the general circulation." *Meteorology and Atmospheric Physics*, 67(1), 181–198.

Zadeh, L. A. (1965). "Fuzzy sets." *Information and control*, 8, 338–353.

Zhu, C., Park, C.-K., Lee, W.-S. and Yun, W.-T. (2008). "Statistical downscaling for multi-model ensemble prediction of summer monsoon rainfall in the Asia-Pacific region using geopotential height field." *Advances in Atmospheric Sciences*, 25(5), 867–884.

LIST OF PUBLICATIONS

Journal Publications

[1] Vathsala.H and Shashidhar G Koolagudi, Long-Range Prediction of Indian Summer Monsoon Rainfall using Data Mining and Statistical Approaches, Theoretical and Applied climatology, Springer.

[2] Vathsala.H, Shashidhar G Koolagudi, Prediction Model for Peninsular Indian Summer Monsoon Rainfall Using Data Mining and Statistical Approaches, Computers & Geosciences.

[3] Vathsala.H, Shashidhar G Koolagudi, Neuro-Fuzzy model for quantifying Summer monsoon rainfall prediction in north interior Karnataka using data mining and soft computing approaches, Engineering Applications of Artificial Intelligence (Communicated)

[4] Vathsala.H, Shashidhar G Koolagudi, Indian summer monsoon - Database, Predictors, Models, Prediction - A review , Theoretical and applied climatology (Communicated)

Conference Publications

[1] H. Vathsala and K. C. Shet, Application of multi dimensional sequential pattern mining algorithm on non sequential multi dimensional climate data, Fifth International Conference on Advances in Recent Technologies in Communication and Computing (ARTCom 2013), 2013.

[2] Vathsala H and Shashidhar G Koolagudi, Closed Item-Set Mining for Prediction of Indian Summer Monsoon Rainfall A Data Mining Model with Land and Ocean Variables as predictors, Eleventh International Multi-Conference on Information Processing-2015 (IMCIP- 2015), 2015.

BRIEF BIO-DATA

Personal Details

Name - Vathsala H

Date of Birth - 10/07/1981

Work Address

Vathsala H
(CDAC) No-1, Knowledge park
Byappanahalli
Bangalore, 560038
Email: vathsala.h@gmail.com

Permanent Address

Vathsala H
Siri Nivasa, No. 2/120
Golikatte, Havanje village
Bellampalli P.O
Udupi, 576124 Karnataka

Qualification

M. Tech. VLSI Design & Embedded Systems, BMS college, Bangalore, Karnataka, 2007.

B. E. Information Technology, GVIT, K.G.F, Karnataka, 2003.

Current Employment

Center for Development of Advance Computing (CDAC), Bangalore(2008 Feb - Till date) Senior technical officer

Previous Work experience

National Aerospace Laboratories, Bangalore (2004 April-2006 March)
as a Research intern