

**Interpretation of Weather Forecasts for
Tornado and Cloudburst
using Data Mining Techniques**

THESIS

Submitted in partial fulfilment
of the requirements for the degree of
DOCTOR OF PHILOSOPHY

by

KAVITA KAPOOR

Under the Supervision of
Prof. Rattan K. Datta



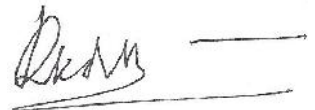
**BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE
PILANI (RAJASTHAN) INDIA
2012**

**BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE
PILANI (RAJASTHAN)**

CERTIFICATE

Research is important for growth of knowledge. If the research can be directly or indirectly used for real life societal application, it has a special significance. After discussions, the author was asked to work on Interpretation of Weather Forecasts for Sub-Grid scale weather systems like Tornado and Cloudburst using Data Mining Techniques. This required not only her understanding of Data Mining tools but also the basics of meteorology and Sub-Grid Scale weather events causing extreme weather. I am pleased to state that Kavita Kapoor has done justice and worked in this nearly virgin area under my guidance.

This is to certify that this thesis entitled **Interpretation of Weather Forecasts for Tornado and Cloudburst using Data Mining Techniques** and submitted by **Kavita Kapoor** ID No **2007PHXF449P** for award of Ph. D. Degree of the Institute embodies original work done by her under my supervision.



Signature in full of the Supervisor: _____

Name in capital block letters: **DR. RATTAN K. DATTA**

Designation - **Director – MERIT**

Hon. Director-Research, CSI

President – GAMS

Vice Chairman- TC5 of IFIP

Date: 4th May, 2012

Acknowledgements

First of all I thank the Almighty God for his blessings throughout the entire sphere of my life.

I would like to express my deepest sense of gratitude to my Supervisor, Dr. Rattan K. Datta, Former Advisor, Department of Science & Technology, Government of India and currently Director, Mohyal Educational Research Institute of Technology for his encouragement, guidance and mentoring. He has always been inspiring me with his aura full of positive energy. Without his support, it would not have been possible for me to take up research in this challenging field. It has been a great learning while working with him on this research area.

I would also like to deeply acknowledge the support of Dr. Ashis Kumar Das, Dean, Research and Consultancy division, Birla Institute of Technology and Science, Pilani. My thanks are also due to the members of my Doctoral Advisory Committee comprising of Dr. Poonam Goyal and Dr. M. K. Rohil for their invaluable suggestions, which significantly improved the quality of this thesis.

I also express my sincere gratitude to Dr. R. N. Saha, Deputy Director (Research), BITS Pilani and Dr. G. Raghurama, Director, BITS Pilani for their support. I wish to acknowledge my profound gratefulness to Professor (Dr.) B. N. Jain, Hon'ble Vice Chancellor, BITS Pilani, for giving me an opportunity to take up research.

I am also extremely grateful to the Head, Agromet Division, Indian Meteorological Department (IMD) Delhi; Head, Research Division, National Center for Medium Range Weather Forecasting (NCMRWF) Noida; Head, Numerical Weather Predictions, IMD Delhi; Director-NHAC, IMD Delhi; Director, National Climate Centre, IMD Pune; Head, Computer & Networking Division, NCMRWF Noida for providing me their valuable inputs and also Numerical Weather Prediction Model forecasts for my research.

I wish to thank my colleague and friend Neetu Narwal for her support and help. Thanks also go to all the faculty members and management of my place of work – Maharaja Surajmal Institute, for their co-operation throughout the research.

No set of acknowledgements can ever be complete without my parents who have always supported me tirelessly, helped me and motivated me beyond measures.

Special gratitude goes to my husband Manoj. I cannot put into words what his love and support has done for me. As a mother, I feel deeply obliged to my small kids, Ujjawal and Nishtha for the long hours which I have deprived them off to do my research.

Thank you one and all.

Abstract

With the advent of digital computers and their continuous increasing processing power, the 'Numerical Weather Prediction' (NWP) models which solve a close set of equations representing atmospheric flow, have been adopted by most of the meteorological services to issue day to day weather forecasts. These forecasts are issued for public in general, farmers, pilots, water works managers, health departments, planners, disaster management services etc.

NWP models continue to improve in resolution (both horizontal and vertical) as well as sophistication in including various atmospheric processes. Despite this, there are various limitations, specifically:

- (i) There are many important processes and scales of motion in the atmosphere especially Sub-Grid scale weather phenomenon that cannot be explicitly resolved with present models. Some of the significant Sub-Grid scale phenomena are Tornado and cloudbursts. Various empirical techniques are used by the experienced forecasters to infer these events.
- (ii) The NWP output consists of mainly flow patterns namely wind, temperature, humidity, and pressure fields at various temporal and spatial levels. The forecast of actual weather elements like rain/ snow etc. are derived from the NWP output products through statistical relationship, known as Model Output Statistics (MOS). But MOS is not a theoretically stable process as it requires longer datasets on time scale to derive the MOS relationships for forecasting of weather elements. Longer and consistent datasets are not available because of frequent revisions of NWP model.

There is thus a strong need for searching alternative tools to MOS for interpretation of weather patterns provided by NWP models into weather elements including cloudburst occurrence and tornado. According to Literature Survey, Intelligent systems have been applied generally for processes like decision making applications of evacuation of public in case of cyclone hit, probability of occurrence of rainfall / snow etc. but there was very limited work done on interpreting and deriving weather elements from the NWP output products except MOS which has limitations as explained already. It has been noted that Data Mining (DM) techniques have been used for prediction of snow/ no snow, to determine the relationship between the trajectories of Mesoscale Convective systems moving out of the plateau of Tibet and their environmental physical field values, to predict temperature and pressure. But Data Mining has not necessarily been used to derive actual weather phenomenon from NWP output products. Especially limited

work has been done for interpretation of Sub-Grid scale related extreme weather like Tornado and Cloudburst. Keeping in view of these significant considerations, the interpretation of Sub-Grid scale weather systems was considered important and following methodology was used.

- (i) Selection of the output products of European Center for Medium-range Weather Forecasting (ECMWF) and Weather Research and Forecasting (WRF) models with reference to recent occurrences of Cloudbursts and Tornado in India.
- (ii) Pre-processing of the NWP model output products as the meteorological data is in standardized meteorological format and is huge and multidimensional. This required implementation of Multidimensional Data Model which considerably improved the storage and retrieval of required weather variables on a selected time and space scale. The important ingredients to tornado and cloudburst formation have been derived using primary output forecasts provided by the models.
- (iii) Application of Clustering technique on the synoptic scale data of well formed Low Pressure Systems (LPS) which form near head Bay of Bengal and move west, west north-west to north-west and produce rainfall in their wake, was considered. The results convinced the use of clustering techniques on sub-grid scale weather processes.
- (iv) Generation of clusters of ensemble (forecast valid for a particular time based on different initial conditions) of derived weather variable *viz.* convergence at various atmospheric pressure levels for a real life case of tornado corresponding to two different NWP models forecasts. Also, four real life cases of cloudbursts have been analyzed using k-means clustering technique.

This approach resulted in locating patterns conducive to formation of cloudbursts four days in advance. The data availability was a limitation but promising advanced signals of formation of patterns have been shown in the study.

The advantage and contribution of the study lies in, (i) the demonstration of the application of Multidimensional Data Model in meteorological research and (ii) the application of new approach of interpretation of NWP output products through DM tools compared to unstable MOS. (iii) The study has also shown a promise of advance signals on the formation of sub-grid scale weather events from NWP output products which can minimize the losses.

An effort has also been made to test other Intelligent systems like Artificial Neural Networks and Adaptive Neuro-Fuzzy Inference System. This is a new field of research and can have further progress. More studies on application of other intelligent tools could form future work for making day to day operational forecast by meteorological services.

Acknowledgement.....	(iii)
Abstract.....	(v)
Table of Contents.....	(vii)
List of Tables.....	(xi)
List of Figures.....	(xii)
List of Abbreviations.....	(xv)

Table of Contents

1. Introduction	
1.1 NWP models and sub-grid scale weather events	1
1.1.1 Sub-Grid Scale and Extreme Weather	2
1.2 Present state of forecast	3
1.3 Intelligent Systems	4
1.3.1 Data mining	5
1.3.2 Fuzzy Logic	5
1.3.3 Genetic Algorithm	5
1.3.4 Artificial Neural Network	6
1.3.5 Bayesian Probabilistic networks	6
1.4 Objective of the Research	7
1.5 Thesis organization	7
2. Literature Review and Problem Definition	
2.1 Data mining Techniques for Weather Forecast.....	8
2.2. Fuzzy Logic Techniques for Weather Forecast.....	15
2.3 Genetic Algorithms for Weather Forecast.....	17
2.4 Artificial Neural Networks for Weather Forecast.....	18
2.5 Bayesian Networks for Weather Forecast.....	20
2.6 Combination of two or more AI techniques for Weather Forecast.....	22
2.7 Problem definition and Objective of the study.....	23
3. Methodology	
3.1 Approach	26
3.1.1 Evolution of Working Methodology.....	27
3.2 Pre-processing of Meteorological datasets	28
3.2.1 Pre-processing of Rainfall datasets	28
3.2.2 Pre-processing of NWP model output products.....	30
3.2.2.1 Pre-processing of ECMWF T-799 model forecasts.....	30

3.2.2.2	Pre-processing of WRF V3.1 model forecasts.....	34
3.3	Multidimensional Data Model	34
3.4	Comparison between the Relational and Multidimensional models.....	37
3.5	Synoptic scale weather system - Low Pressure System	38
3.6	Sub-grid scale weather system - Tornado	38
3.6.1	Tornado genesis and types of Tornadoes	39
3.6.1.1	Supercellular tornadoes	40
3.6.1.2	Multiple vortices from a single-vortex tornado.....	41
3.6.2	Tornado Wind Speed	42
3.6.3	Fujita scale of tornado intensity.....	43
3.6.4	Tornado forecasting.....	44
3.7	Other Sub-grid scale weather system – Cloudburst	45
3.7.1	A conceptual model of the cloudburst	47
3.7.2	Places that suffer from Cloudbursts	49
3.7.3	Forecast of cloudburst	49
4.	Case Studies	
4.1	Analysis of Synoptic scale weather system “Low Pressure System” Movement over Indian Region.....	51
4.1.1	Dataset used.....	51
4.1.2	Technique Used.....	52
4.1.3	Findings.....	53
4.2	Analysis of Rainfall with reference to movement of Low Pressure System over Indian Region	56
4.2.1	Dataset used.....	56
4.2.2	Technique Used and Findings	56
4.3	Artificial Neural Network for Rainfall forecasting	60
4.3.1	Dataset used.....	60
4.3.2	About Artificial Neural Networks	61
4.3.2.1	Network Function	61
4.3.2.2	Training and testing the network	61
4.3.3	Application of ANN to forecast Rainfall	62
4.3.4	Learning Algorithms used	63
4.3.5	Findings	64
4.4	A Multi-Dimensional data model for Meteorological datasets	68
4.4.1	Datasets used	68

4.4.2	Implementation of the 3-Dimensional data cube	69
4.4.3	Findings	71
4.5	A 5-Dimensional data model for Meteorological datasets	73
4.5.1	Datasets used	73
4.5.2	Implementation of the 5-Dimensional data cube	73
4.5.3	Findings	77
4.6	Application of Multidimensional data model for NWP model output products for generating ensembles.....	78
4.6.1	Datasets used	78
4.6.2	Implementation of the Multidimensional data model	79
4.6.3	Findings	82
4.7	Data mining for Interpretation of sub-grid scale weather system –“Tornado” Using NWP output	83
4.7.1	Datasets of ECMWF model under analysis	83
4.7.2	Visualization and interpretation of clusters of convergence	84
4.7.3	Datasets of WRF model under analysis	86
4.7.4	Data mining of WRF output field	86
4.7.5	Visualization of clusters of z-wind component	87
4.7.6	Interpretation of clusters of z-wind component	89
4.8	Data mining for Interpretation of sub-grid scale weather system – “Cloudburst” using NWP (ECMWF) outputs	89
4.8.1	Pre-processing of data – same steps as in section 3.2.2.1	
4.8.2	Technique applied	89
4.8.3	Cloudburst case under consideration- Dhaka	90
4.8.4	Cloudburst case under consideration- Pittorgarh	93
4.8.5	Cloudburst case under consideration- Chamoli	96
4.8.6	Cloudburst case under consideration- Shimla.....	99
4.8.7	Interpretation of visualization of clusters of convergence.....	102
5.	Results and Discussion.....	103
5.1	Synoptic scale weather system “Low Pressure System” Movement over Indian Region	103
5.2	Multidimensional Data Model for Meteorological Datasets	104
5.3	Interpretation of ECMWF and WRF forecast using data mining techniques for tornado forecasting	105
5.4	Interpretation of ECMWF forecast using Data mining techniques for	

Cloudburst forecasting	106
5.5 Artificial Neural Network for Rainfall forecasting	106
6. Conclusions	108
6.1 Pre-processing of NWP model output	108
6.2 Clustering technique for Low Pressure System study	109
6.3 Patterns depicting Tornado formation	109
6.4 Patterns depicting Cloudburst formation	109
6.5 Rainfall Forecasting	110
7. Contributions, Limitations and Future Scope	111
7.1 Contributions	111
7.2 Limitations	112
7.3 Future Scope	112
References.....	114
Appendix A	
(.ctl file describing the .grd file).....	132
Appendix B	
(Program in FORTRAN to convert .grd file (input.grd) to .dat file (output.dat)).....	132
Appendix C	
(Program in Visual Basic to organize gridded rainfall datasets into tabular format).....	133
Appendix D	
(Code in Visual Basic for calculation of vorticity and divergence using Vertical and horizontal wind components of forecast)	134
Appendix E	
(Contents of.ctl file of WRF forecast).....	135
Appendix F	
(Program in C plus plus to convert the wrf forecast file (.dat file) to a .txt file).....	141
Appendix G	
(Implementation of point in polygon algorithm to update fact_rainfall table).....	142
List of Publications	
Journal Publications	144
Conference Publications.....	145
Biography of the candidate	147
Biography of the supervisor	148

List of Tables

3.1	Text file retrieved from .grd file for rainfall in 1989.....	29
3.2	Rainfall for year 1989 organized in tabular format	30
3.3	Calculation of vorticity and convergence based on 72 hour forecast of v and u wind velocities	33
3.4	Fujita scale of tornado intensity.....	44
4.1	A sample of Low Pressure System data for year 1984	52
4.2	A sample of Low Pressure System data (normalized movement) for year 1994.....	56
4.3	A sample of Rainfall for year 1994 corresponding to LPS from 3July1994 to 5July 1994	57
4.4	A Sample of location-wise rainfall for year 1989.....	63
4.5	A sample dataset of dimension table “time”.....	69
4.6	A sample dataset of dimension table “location”.....	70
4.7	A sample dataset of dimension table “low pressure system”.....	70
4.8	A sample dataset of central fact table “fact_rainfall” for 3-Dimensional data model for meteorological datasets.....	71
4.9	A sample dataset of dimension table “river”.....	74
4.10	A sample dataset of dimension table “river details”.....	74
4.11	A sample dataset of dimension table “district”.....	74
4.12	A sample dataset of dimension table “district details”.....	75
4.13	A sample dataset of central fact table “fact_rainfall” for 5-dimensional data model for meteorological datasets.....	75
4.14	A sample of “location” table corresponding to NWP output products	80
4.15	A sample of “time” table corresponding to NWP output products	80
4.16	A sample of central fact table for 2-dimensional data model for NWP model forecasts	81

List of Figures

1.1	Time and Space Scale of Atmospheric Motion	3
3.1	Different Wind velocities	31
3.2	A snapshot of gridded locations to calculate convergence and vorticity.....	32
3.3	Data pre-processing of forecast by ECMWF model.....	33
3.4	A multidimensional cube having 3 dimensions.....	35
3.5	Search of a node corresponding to cuboid in a 5 dimensional data hypercube.....	38
3.6	Vertical wind shear can result in rotation of air around a horizontal axis.....	40
3.7	Vortex tubes with horizontal vorticity are lifted into a vertical position by rising air.....	41
3.8	The four stages in the development of multiple vortices.....	42
3.9	The distribution of winds inside a single-vortex tornado.....	43
3.10	A conceptual model of cloudburst Stage-1: Separate Cells, Stage-2: Merger of Cells and heavy downpour, Stage-3: Dissipation	48
4.1	Area under analysis for LPS movement over Indian region	53
4.2	Clusters of formation and disappearance of LPS during June-July 1984, 85, 88 to 94	54
4.3	Clusters of formation and disappearance of LPS during June-July 1995 to 2003.....	54
4.4	Clusters of formation and disappearance of LPS during Aug-Sept 1984, 85, 88 to 94.	55
4.5	Clusters of formation and disappearance of LPS during Aug-Sept 1995 to 2003.....	55
4.6	Clusters of LPS formation, movement, dissipation during June, July 1984, 85, 88, 89, 90, 91, 92, 93, 94.....	55
4.7	Rainfall with reference to LPS movement from 25June 1994 to 1July 1994.....	58
4.8	Rainfall with reference to LPS movement from 3July 1994 to 5July 1994.....	58
4.9	Rainfall with reference to LPS movement from 3Sept 1994 to 10Sept 1994.....	59
4.10	Rainfall with reference to LPS movement from 25June 1994 to 1July 1994, 3July 1994 to 5July 1994 and 3Sept 1994 to 10Sept 1994	60
4.11	Neuron Model	62
4.12	Result of training ANN with Rainfall data of year 1989 and testing with Rainfall data of year 1990 using learning function traincgf.....	65
4.13	Result of training ANN with Rainfall data of year 1989 and testing with Rainfall	

data of year 1990 using learning function trainrp.....	65
4.14 Result of training ANN with Rainfall data of year 1989 and testing with Rainfall data of year 1990 using learning function trainscg.....	66
4.15 Result of training ANN with Rainfall data of year 1991 and testing with Rainfall data of year 1992 using learning function traincrgf.....	67
4.16 Result of training ANN with Rainfall data of year 1991 and testing with Rainfall data of year 1992 using learning function trainrp.....	67
4.17 Result of training ANN with Rainfall data of year 1991 and testing with Rainfall data of year 1992 using learning function trainscg.....	68
4.18 Star Schema for Rainfall corresponding to Tables 4.5, 4.6, 4.7 and 4.8.....	71
4.19 Analysis of Aggregate Rainfall against date and LPS.....	72
4.20 Aggregate rainfall for June and July months of 1984.....	72
4.21 A 5-D data cube representation of rainfall data, according to dimensions time, gridded-location, Low Pressure system, river catchment area and district.....	76
4.22 Snowflake Schema for Rainfall corresponding to Table 4.5, 4.6, 4.7, 4.9, 4.10, 4.11, 4.12, 4.13.....	76
4.23 Analysis of Aggregate Rainfall against date and LPS.....	77
4.24 Aggregate rainfall for June and July months of 1984 against longitude and Latitude.....	77
4.25 Aggregate rainfall for June and July months of 1984 against River catchment area and district.....	78
4.26 Depicting the creation of ensemble of forecast valid for 1200GMT 8August 2009 with different initial conditions	79
4.27 Star Schema for datasets corresponding to Table 4.14, 4.15 and 4.16.....	81
4.28 Analysis of vorticity against date and Location	82
4.29 Analysis of divergence against date and Location	82
4.30 Forecast of convergence valid for 1200GMT 31 March 09 (Location of tornado: 86°E, 20°N)	85
4.31 3-dimensional visualization of forecast of z-wind, valid for 0600GMT 31 March 09 (Location of tornado: 86°E, 20°N).....	87
4.32 3-dimensional visualization of forecast of z-wind, valid for 1200GMT 31 March 09(Location of tornado: 86°E, 20°N).....	88
4.33 3-dimensional visualization of forecast of convergence, valid for 1800GMT 28 July 09 (Location of cloudburst - Dhaka)	91

4.34	3-dimensional visualization of forecast of convergence, valid for 0000GMT	
	29July 09 (Location of cloudburst - Dhaka)	92
4.35	3-dimensional visualization of forecast of convergence valid for 0600GMT	
	8Aug 09 (Location of cloudburst - Pittorgarh).....	94
4.36	3-dimensional visualization of forecast of convergence valid for 1200GMT	
	8Aug 09(Location of cloudburst - Pittorgarh).....	95
4.37	3-dimensional visualization of forecast of convergence valid for 1200GMT	
	18July 09 (Location of cloudburst - Chamoli).....	97
4.38	3-dimensional visualization of forecast of convergence valid for 1800GMT	
	18July 09 (Location of cloudburst - Chamoli)	98
4.39	3-dimensional visualization of forecast of convergence valid for 1200GMT	
	7Aug 09 (Location of cloudburst - Shimla)	100
4.40	3-dimensional visualization of forecast of convergence valid for 1800GMT	
	7Aug 09 (Location of cloudburst - Shimla).....	101

List of Abbreviations

ACM	Atmospheric Circulation Models
AI	Artificial Intelligence
ALADIN	Aire Limitée et Adaptation Dynamique
ANN	Artificial Neural Network
ARPS	Advanced Regional Prediction System
AUC	Area under the ROC Curve
BN	Bayesian Networks
BPNN	Back Propagation Neural Network
BWER	Bounded Weak-Echo Region
CLARA	Clustering LARge Applications
CLARANS	Clustering Large Applications based upon RANdomized Search
CONQUEST	CONtent-based QUEying in Space and Time
csv	comma separated variables
CTI	Cold Tongue Index
DM	Data Mining
ECMWF	European Center for Medium-range Weather Forecasting
<i>e.g.</i>	<i>exempli gratia</i> ; for example
e-GMDH	enhanced Group Method of Data Handling
ENIAC	Electronic Numerical Integrator And Computer
ENSO	El Niño-Southern Oscillation
ERA-40	ECMWF 40 Year Re-analysis Project
<i>et al.</i>	<i>et alii</i> ; and others
FPAR	Fraction of Photosynthetically Active Radiation
GA	Genetic Algorithm
GANN	Genetic Algorithm Neural Network
GFS	Global Forecast System
GMS	Geostationary Meteorological Satellite
GRIB	GRIdded Binary
GUI	Graphical User Interface
HLAFS	High resolution Limited Area Analysis And Forecasting System
HPCC	High Performance Computing and Communications

ICA	Independent Component Analysis
<i>i.e.</i>	<i>id est.</i> ; that is
IMD	Indian Meteorological Department
JMA	Japan Meteorological Agency
LEAD-CI	Linked Environments for Atmospheric Discovery: A Cyberinfrastructure for Mesoscale Meteorology Research And Education
Matlab	MATrix LABoratory
MCS	Mesoscale Convective System
MDDM	Multi Dimensional Data Models
MLP	Multilayer Perceptron
MOS	Model Output Statistics
MSLP	Mean Sea Level Pressure
NAO	North Atlantic Oscillation
NARR	North American Regional Reanalysis
NASA	National Aeronautics and Space Administration
NASA-CASA	National Aeronautics and Space Administration-Carnegie Ames Stanford Approach
NCAR	National Center for Atmospheric Research
NCEP	National Centers for Environmental Prediction
NCMRWF	National Centre of Medium Range Weather Forecast
NDFD	National Digital Forecast Database
NM	Non Monotonic
NOAA	National Oceanic and Atmospheric Administration
NPP	Net Primary Production
NWF	Numerical Weather Forecast
NWP	Numerical Weather Prediction
PEs	Processing Elements
PoP	Probability of Precipitation
QPF	Quantitative Precipitation Forecast
RIPPER	Repeated Incremental Pruning to Produce Error Reduction
RMSE	Root Mean Square Error
ROC	Receiver Operating Characteristic
SCIT	Storm Cell Identification and Tracking
SLP	Sea Level Pressure
SNN	Shared Nearest Neighbor

SPC	Storm Prediction Center
SPSS	Statistical Package for the Social Sciences
SQL	Structured Query Language
SST	Sea Surface Temperature
SVRG	Support Vector Regression with Genetic algorithms
TC	Tropical Cyclones
UMPL	Unified Model for PoLand area
<i>viz.</i>	<i>videlicet</i> ; namely
WEKA	Waikato Environment for Knowledge Analysis
WRF	Weather Research and Forecasting

Chapter 1

Introduction

1.1 NWP models and sub-grid scale weather events

First attempt to forecast weather through Numerical method was by Richardson during 1920s who worked out the solution of equations through finite differences scheme (Hunt, 1998). Though his results were ridiculous but it led the way for first computer based forecast in 1949 (Lynch, 2007) by Jule Charney, Fjortoft and John von Neumann, using Electronic Numerical Integrator And Computer (ENIAC).

Since then, meteorological communities are more and more depending on Numerical Weather Prediction (NWP) methods for preparing day to day weather forecasts. A number of forecast models, both global and regional in scale, are run to help create forecasts for nations worldwide. Use of model ensemble forecasts helps to define the forecast uncertainty and extend weather forecasting farther into the future than would otherwise be possible. Some of the NWP models are:-

- European Centre for Medium-Range Weather Forecasts (ECMWF) model
- Japan Meteorological Agency (JMA) model
- Fifth-Generation National Centre for Atmospheric Research (NCAR) /Penn State Mesoscale (MM5) Model
- Global Forecast System (GFS) model
- Weather Research and Forecasting (WRF) model
- National Centre of Medium Range Weather Forecast (NCMRWF) Model of India

The significant progress in the observation and collection of weather parameters through various weather monitoring systems viz. satellites, radars, surface and upper air observations, observations from aircrafts, ships and automatic weather stations has been a major contributor for providing initial conditions for the model.

Analysis and forecast of weather data created through NWP models offers an unprecedented opportunity for predicting weather events, provide information and warning of extreme weather events for minimizing losses both to human and property. Such data consist of a sequence of global snapshots of the Earth, typically available at various spatial and temporal intervals including atmospheric parameters over land and ocean (such as temperature, pressure, wind speed, wind direction, sea surface temperature, etc.). Forecast of weather elements like rain/snow, sky conditions etc. at a place are derived through statistical relation popularly known as Model Output Statistics (MOS) proposed by National Weather Service (Hughes, 1976).

Flow patterns are available for a specific grid scale (both horizontal and vertical) according to the resolution of the model. For example ECMWF model provides analysis and forecast at $0.25^\circ \times 0.25^\circ$ in the horizontal scale and 15 atmospheric pressure levels in vertical scale. The phenomenon at the spatial scale lower than the resolution of the model is called sub-grid scale phenomenon. Most of the extreme weather events like cloudburst and tornado/waterspout are sub-grid scale with horizontal coverage of a few hundred meters to kilometer.

The aim of the current research was to work on projects which could lead to significant benefit to society, so the interpretation of sub-grid scale weather events was selected. The development of data assimilation techniques and worldwide scientific validation work has resulted in a sequence of reanalysis with improving quality (Uppala, Dee and Kobayashi, 2008) and also the skill analysis of ECMWF model has improved over the decades.

As stated earlier, despite the continued increase of horizontal and vertical resolution, there are many important processes and scales of motion in the atmosphere that cannot be explicitly resolved with present or future models as their presence is only at the sub-grid scale.

1.1.1 Sub-Grid Scale and Extreme Weather

A meteorological classification of weather systems at time and space scale (Short, 2005) is shown in Figure 1.1. Different weather systems exist at different time and space scales in the atmosphere viz. Hurricanes, Tropical Cyclones exist at synoptic scale (large area of order of 100s to 1000s of km and time from days to weeks), land/sea breeze exist at mesoscale having timescales from hours to days and space scales of 10s to 100s of km, etc. The NWP outputs enable prediction of weather events at a time scale of hours to days to a week or more and at a space scale of 100s of km or more. This is because the grid size of NWP models is of order of 10s to 100s of km. For detection of presence of a weather system of wavelength λ , numerically

the grid-size must be of order less than or equal to $\Delta x/4$. The sub-grid scale weather systems are of space scale 10s to 100s of meters and last for a few minutes only. So, indirect empirical methods are used to delineate the sub-grid scale weather systems as the forecast produced by NWP models can not directly provide an insight. This is where data mining techniques would help to locate patterns of presence of atmospheric features at larger scale that can provide an early signal to formation of sub-grid scale weather systems. The sub-grid scale weather events considered for research work are cloudburst and tornado.

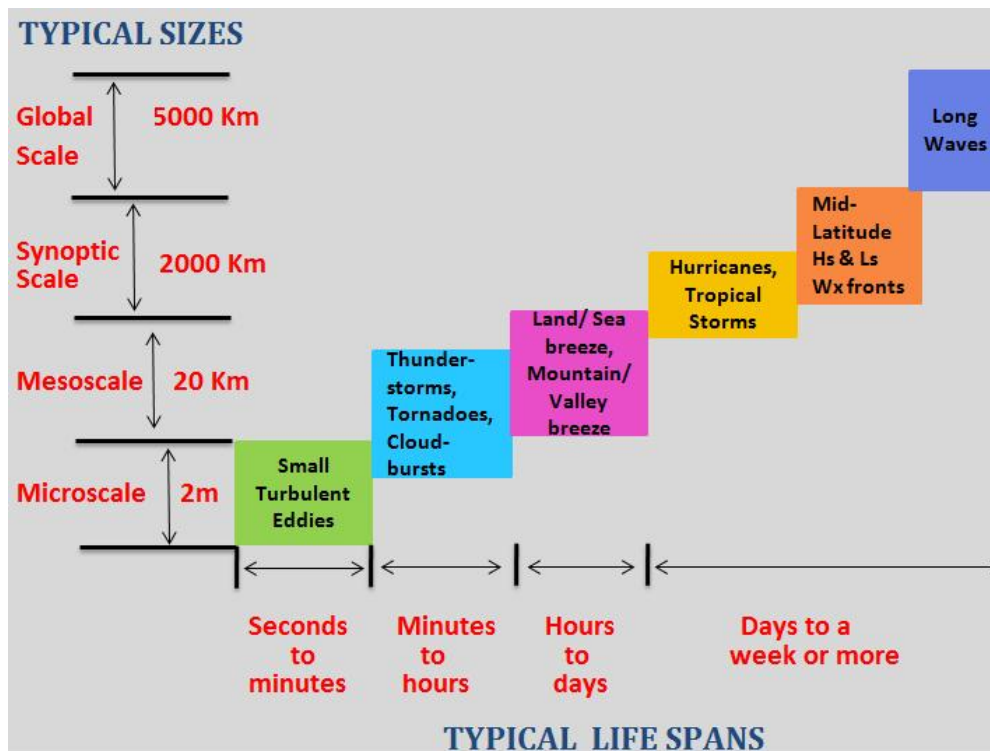


Figure 1.1 – Time and Space Scale of Atmospheric Motion (Source: Short, 2005)

1.2 Present state of forecast

The NWP models do not produce forecast of weather events directly. For interpretation, meteorological community uses MOS to interpret or quantify weather parameters from NWP models. In India, both at National Centre of Medium Range Weather Forecast (NCMRWF) and Indian Meteorological Department (IMD), one or the other form of MOS has been used. General experience is that MOS products show improved skills over the raw model output. Basis of MOS is statistical relationship which requires long term consistent series of NWP products. Since NWP models get upgraded regularly, the series does not remain consistent.

The ability of MOS to provide statistical corrections depends on the ability to identify significant model-observation correlations as suggested by Neiley and Hanson (2004). In order to do so, such correlations must be larger than those that may arise due to the random, stochastic

noise inherent in the forecast system. As NWP models have improved over the decades, both the systematic and stochastic errors in the model have been reduced. However, it is likely that the stochastic errors have not decreased as quickly as the systematic errors since many of the stochastic errors could be from observational errors, not related to the model. The assimilation systems are no doubt working towards getting the best out of the observations. It is believed (Neiley and Hanson,2004) that either (a) the dataset used to derive the MOS relationships must be longer in order to obtain forecasts with equal statistical characteristics or (b) the fractional improvement of the MOS forecasts over the raw NWP models data must increase. As stated earlier (Neiley and Hanson, 2004), the NWP models are being revised very frequently, using longer periods of data to develop regression solutions is not always warranted. Therefore, it seems reasonable to speculate that the fractional improvement of MOS forecasts over raw NWP data might indeed be decreasing.

1.3 Intelligent Systems

In view of above limitation of MOS, it was decided to explore other Intelligent techniques which are becoming available and have been used in other similar disciplines, though in Meteorology, application has been rare & that too in a research mode.

Before discussing these techniques, the type of domain one is venturing in the field of interpretation of NWP products is to be considered. Broadly there are two types of production systems, namely Monotonic & Non Monotonic (NM). By definition monotonic system is a system in which the application of a rule never prevents the later application of another rule that could have been applied at the time first rule was applied. The non-monotonic system is such where change of sequencing of rule totally changes the reasoning. In real life situations, especially in interpretation of weather, we need to apply NM reasoning.

In Artificial Intelligence (AI) one comes across the following tools based on the NM reasoning, which could enhance the capability of interpretation of NWP products:

- i) Data mining techniques
- ii) Fuzzy logic reasoning
- iii) Genetic Algorithm
- iv) Artificial Neural Network
- v) Bayesian probabilistic inferences

1.3.1 Data mining

Data mining, (Witten and Frank, 2005), is the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to analyze important information in data warehouses. Data Mining (DM) is an interdisciplinary field involving Databases, Statistics, High Performance Computing, Machine Learning, Visualization and Mathematics (Han and Kamber, 2006). Data Mining scours databases for hidden patterns, finding predictive information that experts may miss, as it goes beyond their expectations. There is an increasing desire to use this new technology in new domains of application and a growing perception that these large passive databases can be transformed into useful actionable information. Meteorology is one of such domains, where data mining is expected to improve the productivity of its analysts tremendously by transforming their voluminous, unmanageable and prone to ignorable information into usable pieces of knowledge. The survey of studies based on various DM techniques has been discussed in section 2.1.

1.3.2 Fuzzy Logic

Fuzzy logic emerged as a consequence of the proposal of fuzzy set theory by Zadeh (1965) for solving engineering control problems. This technique can be easily used to implement systems ranging from simple, small or even embedded up to large networked ones. It can be used to be implemented in either software or hardware. The key idea of fuzzy logic as mentioned in Stanford Encyclopedia on Fuzzy Logic (2010) is that it uses a simple and easy way in order to get the output(s) from the input(s), actually the outputs are related to the inputs using if-statements and this is the secret behind the easiness of this technique. The most fascinating thing about Fuzzy logic is that it accepts the uncertainties that are inherent in the realistic inputs and it deals with these uncertainties in such a way that their affect is negligible and thus resulting in a precise output. Fuzzy logic is said to be the control methodology that mimics how a person decides but only much faster. One of the many advantages of fuzzy logic is that it really simplifies complex systems. Fuzzy logic has been used to produce forecasts of airport cloud ceiling and visibility, seasonal runoff, formation of radiation fog and detection of bounded weak-echo region applications as discussed in section 2.2.

1.3.3 Genetic Algorithm

Genetic algorithms (GA) as defined by Sivanandam and Deepa (2007) are a biologically inspired technology, are randomized search and optimization techniques guided by the principles of evolution and natural genetics. GAs are executed iteratively on a set of coded

solutions, called population, with three basic operators: selection/reproduction, crossover, and mutation. They use only the payoff (objective function) information and probabilistic transition rules for moving to the next iteration. They are efficient, adaptive, and robust search processes, producing near optimal solutions, and have a large degree of implicit parallelism. Therefore, the application of GAs for solving certain problems of meteorology, which need optimization of computation requirements, and robust, fast and close approximate solutions, appears to be appropriate and natural. GAs have been used to classify rainy and non-rainy days, to find the position of tropical cyclones. These applications of integration of GAs and meteorology have been mentioned in section 2.3.

1.3.4 Artificial Neural Network

An Artificial Neural Network (ANN) is an information processing model that is able to capture and represent complex input-output relationships. The motivation of the development of the ANN technique came from a desire for an intelligent artificial system that could process information in the same way as the human brain (Jena et al. 2009). Its novel structure is represented as multiple layers of simple processing elements, operating in parallel to solve specific problems. ANNs resemble human brain in two respects: learning process and storing experiential knowledge. An artificial neural network learns and classifies a problem through repeated adjustments of the connecting weights between the elements. In other words, an ANN learns from examples and generalizes the learning beyond the examples supplied. ANNs have been used in prediction of weather variables like probability of precipitation, quantitative precipitation forecast and rainfall, as reviewed in section 2.4.

1.3.5 Bayesian Probabilistic Networks

Probabilistic networks are graphical models of (causal) interactions among a set of variables, where the variables are represented as nodes (also known as vertices) of a graph and the interactions (direct dependences) as directed links (also known as arcs and edges) between the nodes (Kjærulff and Madsen, 2005).

Any pair of unconnected/nonadjacent nodes of such a graph indicates (conditional) independence between the variables represented by these nodes under particular circumstances that can easily be read from the graph. Hence, probabilistic networks capture a set of (conditional) dependence and independence properties associated with the variables represented in the network. Bayesian networks contain only random variables, and the links represent direct

dependences (often, but not necessarily, causal relationships) among the variables. BNs have been used for certain meteorological applications, *viz.* damaging winds of high speed and large hail; and to model the spatial and temporal dependencies among the different stations for forecast of rainfall, in section 2.5.

1.4 Objective of the Research

It was noted that the Numerical Weather Prediction models do not produce forecast of actual weather events directly. Model Output Statistics has been used to interpret the weather systems from NWP models. MOS has limitations because its basis is the statistical relationship which requires long term consistent series of NWP products. Since NWP models get upgraded regularly, the series does not remain consistent. Most of the extreme weather events occur at the Sub-Grid scale and this area of research is still virgin, it was decided to evolve methods that can help interpret these disastrous Sub-Grid scale weather systems. In view of these, research work was taken up with the following objectives:-

- To develop alternative methods for the interpretation of forecasts produced by NWP models for Sub-Grid scale weather events.
- To assess that the weather systems form in favored zones.
- To apply some tools of data mining on output products of operational and research NWP models to discover patterns at sub-grid scale from large scale patterns.
- To identify and utilize an appropriate Database Management System for efficient storage and retrieval of multidimensional weather data in Indian context.
- To test other Intelligent systems techniques for their efficacy for forecast at sub-grid scale.

1.5 Thesis organization

The thesis is organized as follows. Chapter 2 is the literature review and problem definition. Chapter 3 describes the methodology used. Chapter 4 describes the case studies. Chapter 5 gives results and discussions. Chapter 6 concludes and chapter 7 finally describes the contributions, limitations and future scope.

Chapter 2

Literature Review and Problem Definition

This dissertation draws on two main aspects under consideration- data mining and sub-grid scale weather events. The related work in the field of data mining and other intelligent systems for weather forecasting is discussed in this chapter.

2.1 Data mining Techniques for Weather Forecast

The weather system can also be thought of as a complex system whose various components interact in various spatial and temporal scales and these states exhibit a great deal of correlations at various spatial and temporal scales, as suggested by Shekhar et al. (2004). Since the weather data are generally voluminous, they can be mined for occurrence of particular patterns that distinguish specific weather phenomenon.

There are many studies that support the applicability of Data mining techniques viz. **clustering, association rule mining and classification techniques** for weather forecasts. These are reviewed as follows:-

Based on **k-nearest neighbor approach**, Singh, Ganju and Singh (2005) have developed a quantitative snowfall forecast model for a station in Jammu and Kashmir, using surface meteorological data of the past 12 winters (1991–92 to 2003–04, excluding the data of winter 1994–95, which was not available). The model predicts weather in terms of snow/no snow day and the amount of snowfall (snowheight in cm) for three consecutive days in advance. The performance of the model has been tested for four winters for day-1, day-2 and day-3 forecasts. For qualitative snowfall forecast, the model performance for day-1, day-2 and day-3 forecasts turns out to be 80–90%, 70–80% and 65–75%. The model estimates the expected snowfall amount at the station for day-1, day-2 and day-3 in advance, and based on the value of the estimated snowfall amount, it is categorized in the expected snowfall range based on the already established criterion. Quantitatively, the model predicts snowfall amount accurately for day-1 and the average accuracy of the model for different ranges of established categories varies from 25 to 55% for day-1 forecast. The model over-predicts the expected snowfall amount for day-2 and day-3 compared to day-1.

Spatial clustering has also been used for the purpose of mining of Geographical data (Koperski, Han and Adhikary, 1998; Han et al. 2001). Clustering Large Applications based upon RANdomized Search (CLARANS) was proposed in Han et al. (2001) so as to improve the quality and scalability of Clustering LARge Applications (CLARA) that is also a **k-medoids** method of spatial clustering. Two spatial data mining algorithms were developed in an approach similar to CLARANS (Ng and Han, 1994). These are spatial-dominant algorithm and non-spatial-dominant algorithm. Both assume that the user specifies the type of rule to be mined and relevant data through a learning request in a similar way as in an experimental database mining system, DBMiner by Fayyad, Piatetsky-Shapiro and Smyth (1996). These algorithms use CLARANS for clustering and find high level non-spatial description of objects in every cluster using attribute oriented induction. The efficiency of clustering algorithms may be significantly improved by using spatial data structures. Also sampling methods help improve the efficiency of clustering algorithm.

The use of **clustering** (Dubes and Jain, 1988) is driven by the intuition that a climate phenomenon is expected to involve a significant region of the ocean or atmosphere, and that it is expected that such a phenomenon will be 'stronger' if it involves a region where the behavior is relatively uniform over the entire area. Shared Nearest Neighbor (SNN) clustering (Ert'oz, Steinbach and Kumar, 2003) has shown to find such homogeneous clusters. Each of these clusters can be characterized by a centroid, i.e., the mean of all the time series describing the ocean points in the cluster, and thus, these centroids represent potential climate indices. This approach offers a number of benefits: (i) discovered signals do not need to be orthogonal to each other, (ii) signals are more easily interpreted, (iii) weaker signals are more readily detected, and (iv) an efficient way is proved to determine the influence of a large set of points, e.g., all ocean points, on another large set of points, e.g., all land points.

The application of data mining to the discovery of interesting and useful earth science patterns has been illustrated by describing some of the results (Kumar et al. 2001; Kumar et al. 2004) in reference of the project *Discovery of changes from the Global Carbon Cycle and Climate System using Data Mining*. The multi-year output of the National Aeronautics and Space Administration-Carnegie Ames Stanford Approach (NASA-CASA) model for predicting Net Primary Production (NPP) and long term global sea surface temperature (SST) anomalies has been used to discover interesting patterns relating changes in NPP to land surface climatology and global climate. The data mining technique used is **SNN clustering** on SST data over the time period from 1958 to 1998. Most of the cluster centroids found is very highly correlated to well-known climate indices like El Nino indices, cold tongue index (CTI index). The correlation of these clusters to their corresponding indices are higher than 0.9. Cluster

centroids that have medium or low correlation with known indices may represent potentially new earth science phenomena (Kumar et al. 2001). Association analysis has been used to derive spatio-temporal relationship hidden in earth science data by Kumar et al. (2001).

In NASA-CASA model, NPP is a direct product of five input factors: the cloud-correlated solar irradiance, Fraction of Photosynthetically Active Radiation (FPAR), maximum light use efficiency, temperature and moisture stress scalars. Two **association rules** extracted by Apriori algorithm are consistent with the predictions made by the NASA-CASA model.

Agee and Zurn-Birkhimer (1998) determined that an axis of increased tornado activity during La Niña years extended from Iowa through Illinois and Indiana into Kentucky and Tennessee, while an axis of increased tornado activity during El Niño years extended from Colorado and New Mexico through the Texas panhandle into Oklahoma and Missouri. They concluded that their findings were a result of geographical shifts in tornado activity, rather than an overall increase or decrease in activity nationwide based on El Niño-Southern Oscillation (ENSO) phase. Bove (1998) found a similar axis of increased activity in La Niña years.

Numerical computations of the local change with respect to time of the local tendency of the vertical component of the vorticity of the surface wind and a stability index were made by Foster (1964) at 3 hour intervals for fifteen tornado days during the period from February to June 1961. Average values of these two parameters were computed with reference to tornado occurrences at the center of the computation grid. These average values are shown for five time periods: 9 to 12 hr., 6 to 9 hr., 3 to 6 hr., 0 to 3 hr. prior to tornado occurrences, and 0 to 3 hr. after tornado occurrences. It has been observed that the tornadoes develop within an area under the influence of increasing cyclonic vorticity tendency and characterized by a conditionally unstable air mass. These conditions apparently exist during a short time interval before tornado occurrence with tornadoes developing only after their combined intensity reaches a critical value.

Meteorological environment of a tornado outbreak in Southern Romania has been studied by Oprea and Bell, 2009. Three tornadoes were reported in Southern Romania on 7May 2005 that caused severe damage. The synoptic and mesoscale conditions associated with these tornadoes were analyzed using ECMWF and ALADIN (Aire Limitée et Adaptation Dynamique) model outputs. The analysis results of ECMWF model confirmed that the vertical profile of the wind showed a veering of the wind at low levels. Also the vertical profiles of the wind obtained from ALADIN model depicted strong vertical shear at low and mid levels.

The traditional technique to detect and anticipate tornadoes has been with the help of radars and each radar system requires the development of its own unique detection/ prediction

algorithms. The lead time is just a few minutes to one or two hours. Furthermore it can predict in terms of high or medium probability only.

Anticipating the formation of tornadoes through data mining has been done by McGovern et al. (2007) by using output products from a NWP model Advanced Regional Prediction System (ARPS), which is a three-dimensional, non-hydrostatic model. It is a very high resolution model specially developed for study of meso scale weather systems including tornadoes. It is interesting to note that the study under reference is conducted by six authors, three from computing school & three from meteorology, one of the meteorologists from US weather service is associated with tornadoes warning centre. A number of fundamental and derived meteorological quantities have been extracted that enable the observation of tornadoes. Vorticity and divergence /vertical motion fields show promise. The authors used a technique of sequential pattern mining as explained by Han and Kamber (2001). The concluding remarks by the authors are interesting and are quoted below;

“Given the small number of simulations, the results reported in the paper are only the beginning of what will a very exciting collaboration between meteorology & computer sciences.....”

During the same period or even earlier author, Pabreja (2005) had presented the concept paper during an international conference at Indian national centre for medium range forecasting (NCMRWF) where a large number of international meteorologists were present. The concept according to the meteorological community was novel and need trial. That marked the beginning of my work & support by Indian Meteorology community.

Spatio-temporal data mining of large geophysical datasets has been experimented by design, implementation and application of CONtent-based Querying in Space and Time (CONQUEST) (Stolorz and Nakamura, 1995) that has been built under the auspices of National Aeronautics and Space Administration (NASA) High Performance Computing and Communications (HPCC) program. CONQUEST supplies a knowledge discovery environment which allows geophysical scientists to easily formulate queries of interest, especially the generation of content-based indices dependent on specified and emergent spatio-temporal patterns. CONQUEST executes these queries rapidly on massive datasets and visualizes the results. The system architecture consists of five basic components namely scientist workbench, query manager, visualization manager, query execution engine and information repository. Application of the system has been demonstrated for the detection of blocking features and cyclone.

Spatial data mining technique based on the **aggregation techniques** and **spatial relations** has been used by Huang and Zhaob (2000) that emulates the way human experts perceive structures in large datasets in order for machines to “see” the same structures. It uses existing domain knowledge, expressed as parameters at various levels of aggregation, to aid the perception and understanding of spatial datasets. This approach has been applied to finding pressure trough features in weather data sets.

The trajectories of Mesoscale Convective system (MCS) over Tibetan Plateau in China, are automatically tracked using Geostationary Meteorological Satellite (GMS) brightness temperature and High Resolution Limited Area Analysis And Forecasting System (HLAFS) data provided by China National Satellite Meteorological Centre from June to August 1998 by Guo, Dai and Lin (2004). Based on these, the relationship between the trajectories of MCSs moving out of the plateau and their environmental physical field values are analyzed by authors using **spatial association rule mining** technique. The results indicated that at the level of 400hPa, the trajectories of MCSs, which move out of the plateau, are mainly influenced by geopotential height, relative humidity, vorticity, divergence and vertical wind speed, while at the level of 500hPa, geopotential height, relative humidity, temperature, vertical wind speed and K index are the main factors which influence the MCS to move out of the plateau.

There are many applications of Knowledge Discovery from Meteorological Databases mentioned by Roiger and Geatz (2003). Tsagalidis and Evangelidis (2010) have used meteorological data from the ECMWF 40 Year Re-analysis Project (ERA-40) and the weather observations data from the Meteorological Station of Mikra (Thessaloniki, Greece) as input to five data mining algorithms with the aim to build **classification models** for the prediction of the occurrence of precipitation in the station. An attempt is being made to find the effect the selection of the training set has on the performance of the algorithms and to determine the minimum training set size that can ensure effective application of the data mining techniques. Statistical Package for the Social Sciences (SPSS) software package is used to process the entire ERA-40 dataset and to produce a new one that consisted of a reduced number of uncorrelated variables. The training datasets were the input to five data mining algorithms, namely, the Decision tree C4.5, the k-Nearest Neighbor, the Multi-layer Perceptron with back-propagation, the Naïve Bayesian and the Repeated Incremental Pruning to Produce Error Reduction (RIPPER). Tsagalidis and Evangelidis (2010) have used the Area under the ROC Curve (AUC) as evaluation metric. The AUC measures the performance of the algorithms as a single scalar. Receiver Operating Characteristic (ROC) graphs are two-dimensional graphs in which the True Positive Rate (the percentage of positive cases correctly classified as belonging to the positive class) is plotted on the Y axis and the False Positive Rate (the percentage of negative cases is

classified as belonging to the positive class) is plotted on the X axis. The AUC is a reliable measure to get a score for the general performance of a classifier. It was observed that the performance on AUC of the decision tree C4.5, k-Nearest Neighbor, and Multilayer Perceptron with backpropagation neural network algorithms does not increase significantly for training set sizes more than 9 years. The performance on AUC of the Naïve Bayesian and RIPPER algorithms is independent from the training set size.

In United States, mesoscale weather events viz. floods, tornadoes, hail, strong winds, lightning, and severe storms cause huge losses and in order to mitigate the impacts of such severe events innovative new software tools and cyber infrastructure through which scientists can monitor data for specific weather events and launch focused modeling computations for prediction and forecasts of these evolving weather events, is being proposed. The model proposed (Li et al., 2008) is based on sophisticated data mining algorithms that apply **classification techniques** to the detection of severe storm patterns. The ideas presented in this paper have been implemented in the Linked Environments for Atmospheric Discovery: A Cyber infrastructure for Mesoscale Meteorology Research And Education (LEAD-CI) prototype (Droegemeier et al., 2004).

The implementation of data mining system for the analysis of operational and reanalysis databases of atmospheric circulation patterns has been done by Cofiño (2003) in the CrossGrid project. The designing of adaptive schemes for distribution of data and computational load according to changing resources available for each grid job submitted has also been done by the author. Self organizing maps and smoothing filters are used in combination with databases of observations to provide downscaled local forecasts from operative model outputs. The ECMWF and National Centers for Environmental Prediction (NCEP)/ National Center for Atmospheric Research (NCAR) reanalysis databases, the mesoscale data archive generated by the operational Unified Model for PoLand area (UMPL) system and local databases with observations from Spanish and Polish meteorological stations posts and radar sites are also being used. Self-Organized maps have been recently applied in several meteorological problems, such as **classifying** climate modes and anomalies in the area of Balkans in Cavazos (2000).

There is another work (Onwubolu et al. 2007) on self-organizing data mining approach called as enhanced Group Method of Data Handling (e-GMDH). In this research, weather data used is daily temperature, daily pressure and monthly rainfall and this data is acquired at the School of Engineering and Physics, University of the South Pacific, Fiji for the city of Suva. Group Method of Data Handling (GMDH) was innovated by Ivakhnenko (1971). GMDH combines the best of both **Statistics and Neural Networks** features while considering a very

important additional principle of induction (Lemke and Müller, 2003; Müller, 2003). This cybernetic principle enables GMDH to perform not only in advanced model parameter estimation but, to perform an automatic model structure synthesis and model validation, too. GMDH creates adaptively models from data in form of networks of optimized transfer functions (active neurons) in a repetitive generation of populations (layers or generations) of alternative models of growing complexity and corresponding model validation and fitness selection until an optimal complex model which is not too simple and not too complex (over-fitted) has been created. Neither, the number of neurons and the number of layers in the network, nor the actual behavior of each created neuron (transfer function of active neuron) are predefined. All these are adjusted during the process of self-organization by the process itself. As a result, an explicit analytical model representing relevant relationships between input and output variables is available immediately after modeling. The weather data being input to the model include daily temperature and pressure observed from 2000-2007 using automated instruments and a chaotic rainfall data set observed for the period of 1990-2002, for the city of Suva.

The **self-organizing GMDH** technique was used for the purpose of extraction and discovery of knowledge of the data acquired. For performance evaluation, network predicted outputs were compared with the actual temperature value. The absolute difference error for the daily temperature is found to be within the range of $\pm 1.5^{\circ}\text{C}$ and the absolute difference error for the daily pressure is found to be within the range of $\pm 0.3\text{bar}$. These two parameters have shown an excellent match between measured and predicted values. For monthly rainfall, the absolute difference error is found to be within the range of $\pm 0.2\text{mm}$. Here, there is not a good match between the measured and predicted values.

The weather variables can be viewed as sources of spatio-temporal signals. The information from these spatio-temporal signals can be extracted using data mining techniques. The variation in the weather variables can be viewed as a mixture of several independently occurring spatio-temporal signals with different strengths. Independent component analysis (ICA) has been widely studied in the domain of signal and image processing where each signal is viewed as a mixture of several independently occurring source signals. Under the assumption of non-Gaussian mixtures, it is possible to extract the independently occurring signals from the mixtures under certain well known constraints (Jayanta, 2004) and (Wang, Li and Li, 2006). Therefore, if the assumption of independent stable activity in the weather variables holds true then it is also possible to extract them using the same technique of ICA.

North Atlantic Oscillation (NAO) has been taken as a typical example and the sea level pressures (SLP) has been mined using **independent component analysis** by Jayanta (2004).

NAO is considered as specific large scale atmospheric feature and is supposed to provide important signal for future weather events. Using the data mining techniques for determining the strongest independent components in the multidimensional data set, it has been observed that the strongest stable patterns as obtained by ICA matched with the physical patterns of oscillation in SLP. The results are also verified by Jayanta (2004) by finding a linear fit of the independent components with the standard NAO index as provided by the meteorological measurements.

Earth Scientists have been using eigenvalue analysis techniques, such as **principal components analysis and singular value decomposition**, to discover climate indices (Storch and Zwiers, 1999). While eigenvalue techniques do provide a way to quickly and automatically detect patterns in large amounts of data, they also have the following limitations (i) all discovered signals must be orthogonal to each other, making it difficult to attach a physical interpretation to them, and (ii) weaker signals may be masked by stronger signals. An alternative clustering based methodology for the discovery of climate indices that overcomes these limitations has been developed (Ert'oz, Steinbach and Kumar, 2001), (Ert'oz, Steinbach and Kumar, 2002) and (Ert'oz, Steinbach and Kumar, 2003).

From this review, it is observed that weather systems involve a significant region of atmosphere where the behavior of certain weather elements is relatively uniform over the entire area. Hence, we believe that the DM tools using clustering technique could form a tool for our study of Interpretation of sub-grid scale weather systems from NWP output products.

2.2 Fuzzy Logic Techniques for Weather Forecast

A system named Weather Prediction system (WIND-1) has been developed (Riordan and Hansen, 2002) that combines fuzzy logic and case-based reasoning to produce forecasts of airport cloud ceiling and visibility. Knowledge about meteorological and temporal features that experienced forecasters use to construct analogous climatological scenarios is encoded in a fuzzy similarity measure. WIND-1 consists of a large database of weather observations from an archive of 315,576 consecutive hourly airport weather observations made at Halifax International Airport during the 36-year period from 1961 to 1996 and a **fuzzy k- nearest neighbors algorithm**. The system has been tested with twelve similarity indicating attributes – date, hour, cloud amount, cloud ceiling height, visibility, wind direction, wind speed, precipitation type, precipitation intensity, dew point temperature, dry bulb temperature and pressure trend. It has been observed that prediction accuracy increases as the number of attributes used for comparison increases. Also by varying the value of k ($k = 1,2,4,8,16,\dots,256$),

the best analog ensemble has been found with $k=16$. Such a fuzzy k -nearest neighbors weather prediction system can improve the technique of persistence climatology by achieving direct, efficient, analogous comparison of past and present weather cases.

Fuzzy k-means clustering has also been validated by Liu and George (2005, 2006) on weather data in the South Central U.S. and global climate data. The algorithm is able to identify and preserve interesting phenomena in the weather data.

Fuzzy logic has been used to forecast seasonal runoff in two basins in Alberta, Canada (Mahabir, Hicks and Robinson, 2003). The potential snowmelt runoff is forecast each spring, for several basins to assess the water supply situation. Water managers need this forecast to plan water allocations for the following summer season. The Lodge Creek and Middle Creek basins, located in southeastern Alberta, are two basins that require this type of late winter forecast of potential spring runoff. The applicability of fuzzy logic modeling techniques for forecasting water supply has been investigated by the authors. Fuzzy variables were used to organize knowledge that is expressed 'linguistically' into a formal analysis. For example, 'high snowpack', 'average snowpack' and 'low snowpack' became variables. By applying fuzzy logic, a water supply forecast was created that classified potential runoff into three forecast zones: 'low', 'average' and 'high'. Spring runoff forecasts from the fuzzy expert systems were found to be considerably more reliable than the regression models in forecasting the appropriate runoff zone, especially in terms of identifying low or average runoff years. Based on the modeling results in these two basins, it is indicated that fuzzy logic modeling may be very practical for forecasting water supply conditions for small basins with limited data.

Fuzzy logic technique has been effectively applied in the field of operational meteorology by Murtha (Murtha, 1995). The forecast of the probability of formation of radiation fog has been discussed and examples of the fuzzy logic methods are presented by Murtha. Radiation fog is one of the physical processes not yet well understood or beyond the resolution of the numerical models and needs alternative methods for its analysis and subsequent prognosis. Within the context of fuzzy logic system, system inputs are those physical variables that are thought to completely determine the solution(s) to the problem, or system outputs. In the current example the system output is the probability of the occurrence of radiation fog within the next (approximately) six hours. System inputs to be used in this case are the values of dew point, dew point spread, the rate of change of the spread, the wind speed and the sky coverage. The results were promising and it was concluded that fuzzy logic has a promising potential for providing reliable forecast in operational meteorology as well as water supply forecasts.

A Fuzzy Rule-based method for the detection of a bounded weak-echo region (BWER) within a storm structure that can help in the prediction of severe weather phenomena is presented (Pal, et al., 2006). A BWER is a radar signature within a thunderstorm characterized by local minima in the radar reflectivity at low levels, which extends upward into and is surrounded by the higher reflectivities aloft. This feature is associated with a strong updraft and is almost always found in the inflow region of a thunderstorm (Markowski, 2002). In fact, a BWER is a representative of a local storm that develops in a strongly sheared environment and tends to a steady-state circulation.

The Fuzzy rule-based approach takes care of the various uncertainties associated with a radar image containing a BWER. This technique automatically finds some interpretable (fuzzy) rules for classification of radar data related to BWER. The radar images are preprocessed to find sub-regions (or segments) that are suspected candidates for BWERs. Each such segment is classified into one of three possible cases: strong BWER, marginal BWER, or no BWER. In this regard, spatial properties of the data are being explored. The method has been tested on a large volume of data that are different from the training set, and the performance is found to be very satisfactory. It is also demonstrated that an interpretation of the linguistic rules extracted by this system can provide important characteristics about the underlying process.

Fuzzy rules can be applied to the study of Interpretation of disastrous sub-grid scale weather systems from output products of NWP models, but we did not find any such reference and did not pursue it. But this technique could be further studied.

2.3 Genetic Algorithms for Weather Forecast

A generic model of weather systems, based on a **GA framework** for finding the tropical cyclones (TC), fronts, troughs and ridges from multidimensional numerical weather prediction data has been designed by Yan, Lap and Wah (2006). The weather systems have been approximated as adjoining segments of arcs. For fronts, the air masses on the two sides of the arc have contrasting pressure or dew point values. Troughs can be seen as lines of convergence of different wind directions. Pressure extrema can be seen as areas enclosed by closed arcs. For TC, the closed arc may represent the eye wall, which marks the boundary between strong wind and calm areas. The problem of identifying weather systems is thus reduced to maximizing the difference between some functions of weather elements on the two sides. The authors have found that their method not only can locate weather systems with 80% to 100% precision, but also discover features that could indicate the genesis or dissipation of such systems that could be ignored by forecasters. A number of studies (Wong and Yip, 2005a; Wong and Yip, 2005b;

Wong et al. 2004) have been done for automatic TC positioning and tracking using satellite or radar data, but literature on automating the process of identifying line-shaped weather systems such as fronts, troughs and ridges are rare.

The **GA** has also been applied for solving the optimum classification of rainy and non-rainy day occurrences based on vertical velocity, dew-point depression, temperature and humidity data at the Lake Van station in eastern Turkey by Sen and Oztopal (2001). The problem involves finding optimum classification based on known data, training the future prediction system and then making reliable predictions for rainfall occurrences. Various statistical approaches require restrictive assumptions such as stationarity, homogeneity and normal probability distribution of the hydrological variables concerned. The GAs do not require any of these assumptions in their applications. Also the amounts of precipitation are predicted with a model similar to a third order Markov model whose parameters are estimated by the GA technique. It has been shown that GAs give better results than classical approaches such as discriminant analysis. This literature review does not document any application of this technique in relation to sub-grid scale weather phenomenon, hence was not pursued in this research.

2.4 Artificial Neural Networks for Weather Forecast

A **neural network**, using input from the Eta Model and upper air soundings, has been developed by Hall, Brooks and Doswell (1999) for the probability of precipitation (PoP) and quantitative precipitation forecast (QPF) for the Dallas–Fort Worth, Texas, area. Forecasts from two years were verified against a network of 36 rain gauges. The resulting forecasts were remarkably sharp, with over 70% of the PoP forecasts being less than 5% or greater than 95%. Of the 436 days with forecasts of less than 5% PoP, no rain occurred on 435 days. Of the 111 days with forecasts of greater than 95% PoP, rain always occurred. The linear correlation between the forecast and observed precipitation amount was 0.95. Equitable threat scores for threshold precipitation amounts from 1 mm to 25 mm are 0.63 or higher, with maximum values over 0.86. Combining the PoP and QPF products indicates that for very high PoPs, the correlation between the QPF and observations is higher than for lower PoPs. In addition, 61 of the 70 observed rains of at least 12.7mm are associated with PoPs greater than 85%. As a result, the system indicates a potential for more accurate precipitation forecasting.

ANNs were designed to mimic the characteristics of the biological neurons in the human brain and nervous system as explained by Zurada (1992). ANN technology is based loosely upon the cellular structure of the human brain. Cell and connections between the cells are

established in the computer. As in the human brain, connections among the cells are strengthened or weakened based upon their ability to yield "productive" results. The system uses an algorithm to "learn" from experience (Koska, 2005).

The analysis of 87 years of rainfall data in Kerala state, the southern part of Indian Peninsula situated at latitude, longitude pairs ($8^{\circ}29$ N, $76^{\circ}57$ E) has been done (Abraham, Philip and Mahanti, 2004). The rainfall data has been standardized and divided into two groups as training set and test set. The data from 1893-1933 has been used for the purpose of training the ANN and after training, the ANN has been tested using data from 1934-1980 as test set. The authors have used a **neuro-fuzzy** system that is capable of adapting the architecture according to the problem after performing some initial experiments to decide the architecture of the neural network. Since rainfall has a yearly periodicity, the author started with a network having 12 input nodes. Further experimentation showed that it was not necessary to include information corresponding to the whole year, but 3-month information centered over the predicted month of the fifth year in each of the 4 previous years would give good generalization properties. The prediction models hence implemented are reliable and the Root Mean Square Error (RMSE) values on test data are very less (0.090). There have been few deviations of the predicted rainfall value from the actual.

ANN has also been used (Chattopadhyay and Chattopadhyay, 2007) to predict the average rainfall over India during summer- monsoon i.e the months of June, July, and August, by exploring the rainfall data corresponding to the summer monsoon months of years 1871-1999. Feed forward neural network, also known as Multilayer Perceptron (MLP), has been used as a predictive tool. First 75% of the available data have been used as training set and the remaining 25% are used as the test set. The multilayer perceptron model is framed with the data set divided into test and training cases. The test cases are arbitrarily chosen for the available dataset so as to maintain the generality. The model has been trained up to 50 epochs. The learning rate parameter is fixed at 0.4 and the momentum rate is chosen 0.9. The weight matrix is framed with the values between -0.5 and 0.5. Least mean squared error is chosen as the stopping criteria for the learning procedure. The performance of this neural net model is compared with conventional persistence forecast. It has been found that the prediction error in case of ANN is 10.2% whereas the prediction error in the case of persistence forecast is 18.3%. Therefore, Neural Net, in the form of Multilayer Perceptron is found to be smarter in the prediction of monsoon rainfall over India.

The literature review has demonstrated applications for forecasting of Rainfall/Precipitation using ANN. But it is not clear whether ANN could be used for getting advance warning of occurrence of a weather event.

2.5 Bayesian Networks for Weather Forecast

The North American Regional Reanalysis (NARR) provides historic weather data over North America with higher spatial and temporal resolution than the global reanalysis (Mesinger, 2002). This dataset has been used by Jaye (2006) to provide a much more accurate assessment of the atmospheric conditions associated with each severe weather event namely presence of a tornado, damaging winds of over fifty knots, and large hail of diameter 0.75" or greater. The author has used NARR data with the Storm Prediction Center's (SPC) storm data which is an archive of the location, type, intensity, and time of every official storm report, to create a climatology of atmospheric conditions associated with severe weather event. NARR and SPC data for the years 1979 through 2003 has been utilized. A selection of seven parameters from fifty possible parameters produced as model output has been made based on their relationship with severe weather, and their correlations with each other. These parameters are Convective Inhibition, 0-CMB wind shear, Lifted Index, W @ Level of Free Convection, 0-CMB storm relative helicity, Surface Streamwise vorticity, Upper Level convergence. The model procedure combines these seven parameters in a Bayesian Framework to devise a conditional probability of severe weather occurrence. This model pinpoints regions of high likelihood of severe weather similar to the SPC convective outlooks. The main differences appear to be a tendency to pinpoint a smaller region for extremely high probabilities, and, of course boundaries in the likelihood contour zones that present themselves much less smoothly.

Bayesian Networks (BNs) have also been used to model the spatial and temporal dependencies among the different stations using a directed acyclic graph (Antonio et al., 2002). This graph is learnt from the available databases and allows deriving a probabilistic model consistent with all the available information. A subset of 100 climatic stations in the North basin of the Iberian Peninsula during winter 1999 has been considered. The variable rainfall has been represented pictorially by a set of nodes; one node for each variable {for clarity of exposition, the set of nodes is denoted as (y_1, \dots, y_n) }. These nodes are connected by arrows, which represent a cause and effect relationship. That is, if there is an arrow from node y_i to node y_j , we say that y_i is the cause of y_j , or equivalently, y_j is the effect of y_i . The independencies from the graph have been translated to the probabilistic model in a sound form.

The practical problem of incomplete knowledge of topology of graph has been managed with the help of learning algorithms and once the model describing the relationships among the set of variables is in place, it can be used to answer queries when evidence becomes available.

Afterwards, the resulting model is combined with numerical atmospheric predictions which are given as evidence for the model. Efficient inference mechanisms provide the conditional distributions of the desired variables at a desired future time. The efficiency of the proposed methodology has been illustrated by obtaining precipitation forecasts for 100 stations in the North basin of the Iberian Peninsula during winter 1999.

Numerical Atmospheric Circulation Models (ACMs) which are daily integrated by different weather services on coarse-grained resolution grids covering wide geographical areas provide a description of several meteorological variables (temperature, humidity, geopotential, wind components, etc.) which define the predicted atmospheric pattern for a given forecast period. The spatial resolution of these models is currently constrained by both computational and physical considerations to scales of approximately 10 to 50 Km. However, meteorological phenomena such as rainfall, vary on much more local scales and therefore, ACMs do not provide a regional detailed description of such relevant phenomena. Due to this limitation, a number of different statistical and machine learning techniques have emerged in the last decade. These techniques mine the information contained in meteorological databases of historical observations to train specific forecast models (regression (Enke and Spekat, 1997); hidden Markov models (Bellone, Hughes and Guttorp, 2000); neural networks (Gardner and Dorling, 1998), etc.). The resulting models predict future outcomes of a given variable based on the past evidence collected in the database. There have also been some attempts for combining both database information and ACMs. This is done by combining the model's predicted patterns with the information available in databases of observations (e.g., rainfall) and predictions (gridded atmospheric patterns). Therefore, **sub-grid detail** in the prediction is gained by post processing the outputs of ACMs using knowledge extracted from the databases (downscaling methods).

The above is a summary of various Artificial Intelligence (AI) techniques used just independently or in combination for forecasting weather elements by researchers. There are also some studies that are based on combination of two or more artificial intelligent techniques.

2.6 Combination of two or more AI techniques for Weather Forecast

An optimal meteorological prediction model based on support vector Regression with genetic algorithms (SVRG) has been developed Xue et al. (2009). GAs are used to determine free parameters of SVR then the proposed SVRG model is applied to forecast the temperature. The experimental results have indicated that SVRG model proposed by the authors, overcomes some shortcomings of the traditional SVR, and can achieve better forecasting accuracy and performance than traditional SVR and Back Propagation Neural Network (BPNN) prediction models. Under the same condition, the experimental results indicate that the method can not only solve the problem that the traditional SVR parameters are commonly determined by users' experience, but also provide better promising prediction results, forecasting accuracy and stability than those of SVR and BP neural network.

The genetic algorithm neural network (GANN) has been used (Lin, Lin and Chen, 2008) to forecast short range precipitation in Guangxi. In order to establish the forecasting model of the GANN, the genetic algorithm has been used to optimize the connection weight and structure of the neural network through application of retaining the best individual in the genetic evolution process. This method overcomes the randomness of the initial weight value, and, avoids the network oscillation as well as its being trapped into the local solution in the determination of the NN structure. The processing of dimensions cut on the massive forecast factors selected from the T213 numerical weather forecast products has been conducted. This GANN technique uses GA to optimize NN connection weight and structure simultaneously, and explores the novel way of interpretation for Numerical Weather Forecast (NWF). The method of interpretation of the NWF products bases on GANN has the learning capability, and the rain forecast accuracy is better than T213 NWF and Japan fine-mesh NWF model, which displays a good prospect for operational precipitation forecasts. This GANN method has achieved the effect of the concentration of the effective information, and established the forecasting model of daily rainfall grades during May-June in Guangxi by combining Japanese fine-mesh NWF predictor. The result are very convincing and have indicated that the accurate rate of 24 hours forecast for the average precipitation equal to or even greater than 10mm within Guangxi with this model is 0.57, 0.50, 0.30, for north, southeast and southwest subdivision respectively, which is 7-15% over the present routine operational T213 and Japanese numerical weather forecast products.

In the proposed model structure by Liu and Lee (1999), a neural-based model is used for multi-station weather prediction. The model includes 1) Genetic Algorithm (GA) for input node selection and network parameter settings; 2) Fuzzy classification for the rainfall parameters; 3) Neural network training using Backpropagation Network (BPN). A time series 6-hourly meteorological data such as wet bulb and dry bulb temperatures, rainfall, mean sea level pressure (MSLP), relative humidity, wind direction and speed are extracted from 11 weather stations in Hong Kong for the period from January 1993 to December 1997. This provides more than 7300 records of data and the basis for rainfall forecast. The author has observed that multiple-station model is better than single-station model. For rainfall prediction, multiple station model has a significant improvement over that of the single-station one (three-folded increase on correlation). To improve the rainfall forecast, two types of techniques were suggested. First, this is by manipulating the data which affect the data distribution. Among the methods, the one which normalizes the rainfall nonlinearly and classifies the rainfall into classes got most significant improvement. Secondly, incorporation of other algorithms like fastpropagation, to the neural network provided positive improvement.

This literature review does not document any application in relation to sub-grid scale weather phenomenon, therefore, was not pursued in this research.

2.7 Problem Definition and Objective of the study

Traditionally, weather forecasting is based mainly on numerical models (Lynch, 2007). This approach attempts to model the fluid and thermal dynamic systems for grid point time series prediction based on meteorological data at boundary. The simulation often requires intensive computations involving complex differential equations and computational algorithms. Besides, the accuracy is bound by certain “inherited” constraints such as the adoption of incomplete boundary conditions, model assumptions and numerical instabilities. This kind of approach is more appropriate for long-term (over 24 hours) forecasting over a large area of several thousand kilometers (Chow and Cho, 1997).

NWP output consists of mainly patterns of wind, temperature and pressure at various horizontal and vertical scales. For the forecast of weather elements like rain/snow, sky conditions, etc. at a place, meteorological community uses MOS to interpret or quantify weather parameters from NWP models output which is depicting large scale features and tries to associate these large scale systems with small scale weather events from certain analogous

situations based on their experience and thereby infer the idiosyncratic behavior of local (small) scale weather events like tornadoes, severe storms, thunderstorms, cloud burst, and intense rain.

MOS products show improved skills over the raw model output (Neiley and Hanson, 2004). But with the Skill of NWP products improving and the NWP models being continuously upgraded, it is hard to get long time consistent data from NWP models for developing MOS. The ability of MOS to provide statistical corrections depends on the ability to identify significant model-observation correlations. The authors have mentioned that in order to do so, such correlations must be larger than those that may arise due to the random, stochastic noise inherent in the forecast system. As stated earlier, the NWP models are being revised very frequently, using longer periods of data to develop regression solutions is not always warranted. Therefore, it seems reasonable to speculate that the fractional improvement of MOS forecasts over raw NWP data might indeed be decreasing.

As explained above, MOS is not a theoretically stable process for weather forecast and hence there is limitation of MOS technique for use of sub-grid scale weather phenomenon (cloudburst, tornado and rainfall). The literature survey definitely shows that the alternative methods for MOS are necessary (Neiley and Hanson, 2004). In the various studies, one finds the use of Intelligent systems for predicting the events of non-monotonic nature including weather. Data mining techniques are becoming widely available and have been used in many disciplines but in Meteorology, application has been rare and that too in a research mode. Different Data mining techniques for forecast of weather at a larger space and time scale or for warning only a few minutes before the occurrence of disastrous weather events have been applied.

Association rule mining has been used by McGovern et al., 2007 to detect tornado 6300 seconds in advance using simulated storm data produced from the ARPS. The relationship between the trajectories of MCSs moving out of the plateau and their environmental physical field values have been analyzed using **spatial association rule mining** technique. Classification techniques have been applied to predict occurrence of precipitation and to detect severe storm patterns.

SNN clustering has been used to find homogeneous clusters whose centroids represent potential climate indices like El Nino indices and cold tongue index. **K-nearest neighbor** technique has been used to develop a quantitative snowfall forecast model for a station in Jammu and Kashmir.

But there is not any application for interpretation of NWP output products for sub-grid scale weather systems like cloudburst and tornado. The reason could be non-availability of data

in a systematic manner. Considering that the area is still virgin, attempt was made to take up the following:-

1. Use of Data Mining techniques specifically clustering, for building up the association of the prediction of various derived variables related to specific weather event.
2. Since data is colossal and complex for processing, it was decided to first develop pre-processing techniques to interpret the meteorological data files and then to restructure data to facilitate faster storage and analysis. A Multidimensional Data Model was developed and used.
3. The data of sub-grid scale events which happened in the last few years in India was captured. Corresponding NWP output from various models was collected and pre-processed. Clustering technique of Data Mining appealed because of the reason that a weather phenomenon is expected to form in clusters in a significant region of the ocean or atmosphere as explained in literature review, so its efficacy for detection of useful signals for foreshadowing the occurrence of Sub-Grid scale events (Cloudbursts and tornadoes) was studied for improved prediction.

Data mining technique was used first for synoptic scale and then for sub-grid scale. With this approach Data mining could become an improved complementary approach to the NWP models which predict weather systems at large scale quite efficiently but have limitations at sub-grid scales.

Chapter 3

Methodology

This chapter explains the approach used in the study for pre-processing of various NWP output products and other meteorological datasets and their organization based on Multidimensional Data Models developed in the study for efficient retrieval. Various scales of weather systems are explained and approach for their interpretation in terms of weather elements using Data Mining tools are discussed.

3.1 Approach

From the literature review it has been observed that various data mining techniques have been applied to forecast weather viz. association rules, clustering, classification methods, fuzzy k-means clustering, pattern mining, etc. **Clustering techniques** have been utilized for associating patterns of weather elements with certain weather systems. These techniques have convincing results as far as forecast at synoptic scale is concerned.

As discussed in literature review (Dubes and Jain, 1988), the use of **clustering** is driven by the intuition that a weather phenomenon is expected to involve a significant region of the ocean, land and atmosphere. Chances of their occurrence will be ‘stronger’ in a region where the environment is favorable.

Shared Nearest Neighbor (SNN) clustering has been used to discover interesting patterns relating changes in NPP to land surface climatology and global climate. Most of the cluster centroids found is very highly correlated to well-known climate indices like El Nino indices, cold tongue index (CTI index).

The traditional technique to detect and anticipate tornadoes has been with the help of radars and each radar system requires the development of its own unique detection/ prediction algorithms. The lead time between the actual occurrence and detection in radar is few minutes whereas theoretically the presence of salient and favorable atmospheric patterns could be used to anticipate tornado formation even days in advance. There are discrete areas of significant updrafts and track around each updraft. It has been observed that the tornadoes and cloudbursts occur within an area under the influence of increasing cyclonic vorticity tendency and

significant updraft (upward vertical motion) and downdraft pattern. To optimize the study of these patterns, an ensemble technique discussed at section 4.6.1 has been utilized.

Since, clustering has been applied in most of the studies to look for homogeneous regions of weather event formation that contribute towards a particular weather phenomenon. Further literature review shows no study that applies Data mining techniques on output products of NWP model for interpretation of sub-grid scale weather systems. Based on these considerations and knowledge gained through discussions with the Meteorological experts, it was decided to study generation of clusters using output products of NWP model so as to look for zones of high vertical motion and hence to derive sub-grid scale phenomenon from the features that depict grid-scale forecasts. It is expected that with Data Mining approach, the predictions at sub-grid scale can be derived from the large scale feature.

3.1.1 Evolution of Working Methodology

Following factors helped in evolving the methodology used:-

- i) Data pre-processing posed a big challenge as the Meteorological data is huge and multidimensional in nature. The processing of approx. 60million data values for just one case of tornado / cloudburst used to take a lot of time, so Multidimensional Data Model was selected for storage in an organized form across time and space scales.
- ii) Discussions with Meteorological experts motivated working first on synoptical Low Pressure Systems (LPS) to analyze clusters of formation and dissipation of LPS movement.
- iii) Application of k-means clustering technique on LPS datasets produced convincing results (Pabreja, 2009) which motivated the author to study the sub-grid scale phenomenon viz. cloudburst and tornado using Data Mining techniques. Also noting that there is practically no study of interpretation of sub-grid scale weather events from NWP models outputs, the Data Mining clustering approach appealed.
- iv) Artificial Neural Networks were useful for forecasting of weather parameters like Rainfall at mesoscale but there were no studies on interpretation of Sub-Grid scale weather events based on ANN. This technique was not pursued.

As stated above, to validate the above hypothesis, Data Mining tools were tested first at the synoptic scale weather systems, before working on the sub-grid scale. This could help to

establish a basis for detection of patterns which are present for well-defined synoptic scale weather system. The details on this are explained in section 3.5 and the case studies using clustering technique in section 4.1 and section 4.2.

Artificial neural networks have been used so as to forecast rainfall after being trained over rainfall of previous years. ANN has shown good forecasting performance in many other weather related forecasts (Hall, 1998; Hayati and Mohebi, 2007; Chattopadhyay, 2007; Paras et al., 2007; Collins and Tissot, 2008). In this thesis a case study has been presented which can later on be extended to forecast the various ingredients that form sub-grid scale weather systems. The datasets used for the study are daily rainfall datasets in .grd format. The pre-processing steps applied on these datasets are explained in section 3.2.1 and the case study in section 4.3.

A typical output of NWP model has many weather variables at every grid point (w.r.t latitude, longitude and atmospheric pressure level) at each time of forecast. These files generated by the models are in .grd (gridded), .bin (binary), .netcdf (network Common Data Form) and .grib (GRIdded Binary) format. These files have been pre-processed so that the numeric values are made available for a data mining tool, as explained in section 3.2.2.

For the storage of these organized weather variables (both direct and derived), a flat 2-D Relational Database system was not efficient because of the multidimensional nature of these datasets. Hence, the need of multidimensional data model was felt that can facilitate the storage and retrieval of required weather variables across selected location and time dimensions in an efficient manner. The limitation of flat Database Management System and the need and advantage of storage of NWP model output products using Data cubes of Multidimensional Data model has been explained in section 3.3 and 3.4. The implementation of various data cubes used for the study is discussed in section 4.4 to section 4.6. The concepts to explain tornado and its formation are explained in section 3.6, case studies to interpret the output products of ECMWF T-799 model and WRF V3.1 model using clustering technique, for a real-life case of tornado is explained in section 4.7 respectively. The concepts based on cloudburst is explained in section 3.7 and application of clustering technique on the forecast made by ECMWF T-799 model for four real life cases of cloudbursts have been done in section 4.8.

3.2 Pre-processing of Meteorological Datasets

3.2.1 Pre-processing of Rainfall datasets

A high resolution ($0.5^\circ \times 0.5^\circ$) daily rainfall (in mm) dataset, for mesoscale meteorological studies over the Indian region, has been provided by IMD (described by

Rajeevan and Bhate (2009)). The dataset is in .grd format, a control file describing the structure of .grd file is explained in Appendix A. There is one .grd file for each year of rainfall.

This dataset consists of daily rainfall data for each year for the period 1984–2003. The data is for the geographical region from longitude 66.5 °E to 100.5 °E and latitude 6.5 °N to 38.5 °N for each day of the year. There are 4485 grid points readings every day and rainfall record for 122 days (June to September) per year are selected for analysis i.e. 5,47,170 records out of a total of 16,37,025 records for one year of rainfall. Steps followed for pre-processing of the .grd so that an intelligent system can be applied, are mentioned below:

1. The .grd file has been converted to .dat file using a FORTRAN programme (Appendix B). This dataset is very huge in size.
2. The .txt files have been exported to Excel worksheet and then to Access database. The data looks like as if a rectangular grid is filled with values of rainfall in mm (a sample of year 1989 rainfall is shown in table 3.1).

		Longitude (°E)									
		75	75.5	76	76.5	77	77.5	78	78.5	79	79.5
Latitude (°N)	38.5	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999
	38	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999
	37.5	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999
	37	-999	-999	-999	-999	0	0	0	-999	-999	-999
	36.5	-999	-999	-999	-999	4.4	1.2	0	-999	-999	-999
	36	-999	-999	-999	9.8	8	1.5	0	1	-999	-999
	35.5	-999	-999	-999	6.4	6.5	1	7.3	2.8	-999	-999
	35	-999	-999	7.7	7.4	24.2	11.7	17.3	24.2	2	-999
	34.5	-999	-999	7.1	10.7	6.2	7.1	27.8	16.2	10.6	5.8
	34	-999	-999	0.4	16	0.3	1.8	0.8	7.7	7.9	0
	33.5	-999	45.5	0	23.5	37.3	5.2	13.6	1.8	0	0
	33	10.6	0.3	15	27.9	17	0	0	0	4.7	1.1
	32.5	0	0	1.8	31.2	8.3	5.9	0.3	7.7	3.6	3.7
	32	0	14.4	24.9	9.3	0	0	0.4	0	2.7	0
	31.5	0	1.6	0.2	3.4	0	1.5	0	0	2.2	0
	31	0	0	6	1.7	0	0	0	0	5	1.6
	30.5	2.4	2.2	4.2	0	0	0	1.6	0	0	0
30	0	26.1	7.1	0	3.2	4.5	0	0	0	0	
29.5	13.6	5.4	11.4	2.1	0	4.8	2.3	0	0	0	
29	6	5.9	1.4	0.2	0	0	1.8	0	0	0	
28.5	11	13.1	0	0	0	0	0	10.7	0	0	

Table 3.1 – text file retrieved from .grd file for rainfall in 1989
(Source: as a result of pre-processing rf1989.grd provided by IMD)

3. Using Visual Basic program (as explained in Appendix C) to organize data in tabular format, as shown in table 3.2.
4. Finally exporting the dataset into .xls format for analysis, by Matlab.

Table1989					
S.No.	Day#	Date	latitude	longitude	Rainfall
404	1	01-Jun-89	34.5	76	7.1
405	1	01-Jun-89	34.5	76.5	10.7
406	1	01-Jun-89	34.5	77	6.2
407	1	01-Jun-89	34.5	77.5	7.1
408	1	01-Jun-89	34.5	78	27.8
409	1	01-Jun-89	34.5	78.5	16.2
410	1	01-Jun-89	34.5	79	10.6
411	1	01-Jun-89	34.5	79.5	5.8

Table 3.2 – Rainfall for year 1989 organized in tabular format
(Source: as a result of pre-processing rf1989.grd provided by IMD)

3.2.2 Pre-processing of NWP model output products

The output products of two NWP models viz. ECMWF T-799 model and WRF V3.1 have been used for the study. The pre-processing done on ECMWF model forecasts and WRF model forecasts are explained in the sections that follow.

3.2.2.1 Pre-processing of ECMWF T-799 model forecasts

The datasets produced as forecast by ECMWF model are in GRIB format which is a mathematically concise data format commonly used in meteorology to store historical and forecast weather data. It is standardized by the World Meteorological Organization's Commission for Basic Systems. The forecast datasets of T-799 model include values for 87 variables (including all atmospheric pressure levels), for latitude -10° to 50° and longitude 50° to 110° at a grid spacing of 0.25° , making it equal to 241×241 grid points i.e. forecast of 87 variables at 58081 grid points. Finally it becomes a huge datasets of 50,53,047 (approx. 5million) values for just one forecast of a particular time. For the purpose of analysis of sub-grid scale weather events, the ensemble of 96hr, 72hr, 48hr, and 24hr forecast valid for 0600GMT, 1200GMT and 1800GMT has been created. So, very huge dataset of 6,06,36,564 (approx. 60million) data values are to be managed for just one case of tornado / cloudburst.

The GRIB files have been converted to (.csv) format by using National Digital Forecast Database - NDFD GRIB2 decoder program of NOAA downloaded from Internet. The model does not provide vertical motion field (w) directly, so this has been derived by using meridian

(v) and zonal (u) component of wind as forecasted by model. The three different wind velocities have been depicted pictorially in Figure 3.1. In the figure, u denotes zonal wind flow (from west to east) in m/s, v denotes meridian wind flow (from south to north) in m/s and z denotes the wind motion vertically in the atmosphere.

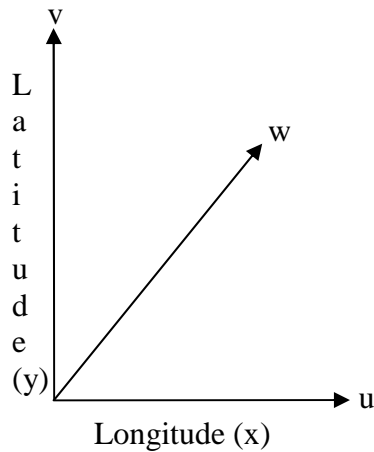


Figure 3.1 – Different Wind velocities

From the conservation of mass equation, as given below:

$$\nabla \cdot \mathbf{V} = 0$$

or

$$\nabla \cdot \mathbf{V}_H + \frac{w}{z} = 0$$

or

$$\frac{u}{x} + \frac{v}{y} = -\frac{w}{z} \dots\dots\dots \text{(Equation 3.1)}$$

So, this term of horizontal convergence is related to the vertical motion as can be seen in the conceptual model of cloudburst in Figure 3.10, section 3.7.1. The other term that has been derived is vorticity that corresponds to rotation of wind, as mentioned below:-

Vorticity formula: $\frac{v}{x} - \frac{u}{y} \dots\dots\dots \text{(Equation 3.2)}$

In the study that follows in section 4.7 and section 4.8, the rotational part (vorticity) did not give any significant signal whereas the field related to divergence part was indicating significant signals of formation of Sub-Grid scale weather systems. The experiments, results and

discussions are thus based on vertical motion field. The vertical wind motion at atmospheric pressure level lower up to 400hPa has been considered. A snapshot of grid points used for calculation is shown in Figure 3.2.

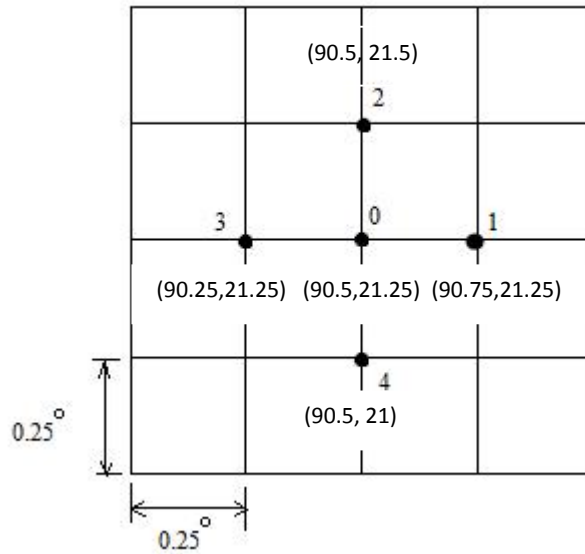


Figure 3.2 – A snapshot of gridded locations to calculate convergence and vorticity, the first term in the bracket corresponds to longitude and second corresponds to latitude. 0.25° corresponds to 25km distance (Source: as a result of pre-processing ECMWF output file in .grib format provided by IMD)

The expressions used for the calculation of convergence and vorticity with reference to equation 3.1 and equation 3.2 respectively, based on the forecast of fundamental variables viz. v wind-velocity and u wind-velocity at grid points of ECMWF model, are mentioned below:-

$$\text{Convergence at grid-point named 0} = \frac{(u_1 - u_3)}{50\text{km}} + \frac{(v_2 - v_4)}{50\text{km}}$$

$$\text{Vorticity at grid-point named 0} = \frac{(v_1 - v_3)}{50\text{km}} - \frac{(u_2 - u_4)}{50\text{km}}$$

A sample of the calculations of convergence and vorticity corresponding to gridded location no. 0 (90.5°E , 21.25°N at 500hPa) in Figure 3.2, for 72 hour forecast valid for 0000GMT 29 July 2009, is depicted in table 3.3.

Point no. (w.r.t. Figure 3.2)	Longitude (°E)	Latitude (°N)	v (m/s)	u (m/s)	vorticity (X10-5 per second)	Convergence (X10-5 per second)
-	90.25	21	7	8	2	-4
4	90.5	21	7	7	2	-2
-	90.75	21	7	7	2	-2
-	90	21.25	7	7	2	-4
3	90.25	21.25	6	7	4	-2
0	90.5	21.25	7	7	4	-4
1	90.75	21.25	7	6	2	-2
-	91	21.25	7	6	2	-2
-	90.25	21.5	6	6	8	-4
2	90.5	21.5	6	6	8	-4
-	90.75	21.5	7	6	4	-4
-	91	21.5	7	5	2	-2

Table 3.3 – Calculation of vorticity and convergence based on 72 hour forecast of v and u wind velocities at 500hPa, valid for 0000GMT 29 July 2009.

(Source: as a result of pre-processing ECMWF output file in .grib format provided by IMD)

The code for calculation of convergence and vorticity has been shown in Appendix D and the complete steps of data pre-processing have been shown in Figure 3.3.

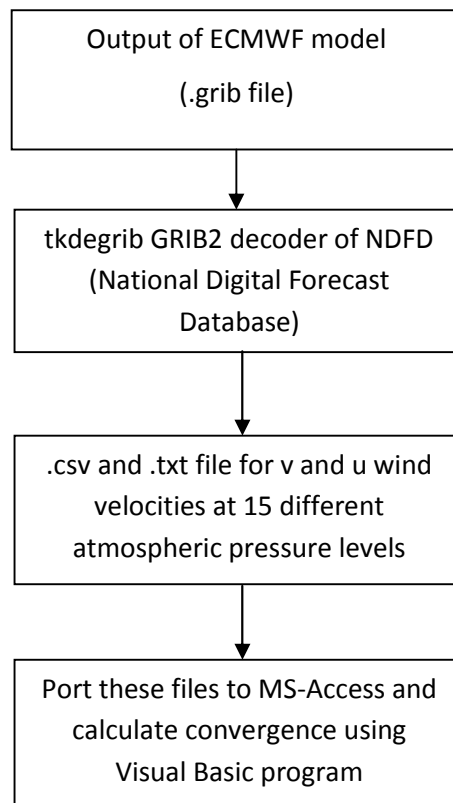


Figure 3.3 – Data pre-processing of forecast by ECMWF model

3.2.2.2 Pre-processing of WRF V3.1 model forecasts

The WRF V3.1 MODEL has been configured based on NCEP analysis of weather variables on 28th, 29th, 30th and 31st March 2009 for forecast of important variables that are indicators of tornado formation. The model configuration at a very fine grid of 9km and run has been done by research scientists at NCMRWF, Noida, Uttar Pradesh.

The forecast at 149 grid points from 76.70129395°E to 89.29870605°E and 139 grid points from 14.39208412°N to 25.41514015°N for 12 z levels viz. 1000hPa, 950hPa, 850 hPa, 750 hPa, 700 hPa, 550 hPa, 500 hPa, 250 hPa, 200 hPa, 150 hPa, 100 hPa, 50 hPa has been made. Forecast of 5 weather variables viz. U (x-wind component (m s-1)), V (y-wind component (m s-1)), W (z-wind component (m s-1)), tc (Temperature (C)) and rh (Relative Humidity (%)) at 12 z levels is being provided. Forecast of 9 weather variables viz. PSFC (SFC PRESSURE (Pa)), U10 (U at 10 M (m s-1)), V10 (V at 10 M (m s-1)), SST (sea surface temperature (K)), RAINC (accumulated total cumulus precipitation (mm)), RAINNC (accumulated total grid scale precipitation (mm)), ws10 (Wind Speed at 10 M (m s-1)), wd10 (Wind Direction at 10 M (Degrees)) and slp (Sea Level Pressure (hPa)) at surface has also been provided. This makes the file containing 14,29,059 values. The dataset is in binary format (.dat) and a corresponding .ctl file that describes the content of .dat file has also been provided by NCMRWF. The content of one of the .ctl files is shown in Appendix E.

The output of model is a .dat file in big-endian format as the model runs on IBM machine and to make it understandable at Intel machine (little endian), the byte-swapping has been done. A program in C plus plus has been written to convert the .dat file to a .txt file containing floating point values. The code for the same is given in Appendix F.

NWP model outputs are huge in size and have multiple dimensions. Hence the need of an organized way for storage of these forecasts was felt. This has been taken up in the following sections.

3.3 Multidimensional Data Model

The NWP model output products have to be stored in a Database System so that a data mining, neural network or any other intelligent technique can be applied. The choices for this are relational or multidimensional databases. Using a relational data model and mastering the concept of table joins is not an easy task. Most ad-hoc query tools do a good job of protecting the user community from the basic relational concepts. Tables and views are usually mapped to concepts such as folders (Oracle White Paper, 2006). An administrator creates all the required

joins and adds additional metadata to enhance the relational objects by changing the names of columns and folder names. However, despite adding additional metadata, users are still exposed to the structural nature of the query language and to some extent the physical storage objects. Also when a particular fact is required to be analyzed across many different dimensions/ column values, the system needs to access all the required tables in query, based on various join conditions and this becomes a sequential access procedure, prone to errors and time-consuming as well.

However, the multi-dimensional model transforms the visualization of a schema into a powerful analytical and fast access environment. A multidimensional data schema is represented by using the concept of a cube. A cube is a logical organization of multidimensional data as shown in Figure 3.4. A cube is derived from a fact table. Edges of cube are referred to as Dimensions that categorize a cube's data. Each dimension is a grouping of common or related columns from one or more tables into a single entity (Han and Kamber, 2006; Berson and Smith, 2005). Dimensions group multiple columns from one or more tables into one single entity organized around one or more hierarchies and ordered by levels. These objects can then be used to provide a very simple query interface.

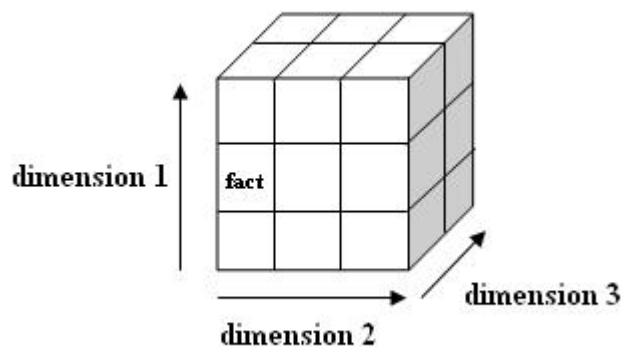


Figure 3.4 – A multidimensional cube having 3 dimensions

A cube contains measures that share the same dimensionality. Measures are just like arrays and are automatically associated to the physical fact table column and related dimension tables. This transformation from fact table column to measure, insulates the user from the complexity of the underlying schema and from the need to understand how the various parts of the schema are joined together. Within an Online analytical processing (OLAP) environment, it is extremely easy to create new measures.

All these structures (cubes, measures and dimensions) interact with each other to provide an extremely powerful reporting environment. Each object adds new levels of interactivity when it is fully exploited.

OLAP environments provide powerful analytical data sources that users can interrogate to follow trends, patterns and anomalies within a cube across any dimension. OLAP provides analytical calculation options such as statistical, forecasting, planning and regression features that help users intelligently manage their data. These concepts allow users to create powerful and focused queries quickly and easily that return consistent and valid analysis as the user extends the query during additional drill-down and pivot operations.

Multidimensional data models (MDDM) have three important application areas within data analysis (Pedersen and Jensen, 2001).

1. *Data warehouses* are large repositories that integrate data from several sources in an enterprise for analysis.
2. OLAP systems provide fast answers for queries that aggregate large amounts of detail data to find overall trends.
3. *Data mining* applications seek to discover knowledge by searching semi-automatically for previously unknown patterns and relationships in multidimensional databases.

Since the NWP output products vary in space and time, thus making them best candidates for storage in a multidimensional database management system. This model has facilitated the storage, selection, analysis and mining of data based on location of occurrence of the event. These models enable seamless usage of algorithms for selection of data for improved interpretation of weather events at sub-grid scale.

Since all of the case studies use MDDM, the implementation of these data cubes from 3-D to 5-D have been explained. A 3-dimensional data model has been implemented to analyze the rainfall across three different dimensions – Gridded location, Time and LPS formation, as explained in section 4.4. In order to understand the impact of an event like cloudburst in a district/state or a river catchment area in a particular time period, the addition of two more dimensions to the cube has been done, making it a 5-D cube (Pabreja 2010b and 2010c), as explained in section 4.5. This paper was awarded the **BEST PAPER** prize during a National level Research Paper Competition (Pabreja, 2010b). Another multidimensional data model for derived weather variables of NWP model output products for generating ensembles is implemented as in section 4.6.

3.4 Comparison between Relational and Multidimensional models

OLAP facility can be implemented as a semantic layer on top of relational store. This layer would provide a multidimensional view, calculation of derived data, drill down intelligence, and generation of the appropriate Structured Query Language (SQL) to access the relational storage. The typical 3rd-normal form representation of data is completely inapplicable in this environment because of the overhead of processing joins and restrictions across a very large number of tables (Colliat, 1996). Instead a denormalized Star Schema has been used to give acceptable performance. The relational schema consists of a Fact table and one table per dimension. The Fact table contains one row for each set of measures and a dimension-id column for each dimension. Rollup summaries are pre-calculated and stored in the Fact table.

The same OLAP model with a server that is based upon a Multidimensional database can be used. The data relevant to the analysis is extracted from a relational Data Warehouse or other data sources and loaded in a multidimensional database which looks like a hypercube. Colliat (1996) has illustrated that the disk space requirement is half in case of MD approach in comparison with relational one. Also in the example considered by the author, there is almost 200 times improvement in the retrieval time of facts.

The access of the fact is very fast as the internal storage of the cube is a tree structure as shown in Figure 3.5. If the total number of dimensions is n , the various unique combinations of different dimensions are 2^n which is equal to the number of unique cuboids. A cuboid can be 1-D, 2-D, 3-D,..... n -D if number of dimensions is n . The time taken to access a fact across any particular combination of dimensions is $O(\log_2 2^n)$.

In the particular case of 5-D data hypercube in section 4.5, the access time of any cuboid containing the fact “rainfall” against any combination of dimensions has been reduced to $O(\log_2 2^5)$ i.e. at the most 5 accesses which otherwise would have been 32 if relational database management system was used.

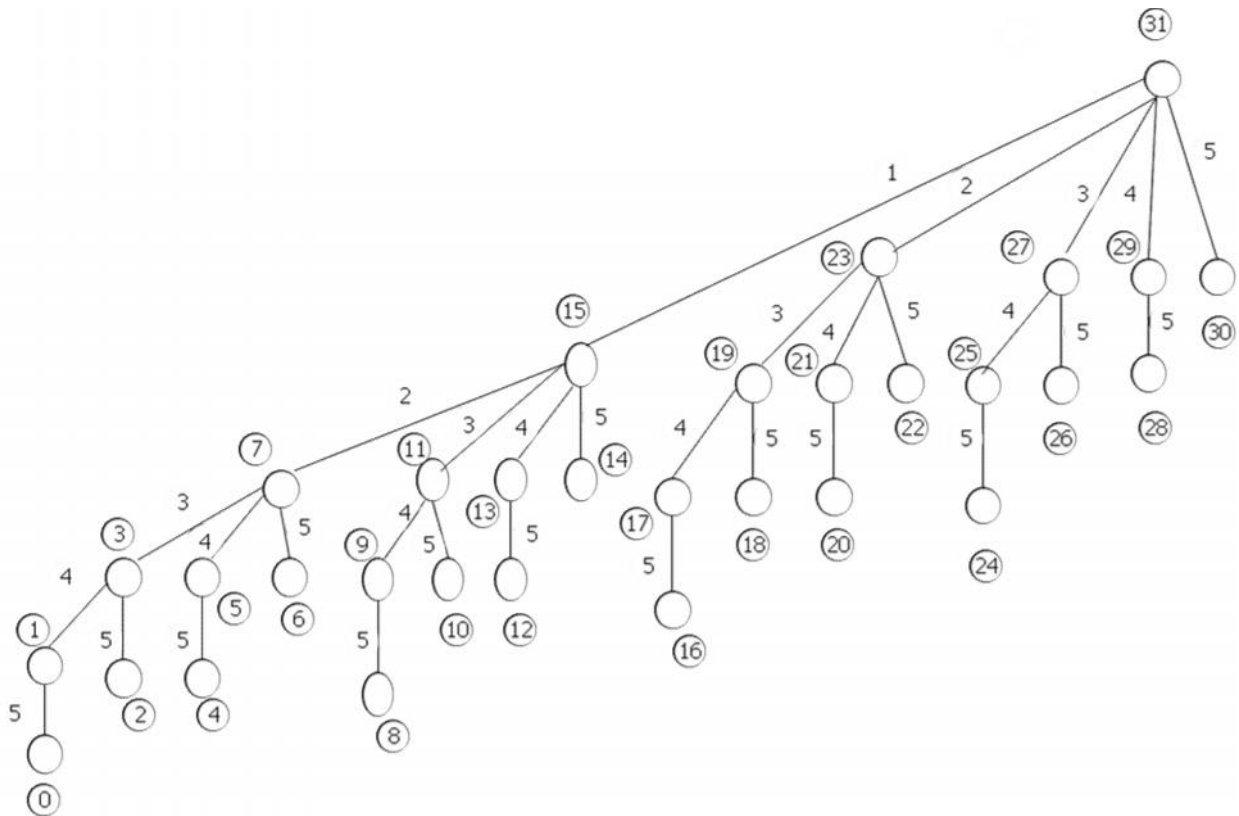


Figure 3.5 – Search of a node corresponding to cuboid in a 5 dimensional data hypercube

3.5 Synoptic scale weather system – Low Pressure System

The synoptic scale in meteorology (with time period of a few days) is a horizontal length scale of the order of 1000 kilometers. This corresponds to a horizontal scale typical of mid-latitude depressions as mentioned in Wikipedia. Most high and low pressure areas seen on weather maps such as surface weather analyses are synoptic-scale systems.

A low pressure area, or "low", is a region where the atmospheric pressure is lowest compared to the surrounding area. As the LPS approach a location, it often results in cloudy weather and precipitation. So, Low Pressure System (LPS) movement during monsoon period, over Indian Region was considered for analysis using k-means clustering so as to locate patterns corresponding to favored zones of formation and dissipation of LPS over Indian sub-continent, as described in section 4.1. This study has been extended to analyze the association of rainfall with reference to movement of LPS over Indian region, as discussed in section 4.2.

3.6 Sub-Grid scale Weather system- Tornado

Tornado is a violently rotating column of air, in contact with the ground, either pendant from a cumuliform cloud or underneath a cumuliform cloud, and often (but not always) visible as a funnel cloud. Tornadoes come in many sizes but are typically in the form of a visible

condensation funnel, whose narrow end touches the ground and is often encircled by a cloud of debris and dust (Grazulis, 2001).

Most tornadoes have wind speeds between 65 km/h and 175 km/h, are approximately 75m across, and travel a few kilometers before dissipating. Some attain wind speeds of more than 480 km/h, stretch more than 1.5 km across, and stay on the ground for few seconds to a few minutes.

Although tornado occurrence is very rare in India but they are very important weather events as they may cause a great damage. Though the physical mechanism for thunderstorm is well understood and it is predictable with modern observational and prediction tools, the detection and prediction of occurrence of tornado over Indian region is still an illusion for the weathermen. Hence, there is a need to understand and predict the tornadoes with potential for devastation. For this purpose, the case studies of tornadoes are very helpful. The major tornadoes which affected India in recent years (Report by IMD, 2009) are listed below:-

- (i) April 19, 1963: A tornado in northwest Assam killed 139 people and left 3,760 families homeless in 33 villages along a 22-mile-long path.
- (ii) March 17-18, 1978: In the northern suburbs of New Delhi, near the University of Delhi, 28 people were killed and 700 were injured by a tornado that swept over the area of 3 miles long and 50 yards wide.
- (iii) April 10, 1978: About 150 people died in a tornado in the Orissa, mostly in the villages of Purunabandha of Keonjhar district.
- (iv) 10 April, 1993: Scores of people were killed when a tornado struck the eastern Indian state of West Bengal, causing devastation in five villages.
- (v) 24 March, 1998: A violent tornado or tornadoes killed 160 people and injured 2000 when it tracked through 20 coastal villages in Midnapore district of West Bengal and Balasore district of Orissa.
- (vi) 31 March, 2009: A tornado accompanied with wind speed of about 250 kmph, thunderstorm, rainfall and hailstorm affected Rajakanika block of Kendrapara district of Orissa in the afternoon as mentioned in (Tornado over Orissa on 31st March 2009 :-A preliminary report by IMD). It caused loss of about 15 human lives and left several injured, apart from causing huge loss of properties.

3.6.1 Tornado genesis and types of Tornadoes

Tornado genesis is the process by which a tornado forms. There are many types of tornadoes, and each type of tornado can have several different methods of formation.

3.6.1.1 Supercellular tornadoes

The supercells were identified as the generators of the deadliest tornadoes, during early 1960s (Grazulis, 2001). The wind shear (change of either the wind speed or the wind direction at different altitudes) gives rise to horizontal rotation near the surface, producing vortex tubes as shown in Figure 3.6. Vertical wind shear can cause a rolling action of the wind. The result is a horizontally rotating column. This horizontal vorticity is very weak. In this situation, the wind is gradually changing direction with altitude. Near the surface the wind might blow from the southeast; at one half mile above ground level it might be blowing from the south; at one mile up it might be blowing from the southwest. If winds were sheared this way, they would cause the horizontal tube to rotate clockwise. The heating of the ground by the sun would then produce updrafts that cause the rotating tube to tilt into a vertical position, as shown in Figure 3.6. It is the spin of the vorticity of the tube that is lifted into a vertical position and is probably the source of rotation for the updraft, the mid-level mesocyclone (presence of a unique area of rotation) and perhaps some of the rotation of the tornado. The vortex tube is a conceptual model that allows us to more easily work with the tricky subjects of wind shear and vorticity. The idea that the vertical axis of rotation in a supercell begins in a horizontal position was suggested by Browning and Landry (1963) and refined by (Barnes, 1968; 1970).

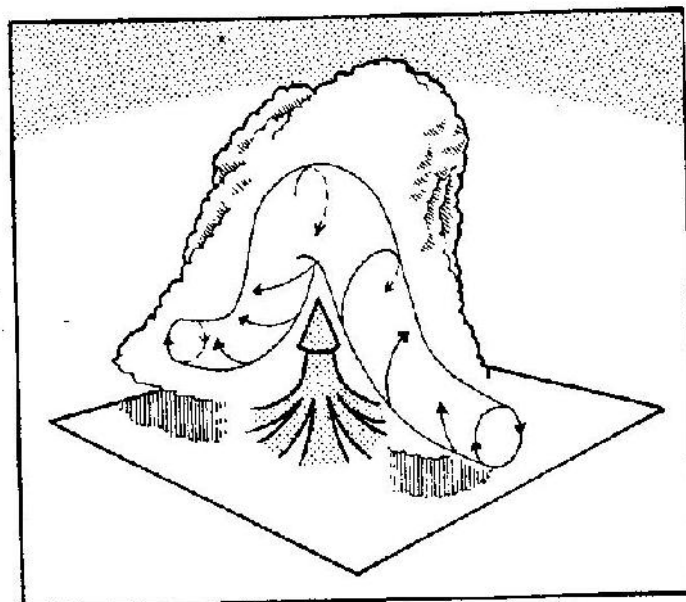


Figure 3.6 – Vertical wind shear can result in rotation of air around a horizontal axis (Drawing by David Hoadley)

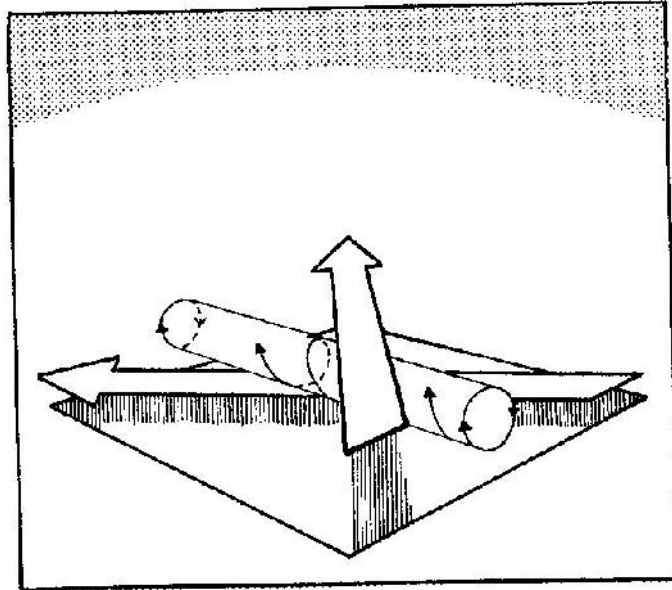


Figure 3.7 – Vortex tubes with horizontal vorticity are lifted into a vertical position by rising air
(Drawing by David Hoadley)

3.6.1.2 Multiple vortices from a single-vortex tornado

Fujita (1970) and Ward (1972) studied and focused upon unique tornadic event of dozens of small funnels at Newton, Kansas on May 24, 1962. The gradual change of a single-vortex tornado into a multiple-vortex form is shown in Figure 3.8. The figure is a series of two-dimensional, vertical cross-sections of a very complex three-dimensional structure. Figure 3.8a shows the cross-section of a single-vortex tornado with the air flowing inward and upward. The air swirls about the axis of the vortex at increasing speed while the pressure in the centre of the vortex continues to drop. Figure 3.8b shows that this pressure drop causes the vortex to fill in from above and the air begins to descend in the core of the funnel. This process is known as vortex breakdown. In Figure 3.8c the descending air reaches the surface and flows outward towards the walls of the tornado. There it meets inward flowing air. This interaction produces three sub-vortices as shown in Figure 3.8d.

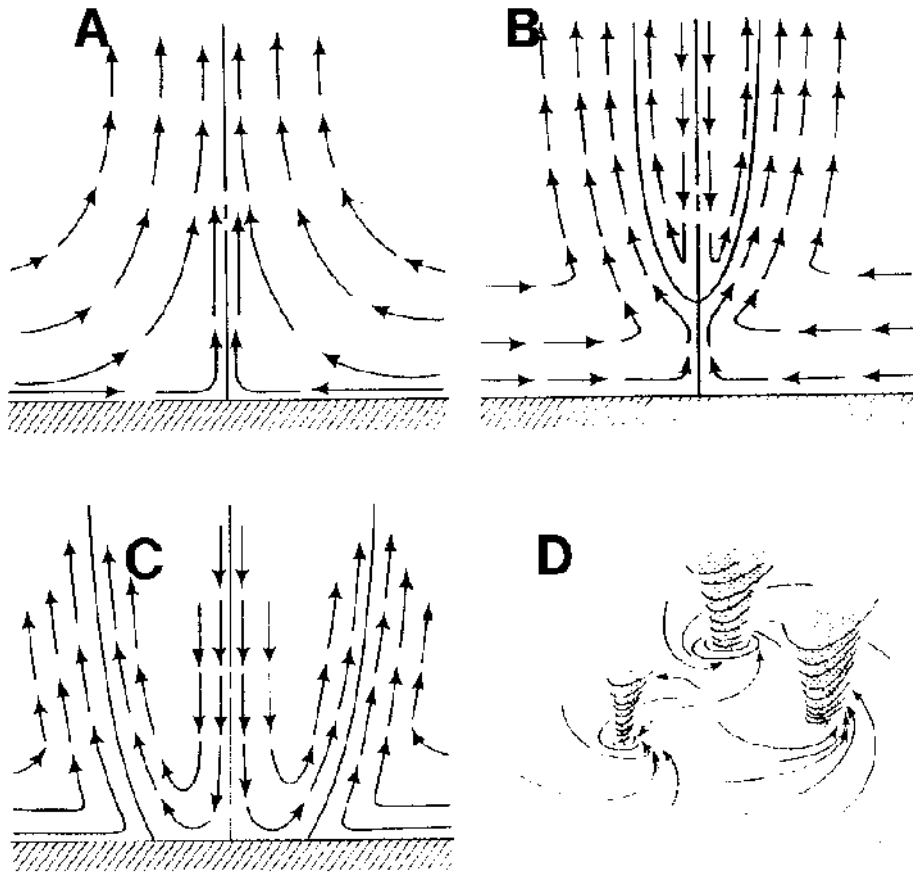


Figure 3.8 – The four stages in the development of multiple vortices (Source: Davies-Jones, 1985)

3.6.2 Tornado Wind Speed

As discussed earlier, the source of tornado formation is wind shear that causes horizontal rotation near surface of earth coupled with heating of the ground that produces updraft which ultimately causes the rotating tube to tilt into a vertical position. Figure 3.9 shows the wind speeds in a rapidly rotating single-vortex tornado that is moving forward at 56 mph. Had the tornado been standing over a single location, the wind speed at the outer edge of the vortex might have been the same on all sides, about 200mph. The forward motion of the tornado is to the right and causes the wind on the right side of the tornado to increase by 56mph. The wind on the left side would decrease by the same amount. This results in a 112mph difference from one side of the tornado to the other. This movement also shifts the area of the calm air from the centre of the tornado to the left.

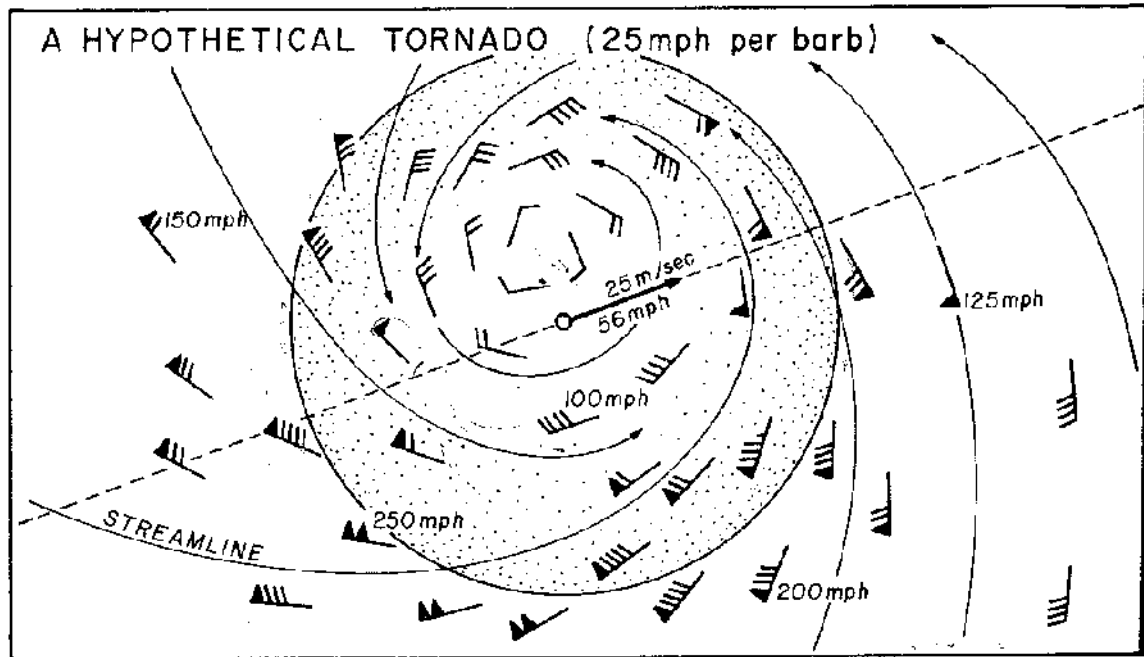


Figure 3.9 – The distribution of winds inside a single-vortex tornado. Each of the flags on the wind direction indicator represents 125mph. Each of the barbs (small lines) is 25mph. (Source: Fujita, 1981)

3.6.3 Fujita scale of tornado intensity

Fujita scale of Tornado intensity was devised by Ted Fujita of the University of Chicago (Grazulis, 2001). The Fujita scale (F-scale) is a classification system analogous to the Richter scale of earthquake intensity. F-scale rating varies from 0 to 6. The Fujita Scale is used to rate the intensity of a tornado by examining the damage caused by the tornado after it has passed over a man-made structure. The worst damage is then compared to the descriptions and the best guess for an F-scale number is made. A wind speed range can then be estimated for this tornado. F-scale rating standards as explained by Grazulis (1993) are listed below in table 3.4.

F-Scale	Tornado	Wind speed	Damaged caused
F0	Gale	less than 116 km/h	Damage to chimneys, damage sign boards, and break branches off of trees and topple shallow-rooted trees.
F1	Moderate	117-180 km/h	Peel surfaces off of roofs, push mobile homes off of their foundations or even overturn them, and push cars off of the road
F2	Significant	181-253 km/h	Can tear the roofs off of light frame houses, demolish mobile homes, overturn railroad boxcars, uproot or snap large trees, lift cars off the ground, and turn light objects into missiles.
F3	Severe	254-332 km/h	Can tear the roofs and walls off of well-constructed houses, uproot the trees in a forest, overturn entire trains, and can throw cars.
F4	Devastating	333-416 km/h	Can level well-constructed houses, blow structures with weak foundations some distances, and turn large objects into missiles.
F5	Incredible	417-509 km/h	Can lift and blow strong houses, debark trees, cause car-sized objects to fly through the air, and cause incredible damage
F6	Inconceivable	above 509 km/h	there would be no objects left to study

Table 3.4 – Fujita scale of tornado intensity (Source: Grazulis (2001))

3.6.4 Tornado forecasting

It's hard to forecast tornadoes. They don't last very long and are also very complicated. Scientists don't really know how they form, but they do know where they tend to form (Grazulis, 2001). Using what they know about the atmospheric conditions from past tornadoes, meteorologists can tell when they may form.

Continued research and advancements in computer technology from the 1960s onwards improved severe weather and tornado forecasting. There are several new observing technologies to which the operational weather services in the United States are already committed i.e. WSR-88D radar network as explained by Charles et al. (1993).

Doppler radar is used to detect storms and their intensity. Doppler radar works by sending out radio waves from an antenna. These radio waves are reflected back to the antenna by objects in the air. Through this process Doppler radar can detect precipitation in the air. It even detects frequency differences based on whether an object is moving away from the antenna or towards it. The frequency will be lower if the object is moving away from the antenna and higher if the object is moving toward the antenna. After the antenna detects an object it sends the information back to a computer that brings up the different frequencies as different colors. The colors used represent speed and direction to the user. Using this method, meteorologists can tell a lot of things about the storms it detects, as explained online by Premium Weather service.

Using this important information meteorologists are then equipped to send out warnings for specific areas. Doppler radar aids meteorologists in sending out tornado warnings but the radar cannot forecast the tornado.

The other technological tool for the future is the numerical prediction model. The tornado is many orders of magnitude smaller than the environmental processes that give it birth, so scale interaction is a crucial question as illustrated in Figure 1.1. However, if model grid-spacing is decreased to less than a km to a few meters (the mesoscale / microscale boundary), this will require enormous computing resources (CyRDAS, 2004). Also for the NWP model to do processing of the huge volumes of observations for nowcasting or forecasting at sub-grid scale is an inherently time constrained task. It is not possible for the forecasters and computers to analyze data for a few hours just for a three-hour prediction.

So, it will be many years (if ever) before a single numerical prediction model will encompass all the time and space scales explicitly and simultaneously. There can be no doubt that in future numerical prediction models will continue to assume ever greater roles in the tornado forecast problem (Charles et al., 1993).

3.7 Other Sub-Grid scale Weather Phenomenon –

Cloudburst

Extreme Weather Events (Singh and Roy, 2002) are the infrequent meteorological events that have a significant impact upon the society or ecosystem at a particular location. They can also be defined as those short term perturbations of the weather that provide magnitudes much outside the normal spectrum or range within the typical averaging period. Usually more than twice the standard deviation is assumed to define an extreme event. If the average climate

changes then the frequency and magnitude of extreme weather events are also likely to change. Predicting the effects of climate change on the magnitude, frequency, timing and duration of extreme weather events is very difficult. There are some extreme weather phenomena that are highly localized, at a spatial scale of 2-20km, defined as meso-gamma scale (Orlanski, 1975) and are very short-lived (approx. 30min -5hours). There are no explicit synoptic conditions associated with these events. Cloudburst is one such event that can bring about large scale changes in the form of soil erosion, landslides and flashfloods.

Cloudbursts or downpours are sub-grid scale weather events that have no strict meteorological definition (Heidom, The Weather doctor, online book). Cloudburst usually signifies a sudden, heavy fall of rain over a short period of time. The rainfall rate in excess of 25 millimeters per hour (1 inch per hour) constitutes a downpour.

Most cloudbursts come from convective, cumulonimbus clouds that form thunderstorms and that the air is generally rather warm in order to contain the amount of moisture needed for a heavy downpour. Besides providing the proper conditions to spawn large quantities of liquid water drops, cumulonimbus clouds have regions of strong updrafts which hold raindrops aloft en masse and can produce the largest raindrops (those greater than 3.5 mm, in diameter).

These updrafts are filled with turbulent wind pockets that toss small raindrops around with surprising force. Within the turmoil of the randomly moving drops, there are more collisions among the drops and many of those close encounters result in their conglomeration into new drops larger in size.

Eventually all updrafts collapse, and when they do, the upheld raindrops descend unimpeded toward the surface, often forming a strong downdraft — such as a downburst or microburst — in the process, an event that appears as if the cloud has burst open. So, not only are the larger drops falling with a terminal velocity of around 12 km/h, but they have the added giddy-up of the downdraft speed, which can easily exceed 80 km/h.

The resulting rainfall is a torrent of water, large raindrops falling at high speed, over a small area. The hard rain characteristic of a cloudburst is caused by a phenomenon known as Langmuir precipitation, in which drops of rain fuse together to create large drops as they fall, falling even more quickly as they grow (Smith, online and Wikipedia). The force and quantity of such downpours can be damaging to vegetation, small animals, and property. When the speed of water accumulation on the ground exceeds the surface's ability to absorb it, localized flooding will occur in low-lying terrain. In hilly or mountainous terrain, the runoff of water can congregate in stream beds or canyons and cause deadly and damaging flash flooding.

3.7.1 A conceptual model of the cloudburst

A conceptual model of the cloudburst based on the development of the vertical shear, vertical motion and the moisture distribution is shown in figure 3.10, which illustrates (Das, Arshit, and Moncrief, 2006) three stages in the lifecycle of the cloudburst. The drift of the cells towards each other and their vertical motion as explained in section 3.2.2.1 can be related to the concept that follows.

In the first stage, the two convective cells are separate and drift towards each other as part of the mean flow (figure 3.10a). Isolated heavy rain occurs during this stage. In the second stage (figure 3.10b), the two convective cells merge. Intensification follows due to strong wind shear and intense vertical motion. Heavy downpour and formation of the anvil also occurs. The storm moves rapidly southward due to strong steering flow. The third stage (figure 3.10c) is one of dissipation in which the two merged cells form one single large cell, which drifts westward and the cloudburst ceases.

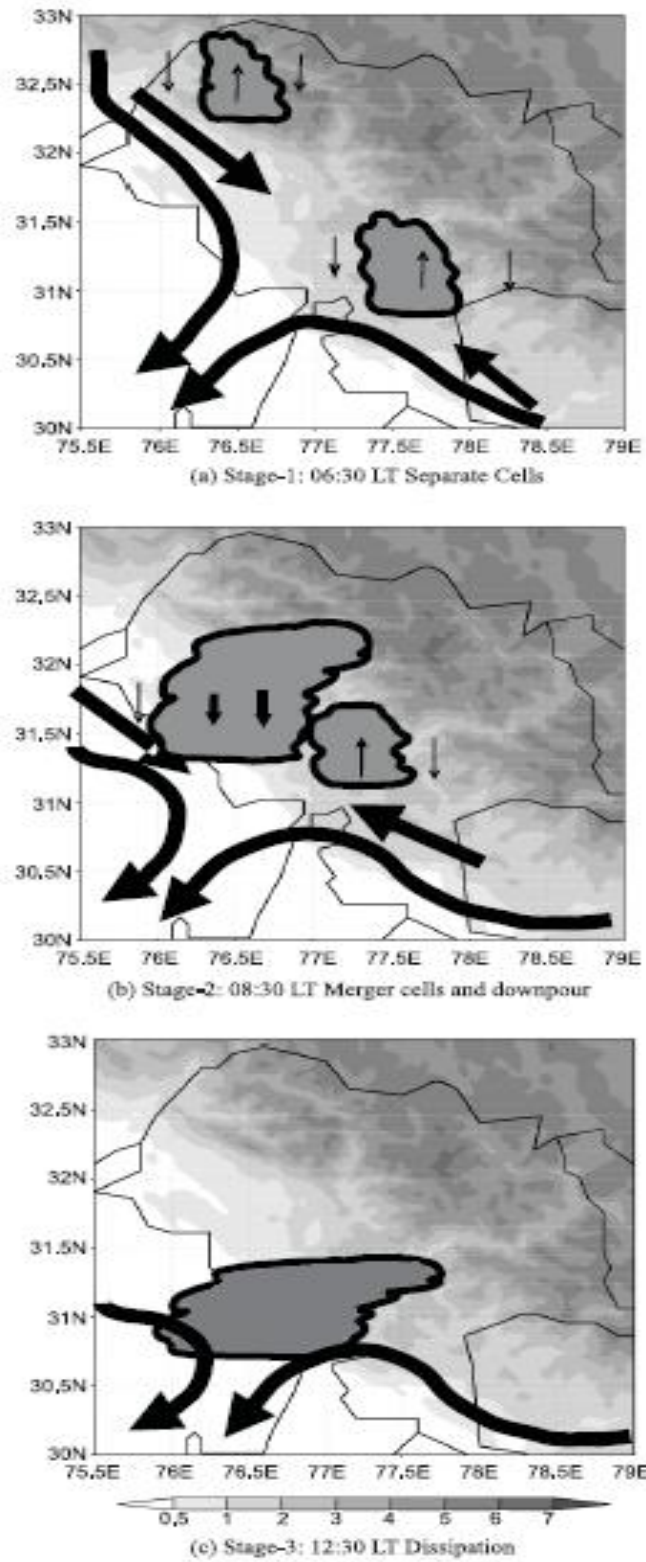


Figure 3.10 – A conceptual model of cloudburst Stage-1: Separate Cells, Stage-2: Merger of Cells and heavy downpour, Stage-3: Dissipation (Source: Das, Arshit, and Moncrief, 2006)

3.7.2 Places that suffer from Cloudbursts

Cloudbursts are primarily observed over coastal areas accompanied with tropical cyclones and over hilly regions associated with interaction of monsoon flow during monsoon and western disturbance in other months *e.g.* February, March. In the hilly areas, when the moisture in the air is drawn into the storm by the violent uprushing currents of the storm, the moisture condenses as it rises on account of the cooling of the air by expansion, but the upward blast is so strong that for a time the water is prevented from falling as rain. If the rising currents are weakened at some point, a large accumulation of water is permitted to fall at one time. This is especially likely to occur (Talman, 1930) when a travelling thunderstorm, which is fed by rising streams of air from overheated ground, passes over the cooler surface of a mountain, so that its supply of warm air is temporarily cut off.

Besides these, cloudbursts can also be perceived in the areas like inner parts of various continents. The cloudbursts are mostly seen during the monsoon months.

3.7.3 Forecast of cloudburst

Numerous studies (Huiming, 1996) have shown that the medium-scale system laying over large-scale or weather scale system like the pressure trough or frontal surface is the weather system that directly produces torrential rain. The medium-scale system ranges from tens of kilometers to hundreds of kilometers and its life time is several hours to a day. The current meteorological observation, especially the high altitude station network used for monitoring large-scale weather system is hard to capture medium-scale system.

The satellite clouds pictures are provided by meteorological satellites and are divided into visible pictures and infrared pictures. In particular, the picture of stationary meteorological satellite with high spatial resolution and time resolution is the most effective tool for monitoring, analyzing and forecasting the medium-scale system which causes torrential rain. But at times there is difficulty in discriminating between a severe hailstorm and a heavy cloudburst (Held, 1981).

Cloudburst causing medium-scale system is expressed as a bright cloud cluster on the satellite picture. The area of the cloud cluster is closely related to the range of cloudburst. Heavy rainfall mostly comes from vigorous convection cloud area that has high cloud top and low temperature at the cloud top. Through the enhancement display of the infrared cloud picture, heavy precipitation center can be seen clearly according to thermal stratification, which becomes an important means for cloud burst forecast, explained by Henderson (2009).

These techniques cannot predict the occurrence of cloudburst well in advance. The technique for 1-2 days advance prediction does not exist in any part of the world. But the prediction of environmental conditions favorable for occurrence of event is possible with some accuracy 3-4 days in advance (Das, 2005; Das, Arshit, and Moncrief, 2006) with the help of high resolution numerical weather prediction model. There is thus a possibility that the prediction skill of these events could improve through the Data Mining clustering techniques based on the NWP output (direct and derived) parameters.

Chapter 4

Case studies

4.1 Analysis of Synoptic scale weather system “Low Pressure System” Movement over Indian Region

4.1.1 Dataset used

Two different datasets corresponding to LPS were examined:

1. Year 1984, 85, 88, 89, 90, 91, 92, 93, 94 (Years 1986 and 1987 were not considered as these were poor rainfall years.)
2. Year 1995 to 2003.

The LPS data has been made available with courtesy of (Mooley and Shukla, 1987) and (Sikka, 2006) and this data is for the monsoon months June to September for each year under consideration. The LPS data contains the serial number of the system in the year, the date of formation of the system (the date is also given on 1 to 122 scale: 1 for 1st June and 122 for 30th Sept), the intensity stage, and the location of the system for each day of existence (latitude °N and longitude °E correct to the first decimal). The intensity stage is given by 1 for Low, 2 for Depression, 3 for Deep Depression, 4 for Cyclonic Storm and 5 for Severe Storm. Data has been computerized, then time interval column is added, the value 1 indicates the formation of LPS and 2, 3, 4 indicate the movement of LPS and finally the last value either 4 or 5 indicates the location of dissipation of the LPS. The datasets are converted to .csv (comma separated variables) format, as required by the WEKA (Waikato Environment for Knowledge Analysis), downloaded from Internet, for data mining (Witten and Frank, 2005). For illustration, a small sample of the LPS data for the year 1984 is shown in table 4.1.

S.No.	Date	Day No.	Longitude	Latitude	Time Interval
8	16-Jul-84	46	89	19.5	1
	17-Jul-84	47	88	20	2
	18-Jul-84	48	83	20.5	3
	19-Jul-84	49	76	22	4
	20-Jul-84	50	72.5	23.5	5
9	30-Jul-84	60	89	19.6	1
	31-Jul-84	61	88.5	19.5	2
	1-Aug-84	62	82.5	22	3
	2-Aug-84	63	79	24	4
	3-Aug-84	64	76	25	5
10	7-Aug-84	68	89	20	1
	8-Aug-84	69	88	22	2
	9-Aug-84	70	84.5	22.5	3
	10-Aug-84	71	83	23	4
	11-Aug-84	72	78	23	5
11	11-Aug-84	72	89	21	1
	12-Aug-84	73	88	21	2
	13-Aug-84	74	89	20	3
	14-Aug-84	75	89.5	20	4
	15-Aug-84	76	90	20	5
12	26-Aug-84	87	90	22	1
	27-Aug-84	88	88	22.5	2
	28-Aug-84	89	85	23.5	3
	29-Aug-84	90	85.5	24	4
	30-Aug-84	91	85	25	5

Table 4.1 – A sample of Low Pressure System data for year 1984 (Source: Sikka, 2006)

4.1.2 Technique Used

An open Source Data Mining tool has been used that offers many techniques for mining of datasets. We have used k-means clustering technique to group a set of objects into clusters so that the objects within a cluster have a high similarity in comparison to one another, but are dissimilar to objects in other clusters.

The k-means method first chooses k points at random as cluster centers. All instances are assigned to their closest cluster centre according to the ordinary Euclidean distance metric. Next the centroid/mean of the instances in each cluster is calculated. These centroids are taken to be new center values for their respective clusters. Finally, the whole process is repeated with the new cluster centers. Iteration continues until the same points are assigned to each cluster in consecutive rounds, at which stage the cluster centers have stabilized and will remain same forever. The groups that are identified are exclusive so that an instance belongs to only one group.

The k-means method of clustering is used to generate clusters corresponding to the day of formation of the system to the day it dies. The tool provides various options to select/ reject

the attributes on the basis of which the clustering should be done. As the focus was to locate favored zones of formation and dissipation so the attribute selected as feature for clustering was time interval and so the number of clusters required has been taken as equal to two. The selection of the LPS dataset has been done so that it contains only the rows corresponding to the first and last day of LPS.

For the other analysis to locate favored zones of LPS movement during months of June and July, here the important feature for clustering is date. The additional field corresponding to month has been added and clustering based on this feature has been done.

4.1.3 Findings

For visualization of clusters, the x-axis is chosen as longitude and y-axis as latitude. The area under consideration is from latitude 15.0°N to 25.0°N and longitude 66.5 °E to 90.0 °E so that focus remains on Indian region, as shown in Figure 4.1.

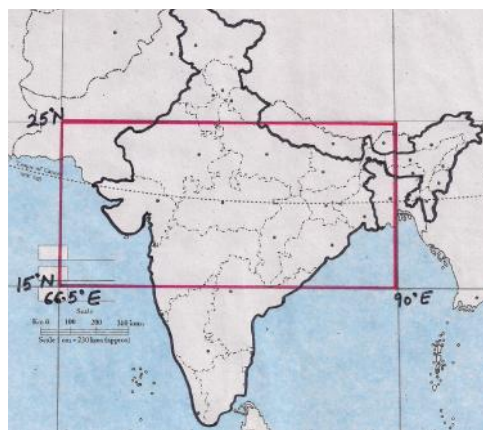


Figure 4.1 – Area under analysis for LPS movement over Indian region

Using data mining technique has resulted in locating a cluster of formation of LPS on 1st day and a cluster of disappearance on 4th or 5th day of formation, on a spatio-temporal scale for the months of June-July and Aug-Sept separately, for two sets of 9 years. These clusters corresponding to June-July for years (year 1984, 85, 88, 89, 90, 91, 92, 93, 94) and (year 1995 to 2003) are shown in Figure 4.2 and Figure 4.3. Similarly clusters corresponding to Aug-Sept for years (year 1984, 85, 88, 89, 90, 91, 92, 93, 94) and (year 1995 to 2003) are shown in Figure 4.4 and Figure 4.5.

There is a strong resemblance between the location of clusters of formation and disappearance in Figure. 4.2 and Figure 4.3 which are for Jun-July months for the two sets of years. There is also a strong resemblance between the location of clusters of formation and disappearance in Figure 4.4 and 4.5 which is for Aug-Sept months, for the two sets of years. Hence the favored zones of formation and disappearance of LPS on a spatio-temporal scale, over

Indian region during June to September could be identified. This case study led to the following conclusions:-

- (i) Data mining clusters represent situations that are acceptable to meteorological community.
- (ii) Clusters of LPS move geographically from south to North corresponding to June to July as shown in Figure 4.6. The changes in location of LPS during June-July to Aug-Sept are reflected in Figure 4.2 to Figure 4.5.

These results were discussed with meteorological community during Intromet 2009 (Pabreja, 2009) and were appreciated a lot. This validated the use of DM tools for well formed synoptic scale weather system.

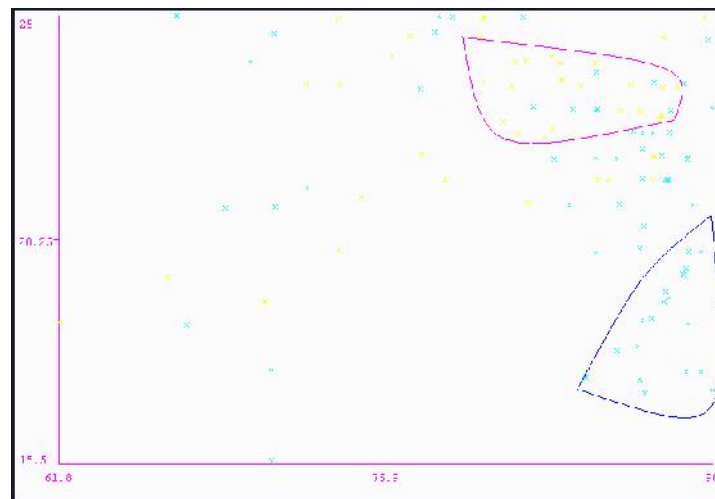


Figure 4.2 – Clusters of formation and disappearance of LPS during June-July 1984, 85, 88 to 94
Blue– data points of formation of LPS. Yellow- data points of disappearance of LPS

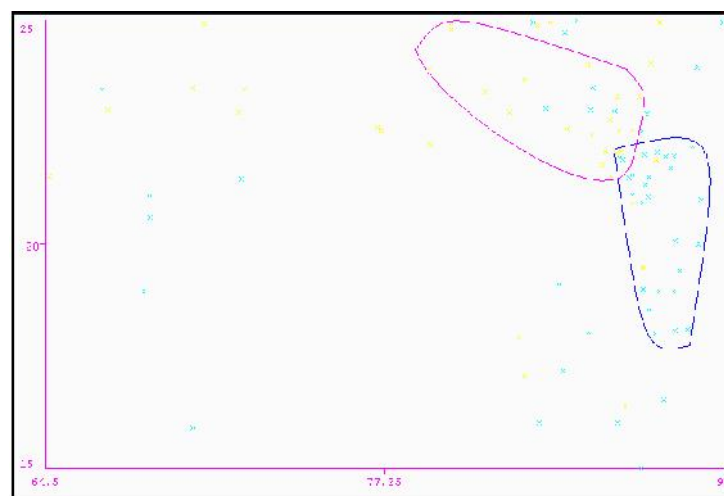


Figure 4.3 – Clusters of formation and disappearance of LPS during June-July 1995 to 2003.
Blue – data points of formation of LPS. Yellow- data points of disappearance of LPS

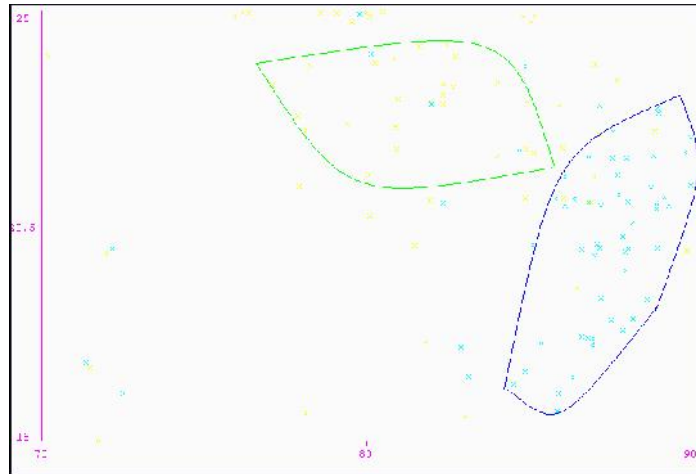


Figure 4.4 – Clusters of formation and disappearance of LPS during Aug-Sept 1984, 85, 88 to 94.
 Blue – data points of formation of LPS. Yellow- data points of disappearance of LPS

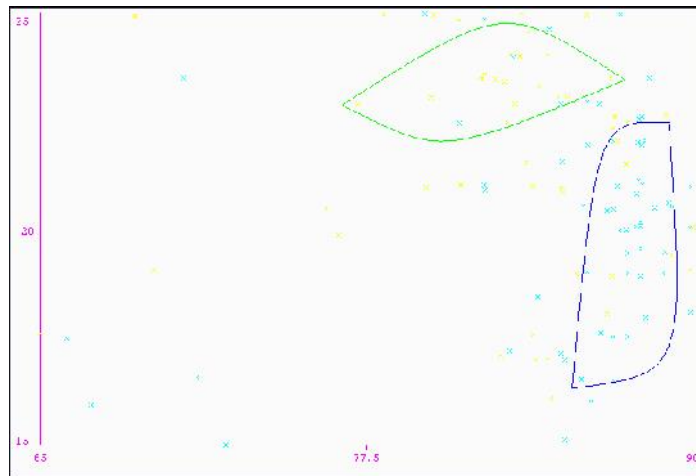


Figure 4.5 – Clusters of formation and disappearance of LPS during Aug-Sept 1995 to 2003.
 Blue – data points of formation of LPS. Yellow- data points of disappearance of LPS

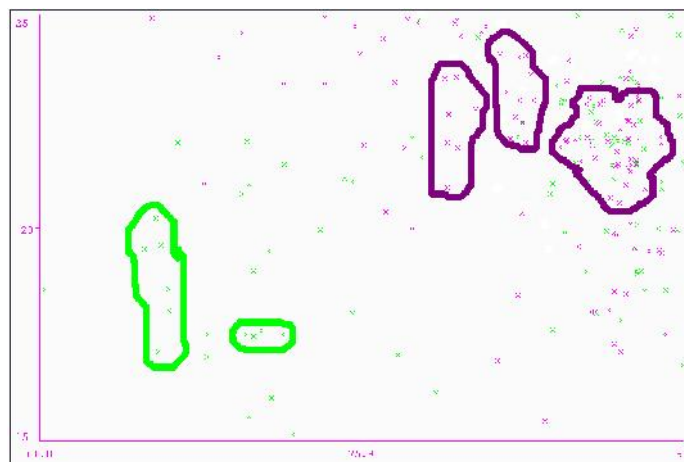


Figure 4.6 – Clusters of LPS formation, movement, dissipation during June, July 1984, 85, 88, 89, 90, 91, 92, 93, 94
 Green represents LPS during June and Pink represents LPS during July

4.2 Analysis of Rainfall with reference to movement of “Low Pressure System” over Indian Region

4.2.1 Datasets used

LPS datasets is same as described in section 4.1.1. In addition to this, the daily gridded rainfall datasets whose source and pre-processing has been explained in section 3.2.1 is used for the study. The datasets under study are for year 1994. The LPS moving in North West direction has been chosen and the corresponding Rainfall because of this LPS movement is being analyzed.

4.2.2 Technique Used and findings

The position of formation of LPS has been mapped to co-ordinate (0, 0) and accordingly the following locations of movement of LPS and dissipation of LPS has been calculated with reference to this point. A snapshot of this pre-processed LPS dataset has been shown in table 4.2 below.

Date	Day number	Longitude (°E)	Latitude (°N)	time interval	normalized longitude	normalized latitude	LPS no.
25-Jun-94	25	85	23	1	0	0	4
26-Jun-94	26	87	22.5	2	2	-0.5	4
27-Jun-94	27	87	22	3	2	-1	4
28-Jun-94	28	86.5	22	4	1.5	-1	4
29-Jun-94	29	84.5	23.5	5	-0.5	0.5	4
30-Jun-94	30	80	24	6	-5	1	4
1-Jul-94	31	76.5	26	7	-8.5	3	4
3-Jul-94	33	87	17	1	0	0	5
4-Jul-94	34	80	22	2	-7	5	5
5-Jul-94	35	77	24.5	3	-10	7.5	5
8-Jul-94	38	88	21.5	1	0	0	6
9-Jul-94	39	86.5	22	2	-1.5	0.5	6
10-Jul-94	40	87	21.5	3	-1	0	6
11-Jul-94	41	87.5	22.5	4	-0.5	1	6
12-Jul-94	42	87.5	21.5	5	-0.5	0	6
13-Jul-94	43	87	21	6	-1	-0.5	6
14-Jul-94	44	85.5	22	7	-2.5	0.5	6
15-Jul-94	45	88	21.5	8	0	0	6

Table 4.2 – A sample of Low Pressure System data (normalized movement) for year 1994 (Source: Sikka, 2006)

Similarly the Rainfall datasets for each LPS have been considered e.g. for LPS number 5 whose starting date is 3 July, 1994 and dissipating date is 5 July, 1994 and location of movement from

(17.0°N and 87.0°E) to (24.5°N and 77.0°E), the Rainfall for the area (longitude \geq 76.5°E And longitude \leq 87.5°E) And (latitude \geq 16.5°N And latitude \leq 25°N) for same dates have been considered. The co-ordinates of Rainfall locations have also been normalized according to what has been done for LPS. The Rainfall values have been given nominal values as low (16mm - 40 mm), good (>40mm to 75mm), heavy (>75mm - 120 mm), very heavy (>120mm -250mm) and extremely heavy (>250mm). The final data with normalized values of co-ordinates and nominal values of Rainfall is shown in table 4.3.

Date	Latitude (°N)	Longitude (°E)	Rainfall in mm	Category of Rainfall	Normalized Latitude	Normalized Longitude
3-Jul-94	19.5	78	39.3	low	2.5	-9
4-Jul-94	21.5	77	39.3	low	4.5	-10
5-Jul-94	21	77.5	39.6	low	4	-9.5
4-Jul-94	23	79	40	low	6	-8
3-Jul-94	24	86.5	60.2	good	7	-0.5
4-Jul-94	23.5	77.5	60.2	good	6.5	-9.5
3-Jul-94	16.5	81	60.3	good	-0.5	-6
3-Jul-94	20	84	60.6	good	3	-3
4-Jul-94	20	80.5	60.9	good	3	-6.5
3-Jul-94	24.5	86	65	good	7.5	-1
3-Jul-94	19.5	80.5	76.3	heavy	2.5	-6.5
3-Jul-94	19.5	81.5	77.5	heavy	2.5	-5.5
4-Jul-94	17.5	80	77.8	heavy	0.5	-7
4-Jul-94	18.5	81	77.8	heavy	1.5	-6
4-Jul-94	19	80.5	80.4	heavy	2	-6.5
3-Jul-94	24	86	81.5	heavy	7	-1
4-Jul-94	19.5	80	134.2	very heavy	2.5	-7
4-Jul-94	19	80	137.8	very heavy	2	-7
4-Jul-94	19.5	79	169.9	very heavy	2.5	-8
4-Jul-94	19.5	78.5	204.9	very heavy	2.5	-8.5

Table 4.3 – A Sample of Rainfall for year 1994 corresponding to LPS from 3July1994 to 5July 1994 organized in tabular format (Source: as a result of pre-processing rf1994.grd provided by IMD)

The clusters corresponding to good, heavy, very heavy, extremely heavy rainfall have been generated using k-means clustering technique. The number of clusters has been taken as 4 as the focus is on the category (amount) of rainfall and hence feature selected for clustering is “Category of Rainfall”. Corresponding to each case of LPS moving in North West direction for the year 1994, the Rainfall clusters have been generated. The plot of LPS movement is superimposed on the corresponding plot of Rainfall. The plots for LPS formed on 25June 1994, 3July 1994 and 11Aug 1994 are shown in Figure 4.7 to Figure 4.9 respectively.

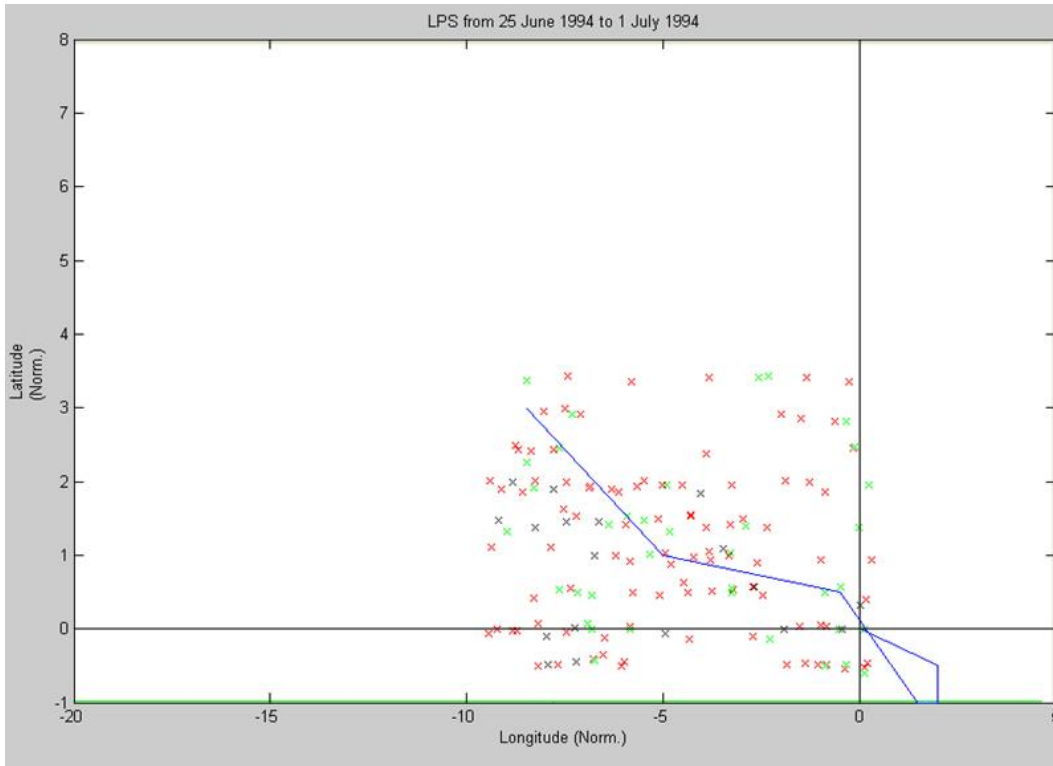


Figure 4.7 – Rainfall with reference to LPS movement from 25 June 1994 to 1 July 1994 (Blue line depicts LPS movement) (Red: Good Rainfall, Green: Heavy Rainfall, Black: very/extremely heavy Rainfall)

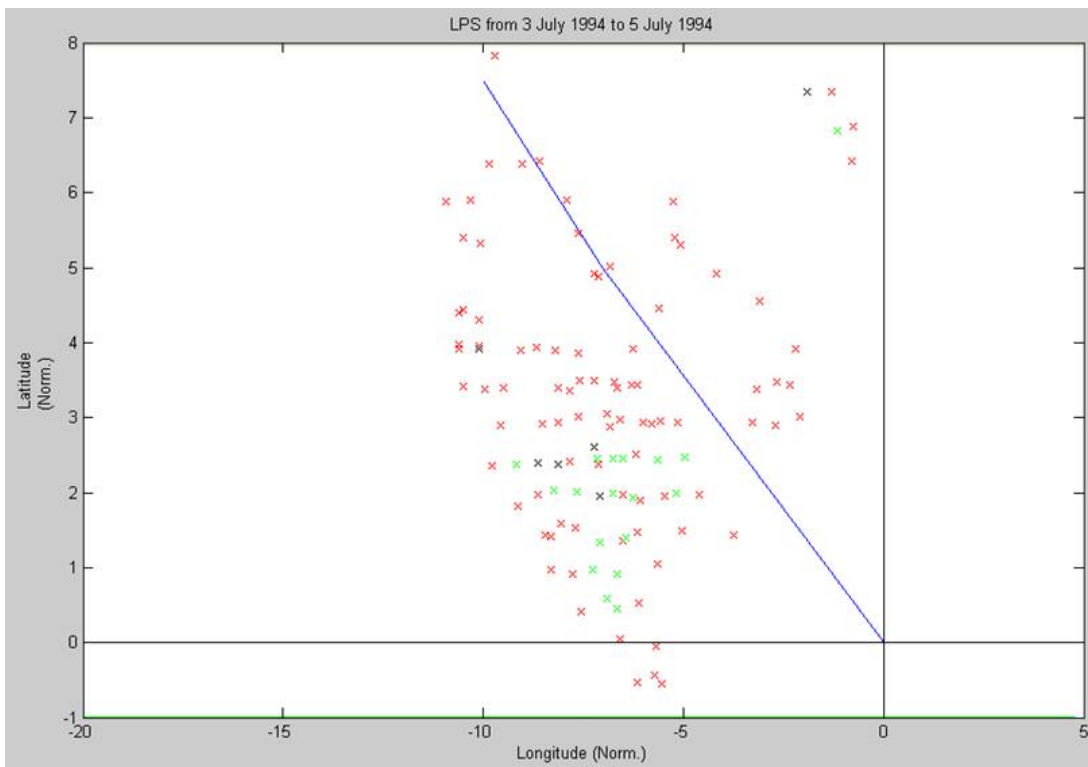


Figure 4.8 – Rainfall with reference to LPS movement from 3 July 1994 to 5 July 1994 (Blue line depicts LPS movement) (Red: Good Rainfall, Green: Heavy Rainfall, Black: very/extremely heavy Rainfall)

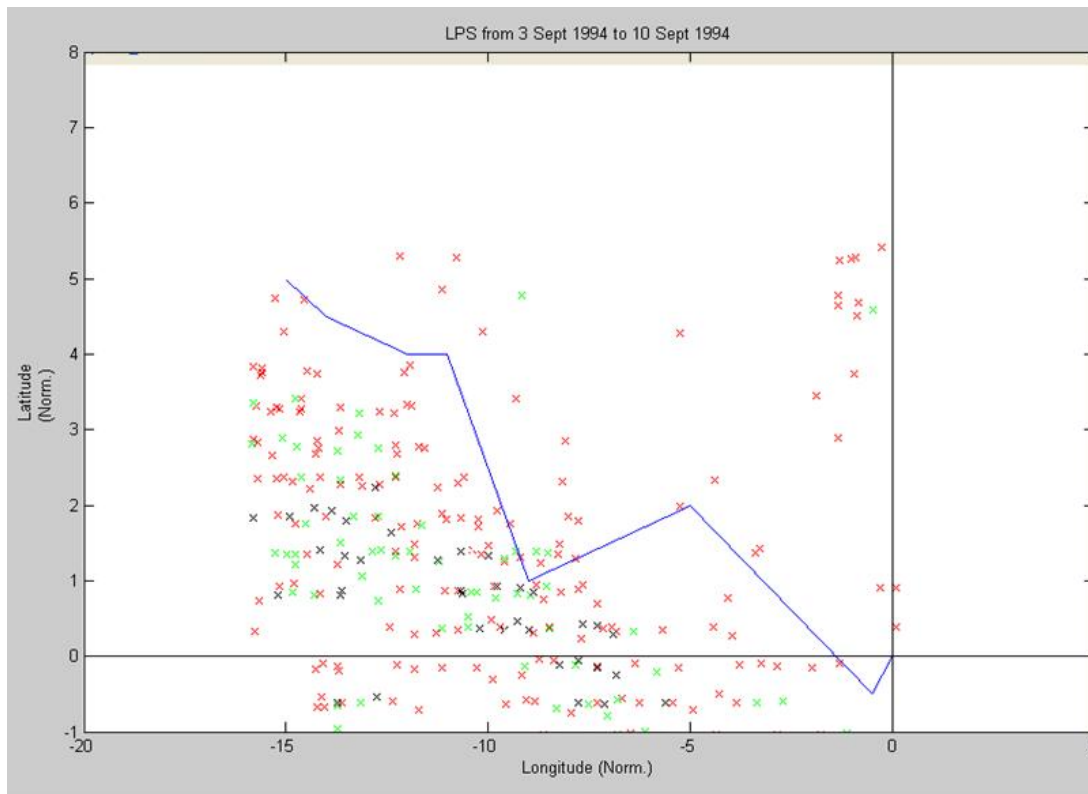


Figure 4.9 – Rainfall with reference to LPS movement from 3Sept 1994 to 10Sept 1994 (Blue line depicts LPS movement) (Red: Good Rainfall, Green: Heavy Rainfall, Black: very/extremely heavy Rainfall)

The compositing of Rainfall for all three cases taken together has also been done and the resulting plot of Rainfall with reference to LPS movement is depicted in Figure 4.10

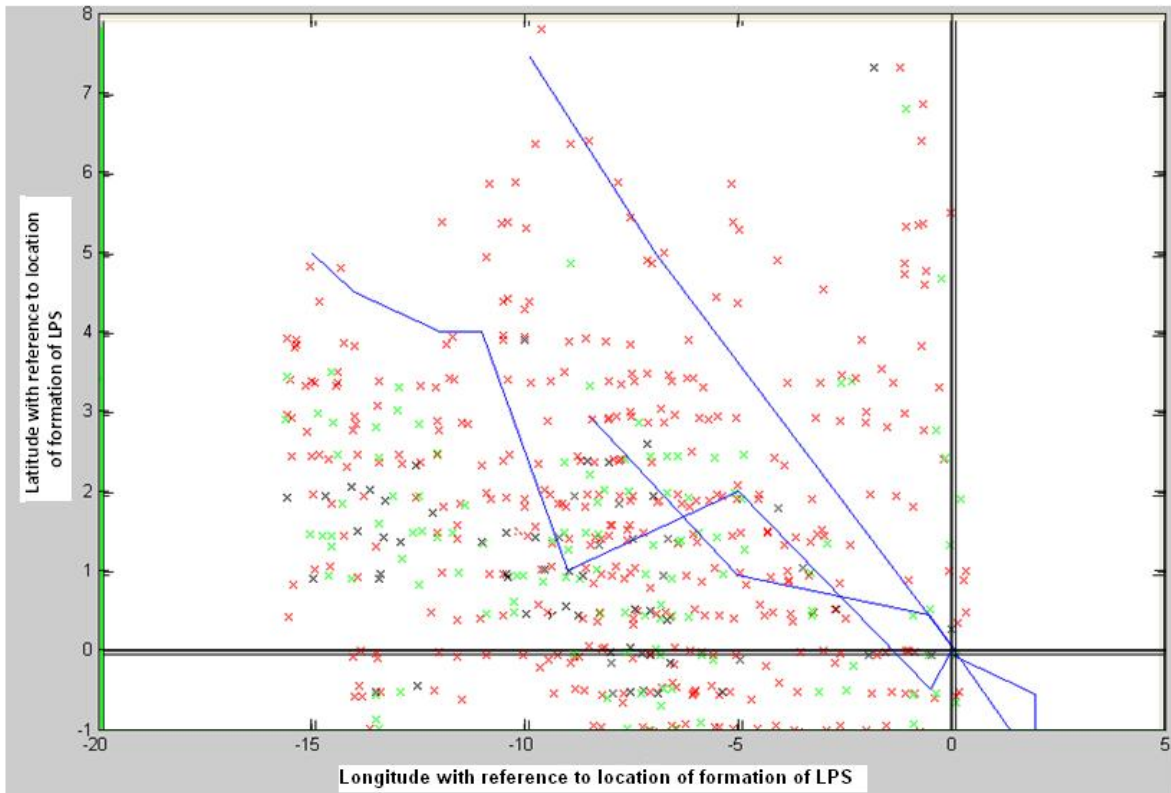


Figure 4.10 – Rainfall with reference to LPS movement from 25June 1994 to 1July 1994, 3July 1994 to 5July 1994 and 3Sept 1994 to 10Sept 1994 (Blue lines depict LPS movement)
 (Red: Good Rainfall, Green: Heavy Rainfall, Black: very/extremely heavy Rainfall)

These results were found acceptable and provided useful information for prediction of areas of heavy rainfall in the wake of movement of LPS.

4.3 Artificial Neural Network for Rainfall forecasting

4.3.1 Datasets used

An effort has been made to make use of ANN for forecasting the rainfall based on previous year's rainfall for the months June to September. The daily rainfall dataset taken into consideration for the training of Neural Network is from longitude 70.5 °E to 90.0 °E and latitude 17.5°N to 37.0°N for the time period June to September for the years 1989 to 1992. The pre-processing of the same has been explained in section 3.2.1.

4.3.2 About Artificial Neural Networks

An ANN is a mathematical model or computational model that is inspired by the structure and/or functional aspects of biological neural networks. A neural network consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation (Sivanandam, Sumathi and Deepa, 2005 and Kosko, 2005). In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase.

4.3.2.1 Network Function

The word '*network*' refers to the inter-connections between the neurons in the different layers of each system. The most basic system has three layers. The first layer has input neurons which send data via synapses to the second layer of neurons and then via more synapses to the third layer of output neurons. More complex systems have more layers of neurons with some having increased layers of input neurons and output neurons. The synapses store parameters called "weights" which are used to manipulate the data in the calculations.

4.3.2.2 Training and testing the network

In an Artificial Neural Network, the system parameters are changed during operation, normally called the training phase. After the training phase, the Artificial Neural Network parameters are fixed and the system is deployed to solve the problem at hand (the testing phase). The Artificial Neural Network is built with a systematic step-by-step procedure to optimize a performance criterion or to follow some implicit internal constraint, which is commonly referred to as the learning rule (Kosko, 2005). The input/output training data are fundamental in neural network technology, because they convey the necessary information to "discover" the optimal operating point. The nonlinear nature of the neural network processing elements (PEs) provides the system with lots of flexibility to achieve practically any desired input/output map, i.e., some Artificial Neural Networks are universal mappers.

An input is presented to the neural network and a corresponding desired or target response set at the output (when this is the case the training is called supervised). An error is composed from the difference between the desired response and the system output. This error information is fed back to the system and adjusts the system parameters in a systematic fashion (the learning rule). The process is repeated until the performance is acceptable. It is clear from this description that the performance hinges heavily on the data. ANN-based solutions are

extremely efficient in terms of development time and resources, and in many difficult problems artificial neural networks provide performance that is difficult to match with other technologies.

MLP Back Propagation Network model which is a supervised network because it requires a desired output in order to learn has been used in the study. The goal of this type of network is to create a model that correctly maps the input to the output using the historical data so that the model then can be used to produce the output when the desired output is unknown.

In this network, shown in Figure 4.11, the input data are fed to input nodes and then they will pass to the hidden nodes after multiplying by a weight. A hidden layer adds up the weighted input received from the input nodes, associates it with the bias and then passes the result on through a nonlinear transfer function. The output node does the same operation as that of a hidden layer. This type of network is preferred as back propagation learning is a popular algorithm to adjust the interconnection weights during training, based upon the generalized delta rule proposed (Kosko, 2005).

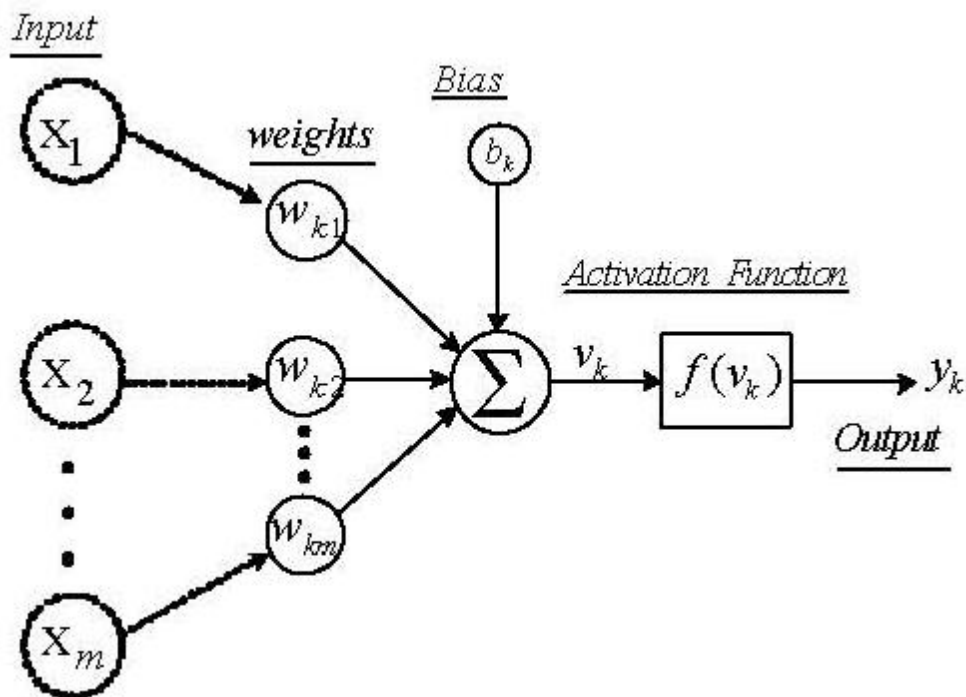


Figure 4.11 – Neuron Model

4.3.3 Application of ANN to forecast Rainfall

ANN in this study was trained and simulated using Matlab 7.0 (matrix laboratory) designed and developed by Math Works Inc. For the training and testing of network, a two layer MLP Back Propagation network has been used. The input dataset comprises of daynumber (day

1 corresponds to June 1, day 2 to June 2 and so on till day number 122 that corresponds to September 30), latitude and longitude. The output data corresponds to rainfall in mm. A sample of dataset is shown in table 4.4. From this table, columns 1 to 3 are used as input and column 4 is used as target.

Day no.	Latitude	Longitude	Rainfall
1	35.5	76.5	6.4
1	35.5	77	6.5
1	35.5	77.5	1
1	35.5	78	7.3
1	35.5	78.5	2.8
1	35	76	7.7
1	35	76.5	7.4
1	35	77	24.2
1	35	77.5	11.7
1	35	78	17.3
1	35	78.5	24.2
1	35	79	2
1	34.5	76	7.1
1	34.5	76.5	10.7
1	34.5	77	6.2
1	34.5	77.5	7.1

Table 4.4 – Sample of location-wise rainfall for year 1989
(Source: as a result of pre-processing rf1989.grd provided by IMD)

Before training, the inputs and outputs have been scaled so that they fall in the range [-1,1]. The following code has been used at Matlab prompt:-

```
[pn1992,minp,maxp,tn1992,mint,maxt] = premnmx(linput,loutput)
```

The original network inputs and targets are given in the matrices linput and loutput. The normalized inputs and targets, pn1992 and tn1992, that are returned, will all fall in the interval [-1,1]. The vectors minp and maxp contain the minimum and maximum values of the original inputs, and the vectors mint and maxt contain the minimum and maximum values of the original targets.

Different transfer functions for hidden and output layers were used to find the best ANN structure for this study. Transfer function used in hidden layer of the back propagation network is tangent-sigmoid while pure linear transfer function is used in output layer.

4.3.4 Learning Algorithms used

ANN developed for prediction of rainfall is trained with different learning algorithms, learning rates, and number of neurons in its hidden layer. The aim is to create a network which

gives an optimum result. The network was simulated using 3 different Back propagation learning algorithms. They are Resilient Backpropagation (*trainrp*), Fletcher-Reeves Conjugate Gradient (*traincgf*) and Scale Conjugate Gradient (*trainscg*).

The Resilient Back propagation (*trainrp*) eliminates the effect of gradient with small magnitude. As magnitudes of the derivative have no effect on the weight update, only the sign of the derivative is used to determine the direction of the weight update. *Trainrp* is generally much faster than standard steepest descent algorithms, and require only a modest increase in memory requirements which suits network with sigmoidal transfer function.

Fletcher-Reeves Conjugate Gradient (*traincgf*) generally converges in lesser iteration than *trainrp*, although there is more computation required in each iteration. The conjugate gradient algorithms are usually much faster than variable learning rate back propagation, and are sometimes faster than *trainrp*. *Traincgf* also require only a little more storage than simpler algorithms, thus they are often a good choice for networks with a large number of weights.

The third algorithm, Scale Conjugate Gradient (*trainscg*) was designed to avoid the time-consuming line search. This differs from other conjugate gradient algorithm which requires a line search at each iteration. The *trainscg* routine may require more iteration to converge, but the number of computations in each iteration is significantly reduced because no line search is performed. *Trainscg* require modest storage.

4.3.5 Findings

Daily rainfall data for 122 days in a year i.e. months June to September were chosen for training and testing. Networks were trained with data of year 1989 and tested using rainfall data of the year 1990. The training has been done using three different training functions as mentioned before: *traincgf*, *trainrp* and *trainscg*. Figure 4.12 to Figure 4.14 demonstrate the result of training with year 1989 dataset and testing with year 1990 datasets. The results are convincing and the network once trained has been tested with year 1990 datasets and the error comes out to be less than 0.005 in 5 epochs for training functions *trainscg* and *traincgf*. With *trainrp* function, it takes 35 iterations to train.

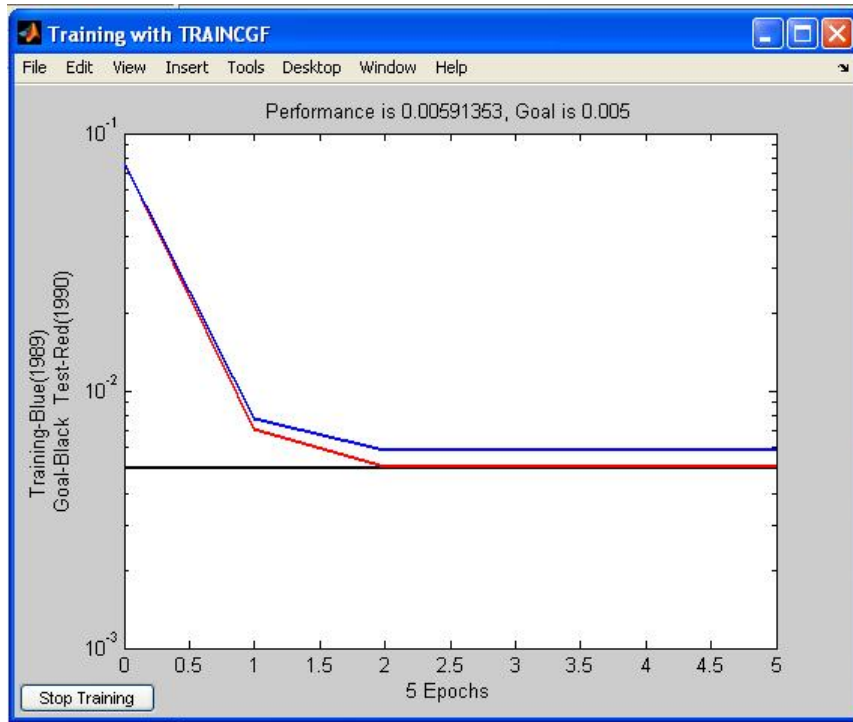


Figure 4.12 – Result of training ANN with Rainfall data of year 1989 and testing with Rainfall data of year 1990 using learning function traincgf

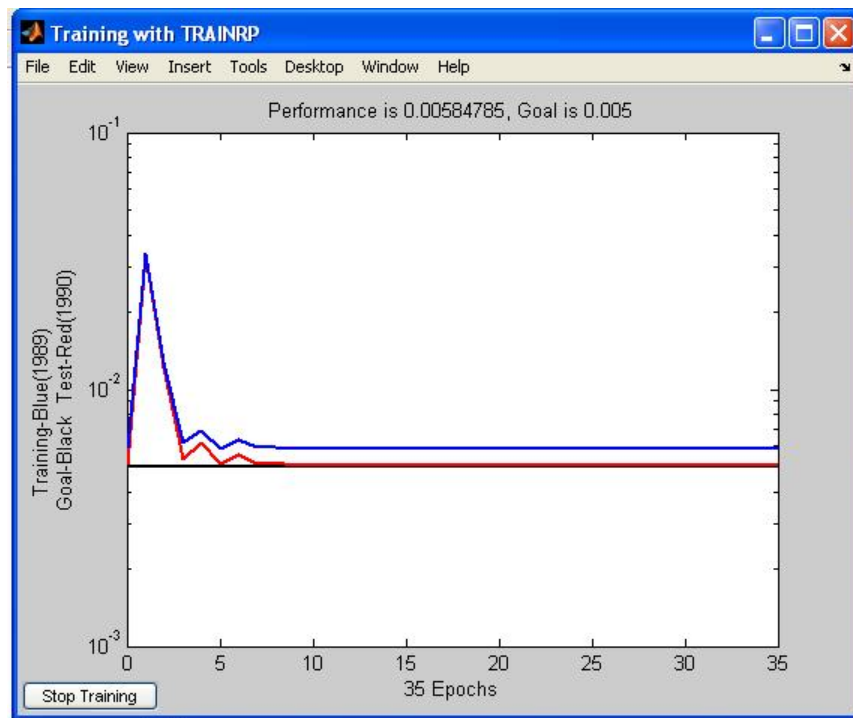


Figure 4.13 – Result of training ANN with Rainfall data of year 1989 and testing with Rainfall data of year 1990 using learning function trainrp

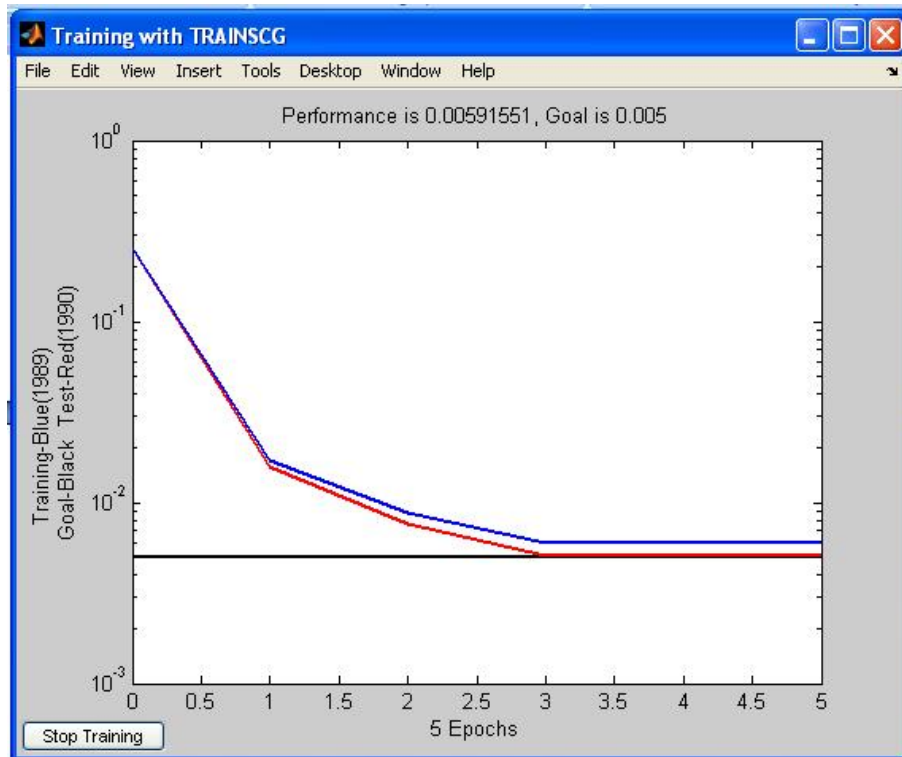


Figure 4.14 – Result of training ANN with Rainfall data of year 1989 and testing with Rainfall data of year 1990 using learning function trainscg

Another rainfall dataset is for the year 1991 and 1992, training with 1991 and testing with 1992. Figure 4.15 to Figure 4.17 demonstrate the result of training with year 1991 dataset and testing with year 1992 datasets. Here again, the results are convincing and the network once trained has been tested with year 1992 datasets and the error comes out to be less than 0.005 in 3 epochs for training functions trainscg and traincgf. With trainrp function, it takes 13 iterations to train.

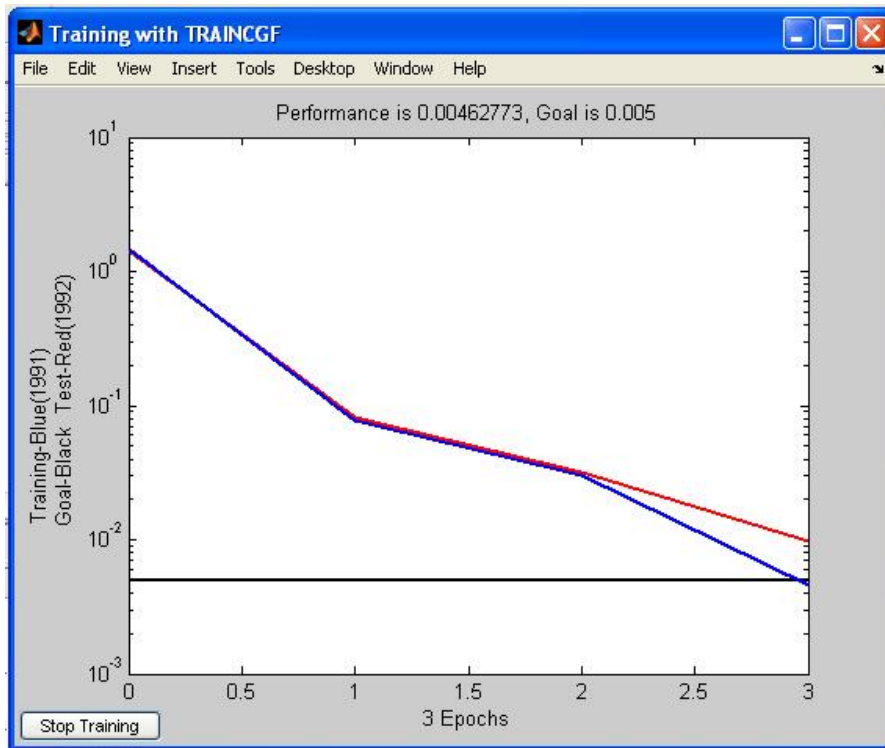


Figure 4.15 – Result of training ANN with Rainfall data of year 1991 and testing with Rainfall data of year 1992 using learning function traincgf

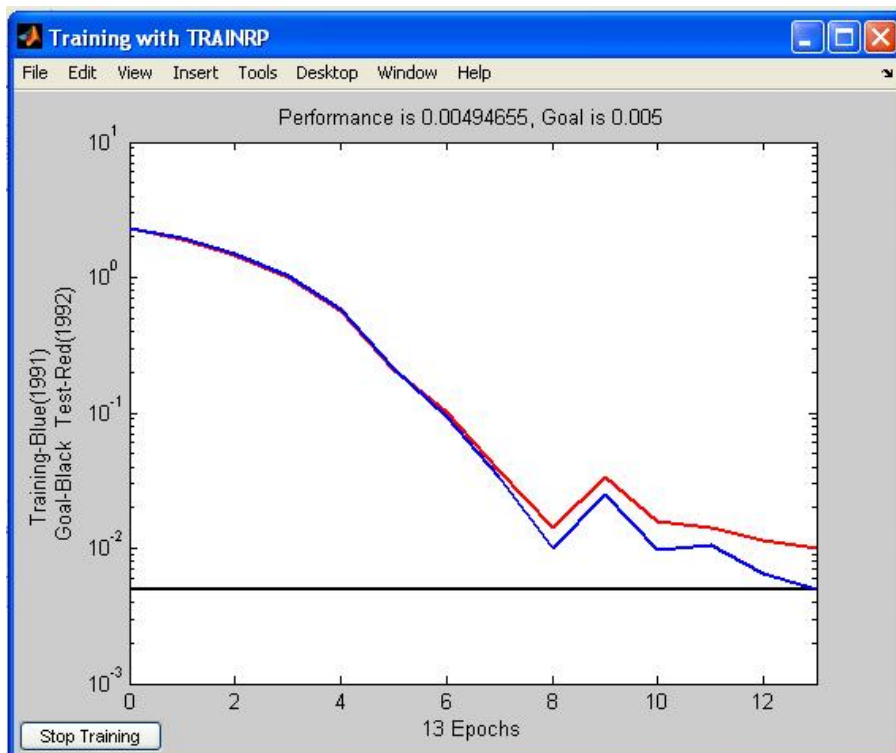


Figure 4.16 – Result of training ANN with Rainfall data of year 1991 and testing with Rainfall data of year 1992 using learning function trainrp

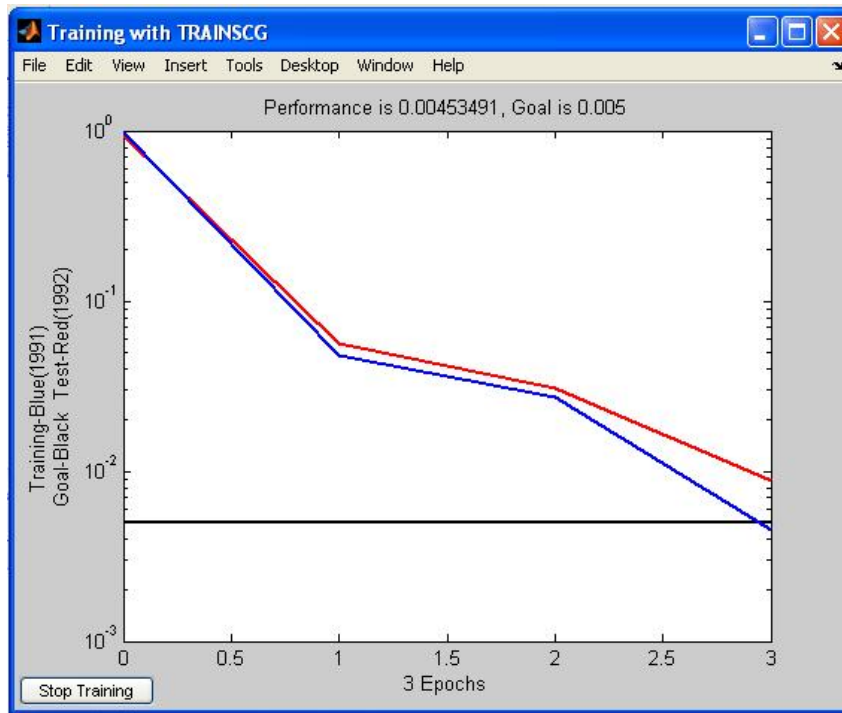


Figure 4.17 – Result of training ANN with Rainfall data of year 1991 and testing with Rainfall data of year 1992 using learning function trainscg

It is found that ANN has demonstrated promising results and is very suitable for solving the problem of rainfall forecasting for synoptic scale systems but it was not clear if ANN could be used for getting advanced warning of occurrence of sub-grid scale weather events from the NWP model output products. Hence ANN method was not pursued further.

4.4 A Multidimensional data model for Meteorological datasets

As explained in section 3.4, MDDM enables efficient access of facts across multiple dimensions in comparison to relational databases, so in order to analyze rainfall with reference to a particular time period of a year and in a selected geographical region because of movement of LPS system, a 3-Dimensional data cube was implemented.

4.4.1 Datasets used

The dataset corresponding to daily rainfall data for each year for the period 1984–2003 has been used. The data is for the geographical region from longitude 66.5 °E to 100.5 °E and latitude 6.5 °N to 38.5 °N for each day of the year, provided by IMD and pre-processed in section 3.2.1. The LPS datasets as discussed in section 4.1 has also been used.

4.4.2 Implementation of the 3-Dimensional data cube

For implementation of 3-D cube, the data has been restructured so as to adapt itself to Multidimensional OLAP environment. This cube has three dimensions – Location, Time and Low Pressure System. The fact/ measure is Rainfall. Three tables corresponding to three dimensions are created as illustrated in Pabreja (2010a). Each dimension table contains a primary key and a set of attributes. The primary keys are time_key as primary key for Time dimension, loc_key as primary key for Location dimension, (LPSno + Date of LPS) as primary key for Low Pressure System dimension. The schema contains a central fact table for Rainfall that contains foreign keys to each of the three dimensions, along with measure for Rainfall. The sample data values of dimension tables namely Time, Location, Low Pressure Systems and central fact table Rainfall are shown in Table 4.5, 4.6, 4.7 and 4.8 respectively.

Microsoft SQL Server Business Intelligence Development Studio (Vatuiu and Popeanga, 2007) has been used to implement the OLAP model for the mentioned datasets. It offers a Graphical User Interface (GUI) based interface for creating dimension and fact tables for the cube (Harinath and Carroll, 2009).

The multidimensional data of Rainfall and LPS over Indian region for years 1984-2003 has been implemented as a star schema, shown in Figure 4.18. The OLAP operations Roll-up (climbing up on a concept hierarchy for a dimension) and drill-down (stepping down a concept hierarchy), slice and dice have been performed on the cube.

time_key	daynumber	date
1	1	1-Jun-84
2	2	2-Jun-84
3	3	3-Jun-84
4	4	4-Jun-84
5	5	5-Jun-84
6	6	6-Jun-84
7	7	7-Jun-84
8	8	8-Jun-84
9	9	9-Jun-84
10	10	10-Jun-84
11	11	11-Jun-84
12	12	12-Jun-84
13	13	13-Jun-84
14	14	14-Jun-84
123	1	1-Jun-85
124	2	2-Jun-85

Table 4.5 – A sample dataset of dimension table “time”
(Source: as a result of pre-processing LPS data from Sikka (2006))

loc_key	latitude	longitude
1	38.5	66.5
2	38.5	67
3	38.5	67.5
4	38.5	68
5	38.5	68.5
6	38.5	69
7	38.5	69.5
8	38.5	70
9	38.5	70.5
10	38.5	71
11	38.5	71.5
12	38.5	72
13	38.5	72.5
14	38.5	73

Table 4.6 – A sample dataset of dimension table “location”
 (Source: As a result of pre-processing rf1984.grd file provided by IMD)

lpsno	date of lps	Day number	intensity	time interval	loc_key
1	2-Jun-84	2	1	1	1726
1	3-Jun-84	3	1	2	1824
1	4-Jun-84	4	1	3	1627
1	5-Jun-84	5	1	4	1529
1	6-Jun-84	6	1	5	1481
2	5-Jun-84	5	1	1	1923
2	6-Jun-84	6	1	2	1778
2	7-Jun-84	7	1	3	1826
2	8-Jun-84	8	1	4	1825
3	9-Jun-84	9	1	1	1185
3	10-Jun-84	10	1	2	1186
3	11-Jun-84	11	1	3	1138

Table 4.7 – A sample dataset of dimension table “low pressure system”
 (Source: As a result of pre-processing rf1984.grd file provided by IMD)

loc_key	time_key	Rainfall (in mm)	lpsno	date of lps
1624	4	54.9	no lps	no lps
1625	4	29	no lps	no lps
1626	4	41	no lps	no lps
1627	4	62.4	1	4-Jun-84
1628	4	76.1	no lps	no lps
1629	4	56.5	no lps	no lps
1528	5	17.6	no lps	no lps
1529	5	25.2	1	5-Jun-84
1530	5	56.5	no lps	no lps

Table 4.8 – A sample dataset of central fact table “fact_rainfall” for 3-Dimensional data model for meteorological datasets (Source: As a result of pre-processing rf1984.grd file provided by IMD and LPS datasets provided by Sikka,2006)

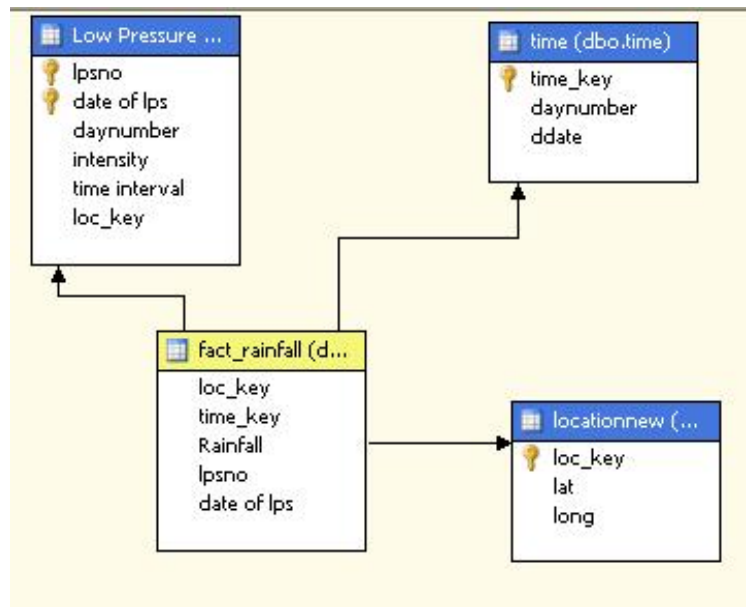


Figure 4.18 – Star Schema for Rainfall corresponding to Tables 4.5, 4.6, 4.7 and 4.8

4.4.3 Findings

This MDDM has been browsed to view rainfall value against LPS vs. time vs. location dimensions as shown in Figure 4.19. SQL server offers facility to apply any filter based on any condition on one/more attributes of dimensions of the cube.

The cube has been browsed to view aggregate rainfall at all locations separately, month-wise. Here aggregate rainfall for June and July months of 1984 at 4485 different grid points has

been shown in Figure 4.20. The analysis of rainfall associated with LPS formation, by rolling up time dimension to year level, has also been done and further drilling down the time dimension to months have also been done. Any number of additional dimensions can be added to the cube to make it a Hypercube which has been done in next section 4.5.

Dimension	Hierarchy	Operator	Filter Expr
<Select dimension>			
Drop Filter Fields Here			
	Ddate ▾		
	1984-06-23 00:00:00	1984-06-24 00:00:00	1984-06-25 00:00:00
Daynumber ▾	Rainfall	Rainfall	Rainfall
22			
23	75.3		
24		13.6	
25			17.5
26			
27			
28			
29			
30			

Figure 4.19 – Analysis of Aggregate Rainfall against date and LPS

Dimension	Hierarchy	Operator	Filter Expression
Time	Daynumber	Range (Inclus...	1 : 61
<Select dim...			
Drop Filter Fields Here			
	Long ▾		
	75.5	76	76.5
	77	77.5	78
Lat ▾	Rainfall	Rainfall	Rainfall
15	74.2	170.5	298.8
15.5	22.5	123.5	38.2
16	105.8	106	273.7
16.5	122.2	113.4	158.2
17	132.9	142.5	104.6
17.5	162.5	137.3	234.1
18	115	154.6	186.3
18.5	82.2	77	257.4
19			

Figure 4.20 – Aggregate rainfall for June and July months of 1984

4.5 A 5-dimensional data model for Meteorological datasets

In order to understand the impact of an event like cloudburst in a district/state or a river catchment area in a particular time period, the addition of two more dimensions to the cube has been done, making it a 5-D cube (Pabreja 2010b and 2010c).

4.5.1 Datasets used

In addition to datasets used in section 4.4.1, the datasets corresponding to river catchment area and district with state has been added. The river catchment area has been described by considering river catchment area to be a polygon with the shape being described by latitude, longitude values. Similarly districts are described using polygon shapes. In order to find the presence of different gridded location points inside or outside of river catchment area or district, Point in Polygon algorithm has been implemented in Visual Basic as shown in Appendix G.

4.5.2 Implementation of 5-Dimensional data cube

In order to analyze the rainfall across five different dimensions – Gridded location, Time, LPS formation, River Catchment area and District, a 5 dimensional data cube has been implemented. That is the rainfall during a particular time period of a year, in a particular district or districts, in a particular river catchment area or areas because of movement of LPS system (LPS data from Mooley and Shukla, 1987; Sikka, 2006) can be analyzed. For this purpose, the data has been restructured so as to adapt itself to Multidimensional OLAP environment. Three dimension tables are same as in section 4.4.2 and for two more dimensions- river catchment area and district, new tables are created. For these two dimensions – River catchment area and district are normalized tables represented by two tables per dimension. Here, the primary keys are riverid for river table and districtid for district table. Sample datasets for the tables *viz.* river, riverdetails, district, district details are shown in Tables 4.9 to Table 4.12 respectively. A 5-Dimensional cube for the data to be examined is shown in Figure 4.21.

Snowflake schema has been used to model the dimensions datasets and the fact dataset as shown in Figure 4.22, using Microsoft SQL Server Business Intelligence Development Studio. The schema contains a central fact table for Rainfall that contains foreign keys to each of the five dimensions' primary keys, along with measure for Rainfall. The primary keys are time_key as primary key for Time dimension, loc_key as primary key for Location dimension, (LPSno + Date of LPS) as primary key for Low Pressure System dimension, riverid for river table and districtid for district table.

riverid	rivername
1	ganga
2	kosi
3	yamuna

Table 4.9 – A sample dataset of dimension table “river”

riverid	longitude	latitude
1	86.8	22.8
1	84.4	23.3
1	88	23.7
1	81.8	23.8
1	86.7	24
1	80.8	24.8
1	85.8	24.9
1	84.9	25.6
1	79.7	26
1	83.8	26.2
1	82.4	26.8
1	79.1	26.85
1	78.5	27.7
1	80.1	28.6
1	77.6	28.8

Table 4.10 – A sample dataset of dimension table “river details”

districtid	districtname	state
1	Agra	Uttar Pradesh
2	Varanasi	Uttar Pradesh
3	Lucknow	Uttar Pradesh
4	Gaya	Bihar
5	Kishanganj	Bihar
6	Kathihar	Bihar
7	Naya Dumka	Bihar

Table 4.11 – A sample dataset of dimension table “district”

districtid	longitude	latitude
1	77.9	26.8
1	78.2	27
1	77.8	27.2
1	78.3	27.5
1	78	27.6
2	83.3	24.6
2	82.8	24.7
2	82.7	24.8
2	83.5	25
2	83.4	25.2
2	83	25.6
3	80.1	25.8
3	80.3	26
3	79.6	26.2

Table 4.12 – A sample dataset of dimension table “district details”

loc_key	time_key	Rainfall	Lpsno	date of lps	districtid	riverid
1128	37	1.7	14	10-Sep-84	1	3
1128	36	3.8			1	3
1080	79	0	14	14-Sep-84	1	3
1080	36	1			1	3
1128	38	0.9			1	3
1080	40	0.3	14	18-Sep-84	1	3
1128	40	0.4			1	3
1128	43	4			1	3
1080	42	1.5	12	29-Aug-84	1	3
1128	42	3.6			1	3
1080	44	25.2			1	3
1080	43	1.8			1	3

Table 4.13 – A sample dataset of central fact table “fact_rainfall” for 5-dimensional data model for meteorological datasets (Source: As a result of pre-processing rf1984.grd file provided by IMD and LPS datasets provided by Sikka(2006))

This 5-D cube comprises of a total of 32 cuboids (2^5) and search of a particular cuboid has been represented in Figure 3.5 where each node corresponds to a unique cuboid. Statistically the search of a node would save a lot of time in comparison of flat 2-D data storage in Relational databases.

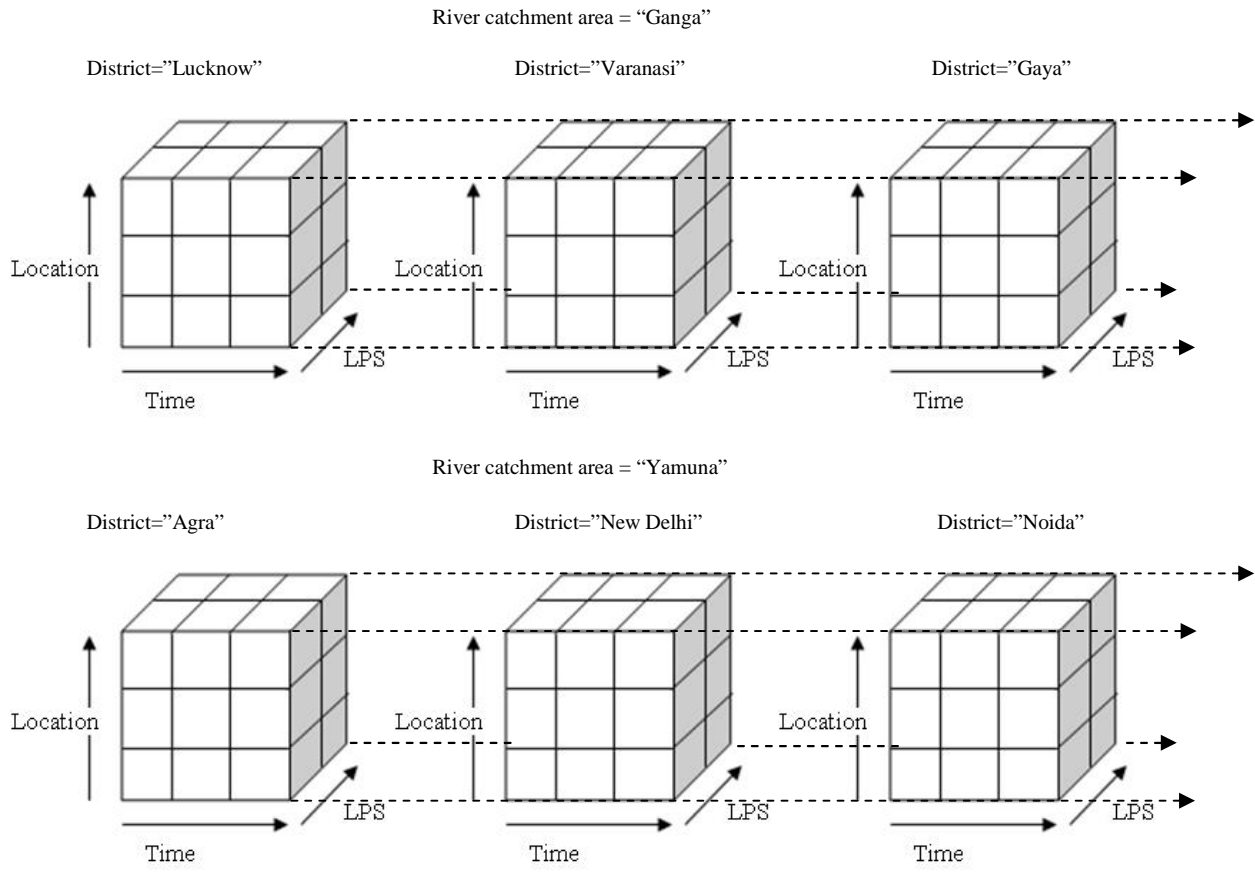


Figure 4.21 – A 5-D data cube representation of rainfall data, according to dimensions time, gridded-location, Low Pressure system, river catchment area and district

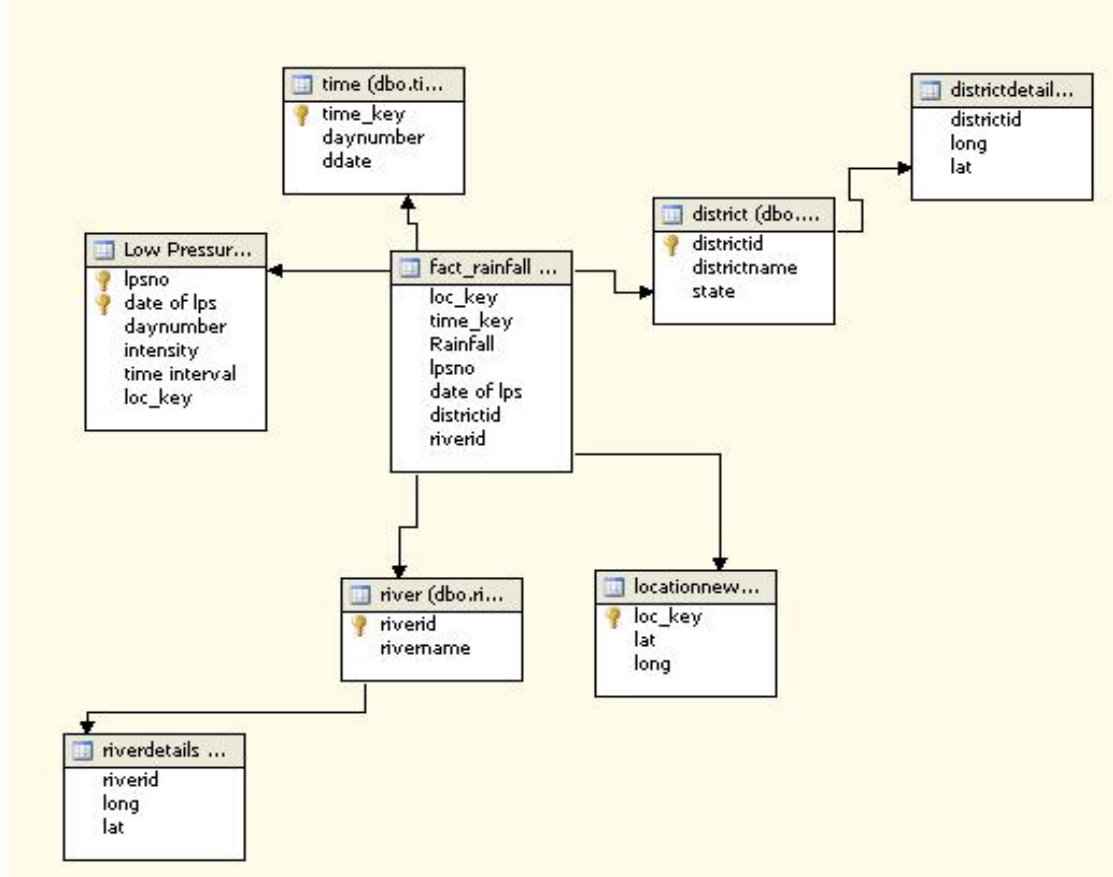


Figure 4.22 – Snowflake Schema for Rainfall corresponding to Table 4.5, 4.6, 4.7, 4.9, 4.10, 4.11, 4.12, 4.13

4.5.3 Findings

This MDDM has been browsed to view rainfall value against LPS vs. time vs. location dimensions as shown in Figure 4.23. SQL server offers facility to apply any filter based on any condition on one/more attributes of dimensions of the cube. The cube has been browsed to view aggregate rainfall at all locations separately, month-wise. Here aggregate rainfall for June and July months of 1984 at 4485 different grid points has been shown in Figure 4.24.

The cube has also been browsed to view aggregate rainfall falling in different river catchment areas against various districts, month-wise. Here aggregate rainfall for June and July months of 1984 at 4485 different grid points has been shown in Figure 4.25. These results serve as input to Data mining tools and depict efficient mode of pre-processing of meteorological datasets.

Dimension	Hierarchy	Operator	Filter Expr
<Select dimension>			
Drop Filter Fields Here			
	Ddate		
	1984-06-23 00:00:00	1984-06-24 00:00:00	1984-06-25 00:00:00
Daynumber	Rainfall	Rainfall	Rainfall
22			
23	75.3		
24		13.6	
25			17.5
26			
27			
28			
29			
30			

Figure 4.23 – Analysis of Aggregate Rainfall against date and LPS

Dimension	Hierarchy	Operator	Filter Expression
Time	Daynumber	Range (Inclus...	1 : 61
<Select dim...			
Drop Filter Fields Here			
	Long		
	75.5	76	76.5
	77	77.5	78
Lat	Rainfall	Rainfall	Rainfall
15	74.2	170.5	298.8
15.5	22.5	123.5	38.2
16	105.8	106	273.7
16.5	122.2	113.4	158.2
17	132.9	142.5	104.6
17.5	162.5	137.3	234.1
18	115	154.6	186.3
18.5	82.2	72	257.4
19			

Figure 4.24 – Aggregate rainfall for June and July months of 1984 against longitude and latitude

Dimension	Hierarchy	Operator	Filter Expression
Time	Daynumber	Range (Inclus...	: 61
<Select dimension>			

Time	Districtname	Rainfall	Rainfall
All			
	Agra	Gaya	Lucknow/Nava Dumka
			Varanasi
			Grand Total
Rivername	Rain	Districtname (Districtname)	Rainfall
ganga	363	395.8	728.9
yamuna	664		
Grand Total	664	363	395.8

Figure 4.25 – Aggregate rainfall for June and July months of 1984 against River catchment area and district

4.6 Application of Multidimensional data model for NWP model output products for generating ensembles

The MDDM was used in section 4.4 and 4.5 for storage, retrieval and analysis of rainfall across three and five dimensions respectively and the results were very convincing so the MDDM was further utilized for storing the colossal outputs of NWP model forecasts after lot of pre-processing steps as these files are in a format that a database management system cannot interpret.

4.6.1 Datasets used

The output product of ECMWF model has been provided with courtesy of IMD. Forecast data corresponding to various cases of cloudburst over Indian region, during year 2009 have been considered. ECMWF model outputs for forecast of various weather variables for 12 hour, 36 hour, 60 hour, and 84 hour at various atmospheric pressure levels viz. 1000hPa, 925hPa, 850hPa, 700hPa, 500hPa, 400hPa and 300hPa, valid for the following dates of cloudburst have been considered.

1. 18 July, 2009, Chamoli district of Uttarakhand
2. 29 July 2009, Dhaka
3. 7 August 2009, Shimla, Himachal Pradesh
4. 8 August 2009, Pittorgarh district of Uttarakhand

For a particular case of cloudburst that occurred on 8 August 2009 at Pittorgarh district of Uttarakhand, the creation of ensemble of forecast valid for 1200GMT 8August 2009 with different initial conditions of NWP model is depicted in Figure 4.26.

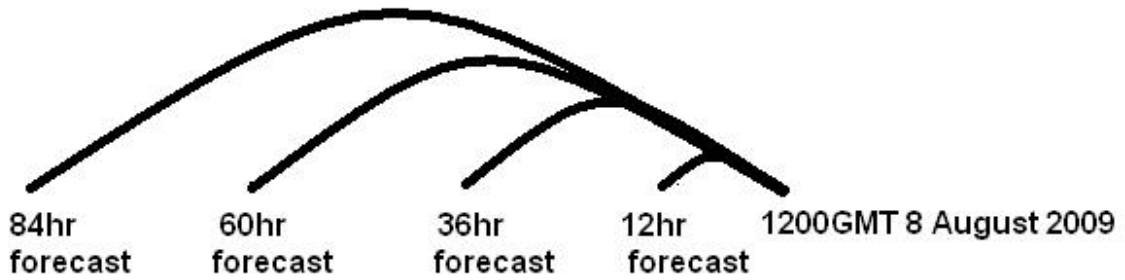


Figure 4.26 – Depicting the creation of ensemble of forecast valid for 1200GMT 8 August 2009 with different initial conditions

The output product of the model after pre-processing to derive convergence and vorticity (as explained in section 3.2.2.1), at atmospheric pressure levels 1000hPa, 925hPa, 850hPa, 700hPa, 500hPa, 400hPa and 300hPa are stored in the data cube.

The data hence obtained is restructured so that the dimensions of time and location are in two separate tables and the facts – vorticity and divergence forecast valid for 0600GMT and 1200GMT can be analyzed for any particular case of cloudburst that affected a particular location ($2.5^{\circ} \times 2.5^{\circ}$ window surrounding the location of occurrence of cloudburst.). The creation of ensemble to be presented to data mining tool has been facilitated which otherwise is very time consuming and hence not efficient when flat 2-D tables are used for storage.

4.6.2 Implementation of Multidimensional data model

The analysis of vorticity and divergence across two different dimensions – Location and time (Pabreja, 2010d) has been done. For this purpose, the data has been restructured so as to adapt itself to Multidimensional OLAP environment. Two tables corresponding to two dimensions are created. Each dimension table contains a primary key and other attributes. The primary keys are `input_time_key` as primary key for Time dimension, `loc_key` as primary key for Location dimension.

The schema contains a central fact table for vorticity and divergence that contains foreign keys to each of the two dimensions. The sample data values of dimension tables namely Location, Time and central fact table are shown in Table 4.14, 4.15 and 4.16 respectively.

The multidimensional data of ECMWF forecast over Indian region for year 2009 has been implemented as a star schema, shown in Figure 4.27. The OLAP operations Roll-up (climbing up on a concept hierarchy for a dimension) and drill-down (stepping down a concept hierarchy), slice and dice have been performed on the cube.

final locdim			
loc_key	latitude	longitude	Atmospheric Pressure level
38788	30	106.75	300
38789	30	107	300
38790	30	107.25	300
38791	30	107.5	300
38792	30	107.75	300
38793	30	108	300
38794	30	108.25	300
38795	30	108.5	300
38796	30	108.75	300
38797	30	109	300
38798	30	109.25	300
38799	30	109.5	300
38800	30	109.75	300
38801	30	110	300
38802	30.25	50	300
38803	30.25	50.25	300
38804	30.25	50.5	300
38805	30.25	50.75	300
38806	30.25	51	300
38807	30.25	51.25	300
38808	30.25	51.5	300
38809	30.25	51.75	300
38810	30.25	52	300

Table 4.14 – A sample of “location” table corresponding to NWP output products
(Source: as a result of pre-processing ECMWF output file in .grib format provided by IMD)

final timedim		
input_time_key	date (mm/dd/yy) on which forecast was made	time at which forecast was made
1	8/5/2009	0000GMT
2	8/6/2009	0000GMT
3	8/7/2009	0000GMT
4	8/8/2009	0000GMT

Table 4.15 – A sample of “time” table corresponding to NWP output products

forecast of vorticity valid for 0600GMT 8Aug 09 (X10 ⁻⁵ per second)	forecast of divergence valid for 0600GMT 8Aug 09 (X10 ⁻⁵ per second)	forecast of vorticity valid for 1200GMT 8Aug 09 (X10 ⁻⁵ per second)	forecast of divergence valid for 1200GMT 8Aug 09 (X10 ⁻⁵ per second)	input_time_key	loc_key
6	-2	8	-10	4	98926
2	0	6	-10	1	98930
2	-2	4	-10	4	98691
2	0	2	-10	4	98933
6	0	2	-10	3	98204
4	-2	2	-10	1	99172
2	0	0	-10	2	98697
4	-8	0	-10	3	97964
4	-4	14	-8	4	98686
8	-4	10	-8	4	98685
2	-2	10	-8	4	98445
4	-4	8	-8	4	99174

Table 4.16 – A sample of central fact table for 2-dimensional data model for NWP model forecasts (Source: as a result of pre-processing ECMWF output files in .grib format provided by IMD)

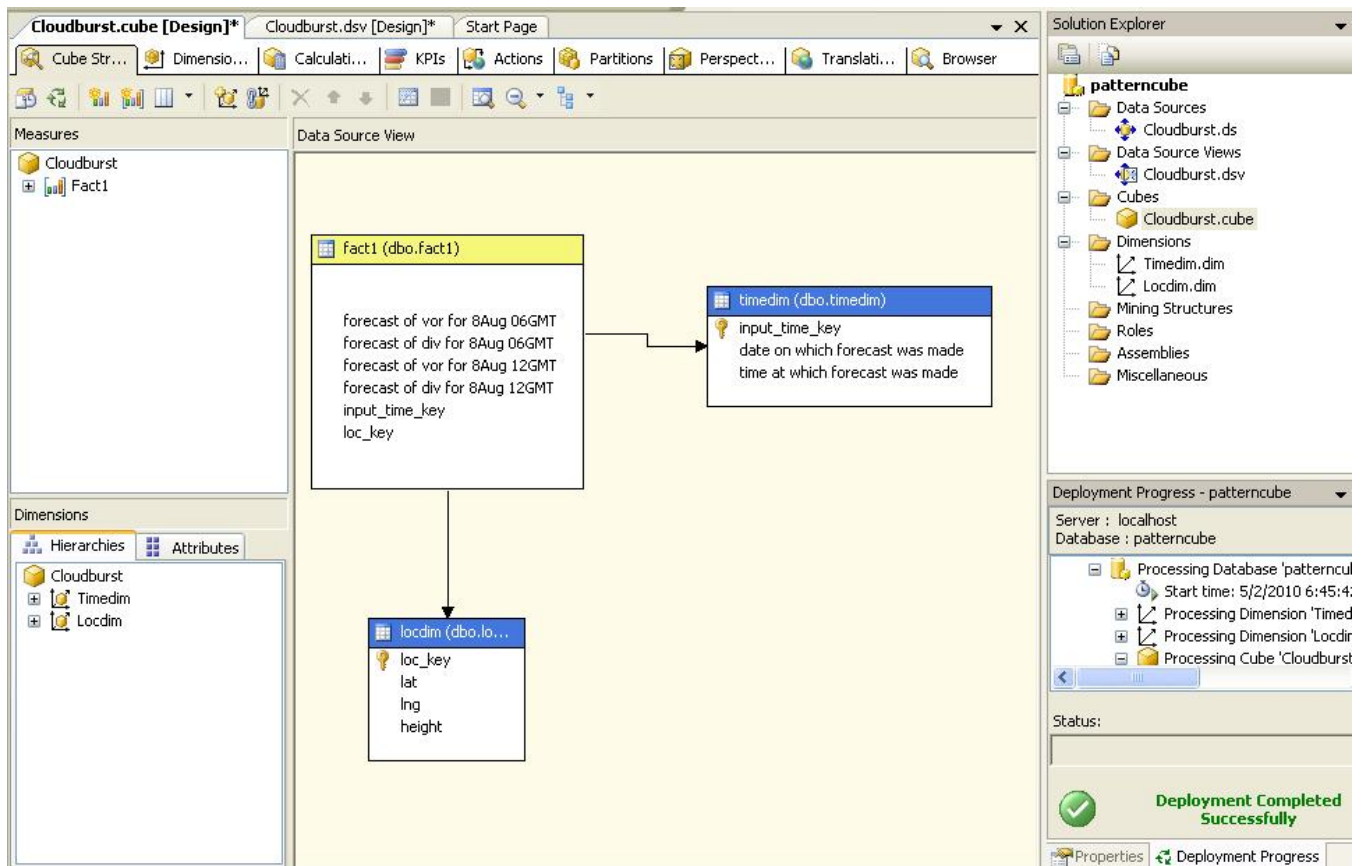


Figure 4.27 – Star Schema for datasets corresponding to Table 4.14, 4.15 and 4.16

4.6.3 Findings

This MDDM can be browsed to view vorticity and divergence forecast valid for 0600GMT/ 1200GMT against time and location dimensions as shown in Figure 4.28 and Figure 4.29 respectively. SQL server offers facility to apply any filter based on any condition on one/more attributes of dimensions of the cube. Any number of additional dimensions can be added to the cube to make it a Hypercube. Thus the ingredients to formation of a sub-grid scale weather event has can be selected based on location and time and fed to Data Mining tool.

Dimension	Hierarchy	Operator	Filter Expression
Timedim	Date On Which Fore...	Equal	{ 2009-08-08 ... }
Locdim	Lng	Range (I...	80.5 : 82.5
Locdim	Height	Equal	{ 300 }

Date On Which Forecast Was Made				
All				
	Lng			
	80.5	80.75	81	81.25
Lat	Forecast Of Vor For 8	Forecast Of Vor For 8	Forecast Of Vor For 8	Forecast Of Vor F
27.5	-6	-8	-8	-6
27.75	-4	-6	-4	-6
28	-2	-2	-2	-4
28.25	0	0	-2	0
28.5	-4	-4	-2	0
28.75	-8	-6	-6	-2
29	-4	-4	-6	-6
29.25	-6	-6	-4	-6
29.5	-4	-6	-6	-4
29.75	-4	-8	-8	-8
30	-10	-4	-4	-6
30.25	-4	-4	-4	-2
30.5	-6	-8	-6	-4
30.75	-4	-6	-6	-8
31	-4	-4	-6	-4
Grand Total	-70	-76	-74	-66

Figure 4.28 – Analysis of vorticity against date and Location

Dimension	Hierarchy	Operator	Filter Expression
Timedim	Date On Which Fore...	Equal	{ 2009-08-07 00:(
Locdim	Lng	Range (Inclus...	80 : 82
Locdim	Height	Equal	{ 400 }

Date On Which Forecast Was Made			
All			
	Lng		
	80	80.25	80.5
Lat	Forecast Of Div For 8 Aug 12GMT	Forecast Of Div For 8 Aug 12GMT	Forecast Of Div F
27.5	0	2	2
27.75	-2	-2	2
28	0	0	0
28.25	2	-2	-4
28.5	0	0	0
28.75	2	2	0
29	-4	-4	0
29.25	-6	-6	-2
29.5	0	-2	-6
29.75	0	-10	-6
30	-4	-6	-4
30.25	-4	-4	-10
30.5	-10	-4	0
30.75	2	4	0
31	8	-2	0
31.25	8	6	14
31.5	-6	18	30
31.75	-4	4	6
32	-6	-16	-8
32.25	-4	-4	-10
32.5	2	2	-10
Grand Total	-26	-24	-6

Figure 4.29 – Analysis of divergence against date and Location

4.7 Data mining for Interpretation of sub-grid scale weather system –“Tornado” using NWP output -

A Case Study of Tornado in Orissa on 31st March 2009

A tornado accompanied with wind speed of about 250 kmph, thunderstorm, rainfall and hailstorm affected Rajakanika block of Kendrapara district of Orissa in the afternoon of 31st March 2009 (Tornado over Orissa on 31st March 2009 :A preliminary report by IMD). It caused loss of about 15 human lives and left several injured, apart from causing huge loss of properties. The special features of this tornado are as follows.

- (i) It was embedded in a thunder squall line.
- (ii) According to doppler weather radar (DWR), Kolkata, the convective cloud cluster moved from north-northwest to south-southeast during its genesis, growth and maturation stage like a Nor'wester. It is almost supported by the orientation of the area affected by the tornado.
- (iii) The maximum surface wind speed was about 250 kmph.
- (iv) The convection developed after 1430 hrs IST and dissipated over the sea by 1730 hrs IST with a life period of less than three hours. The tornado survived for about 10 minutes, hitting ground at 1640 hrs IST, according to eye witnesses.

This real-life tornado has been analyzed with two different NWP models' output products viz. ECMWF and WRF as explained in following sub-sections.

4.7.1 Datasets of ECMWF model under analysis

With the courtesy of IMD, various NWP model outputs based on weather variables input on 0000GMT 29 March 2009, 0000GMT 30 March 2009 and 0000GMT 31 March 2009 for forecast valid for 0600 GMT 31 March 2009 and 1200GMT 31 March 2009 have been obtained. IMD has provided ECMWF, JMA and GFS models outputs. Of all these, ECMWF T-799 model output is of the highest resolution i.e. 0.25° X 0.25° (latitude X longitude) (approx. 25km X 25km) and has reputation of best skill.

The model output has been pre-processed as explained in section 3.2.2.1. For this case of tornado, the location was Rajakanika block of Kendrapara district of Orissa which is at 86° East longitude and 20° North latitude so the convergence and vorticity have been considered for a window of 2.5° around the sides i.e. the area 83.5° longitude to 88.5° longitude and 17.5 ° latitude to 22.5 ° latitude. There are no missing values in the datasets.

We have tried to pick up updraft of air mass that is primary cause of formation of tornado, represented by convergence in the dataset. Hence clustering technique has been applied for feature viz. convergence. This field has positive and negative values depicting downdraft and updraft of air mass. So, the number of clusters has been taken as equal to two.

4.7.2 Visualization and interpretation of clusters of convergence

The derived value of convergence based on weather variables input as on 0000GMT 29 March 2009, 0000GMT 30 March 2009 and 0000GMT 31 March 2009 to ECMWF model for the forecast valid for 0600GMT on 31 March 2009 and 1200GMT on 31 March 2009 have been considered for analysis. The ensemble of convergence for 54hr forecast made based on 0000GMT 29 March 09 weather variables, 30hr forecast made based on 0000GMT 30 March 09 and 6hr forecast based on 0000GMT 31 March 09, valid for 0600GMT 31 March 09 is created. Similarly, the ensemble of convergence for 60hr forecast made based on 0000GMT 29 March 09 weather variables, 36hr forecast made based on 0000GMT 30 March 09 and 12hr forecast based on 0000GMT 31 March 09, valid for 1200GMT 31 March 09 is created.

The visualization of cluster of 0600GMT 31 March 09 does not depict the features conducive to formation of tornado but visualization of cluster of 1200GMT 31 March 09 illustrates the vertical wind motion in the area surrounding the location of tornado.

The clusters of convergence (green and orange points) and divergence (black points) for forecast on 1200GMT 31 March 09 have been plotted as shown in Figure 4.30. The clusters indicate a very large vertical motion field based on every forecast from 29 March 09 onwards, up to atmospheric pressure level 700hPa. The presence of strong vertical motion field depicts that pattern conducive to formation of tornado is there and a broad interpretation can be made. From the 60hr forecast itself, the indication is provided which starts becoming clearer as we move forward to 36hr and 12hr forecast. Although a forecast made with the help of radar reflectivity indicates hook like image, but it is only a few minutes advance warning, whereas with this technique even 3-4days in advance one can observe that some signal conducive to formation of tornado is developing.

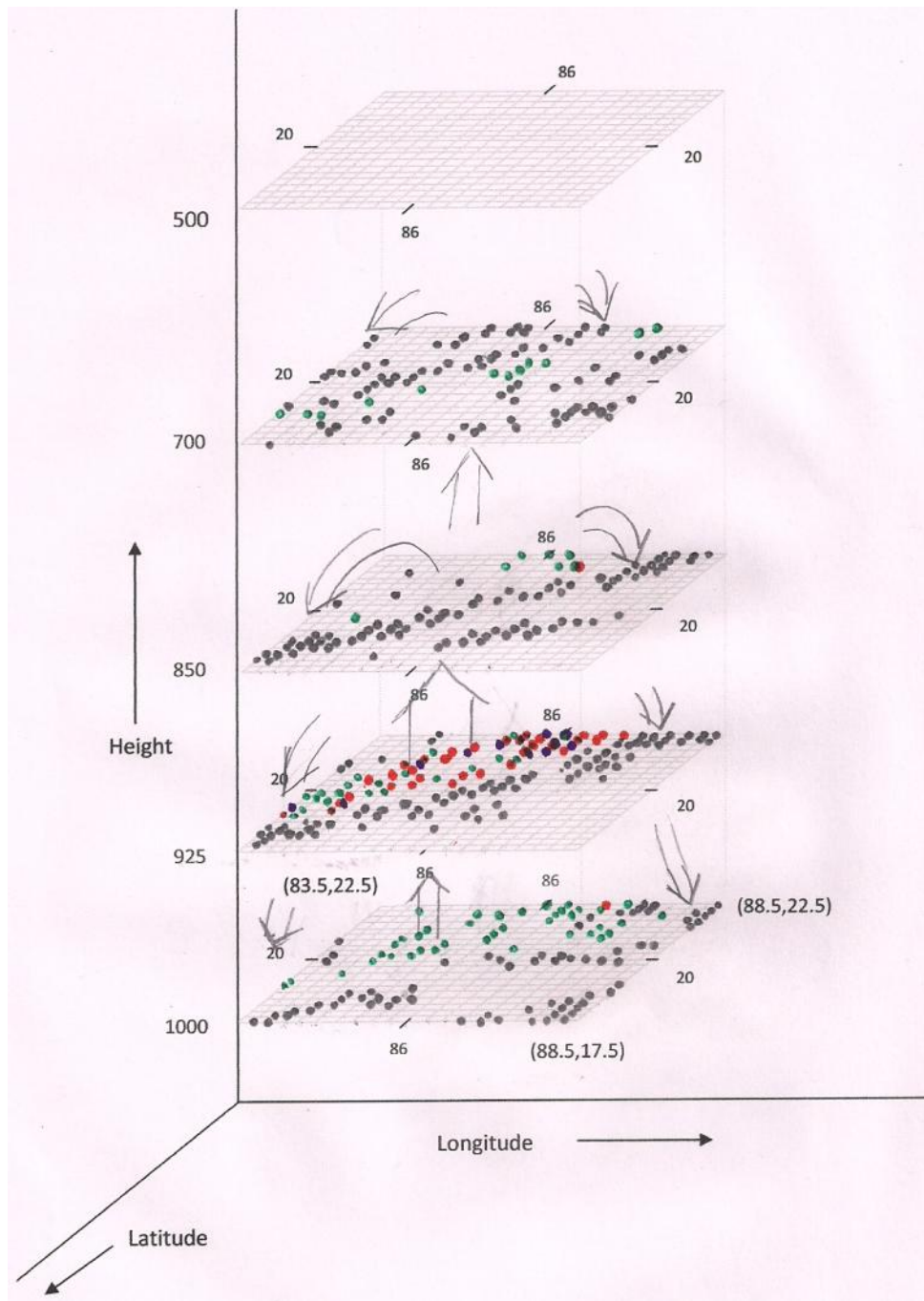


Figure 4.30 Forecast of convergence valid for 1200GMT 31 March 09

(Location of tornado: 86°E, 20°N)

- (green : -12×10^{-5} per second or -14×10^{-5} per second
- red : -16×10^{-5} per second or -18×10^{-5} per second or -20×10^{-5} per second
- purple : -22×10^{-5} per second or -24×10^{-5} per second
- black : 10×10^{-5} per second or 8×10^{-5} per second or 6×10^{-5} per second or 4×10^{-5} per second)

4.7.3 Datasets of WRF model under analysis

With the courtesy of research scientists at NCMRWF, the WRF model configuration at a very fine grid of 9km and run has been done. This model has been initialized with inputs from NCEP analysis. The pre-processing done on these datasets is explained in section 3.2.2.2. The window under analysis has been selected as 82° longitude to 90° longitude and 14° latitude to 26° latitude. This has been done because the location of tornado was Rajakanika block of Kendrapara district of Orissa which is at 86° East longitude and 20° North latitude. There are no missing values in the forecast data.

Now, for data mining, the z-wind component at atmospheric pressure levels of 850 hPa, 750 hPa, 700 hPa and 550 hPa has been selected as the convergence at these levels are good indicative of development of tornado.

The ensembles of vertical wind component forecast for 78hr forecast made based on 0000GMT 28 March 09, 54hr forecast made based on 0000GMT 29 March 09 weather variables, 30hr forecast made based on 0000GMT 30 March 09 and 6hr forecast based on 0000GMT 31 March 09, valid for 0600GMT 31 March 2009 is created for 4 different atmospheric pressure levels. Similarly, the ensembles of vertical wind component forecast for 84hr forecast made based on 0000GMT 28 March 09, 60hr forecast made based on 0000GMT 29 March 09 weather variables, 36hr forecast made based on 0000GMT 30 March 09 and 12hr forecast based on 0000GMT 31 March 09, valid for 1200GMT 31 March 09 is created for 4 different atmospheric pressure levels.

4.7.4 Data mining of WRF output field

The two ensembles of vertical wind motion values made for forecasts valid for 0600GMT 31 March 09 and 1200GMT 31 March 09 as mentioned above have been analyzed (Pabreja, 2010e) using clustering technique of data mining. For each ensemble, the k-means method of clustering is used to generate clusters corresponding to the positive and negative values to vertical wind at every atmospheric pressure level under consideration. We have selected z-wind component as the attribute on the basis of which the clustering should be done. The number of clusters is kept as two. The groups that are identified are exclusive so that an instance belongs to only one group. There are other research works (Singh, Ganju and Singh, 2005; Stojanova, Panov, and Koblar, 2006) that are based on clustering technique of data mining for prediction of weather events.

4.7.5 Visualization of clusters of z-wind component

The 3D view of clusters of ensemble of vertical wind component for forecast valid for 0600GMT 31 March 09 has been shown in Figure 4.31.

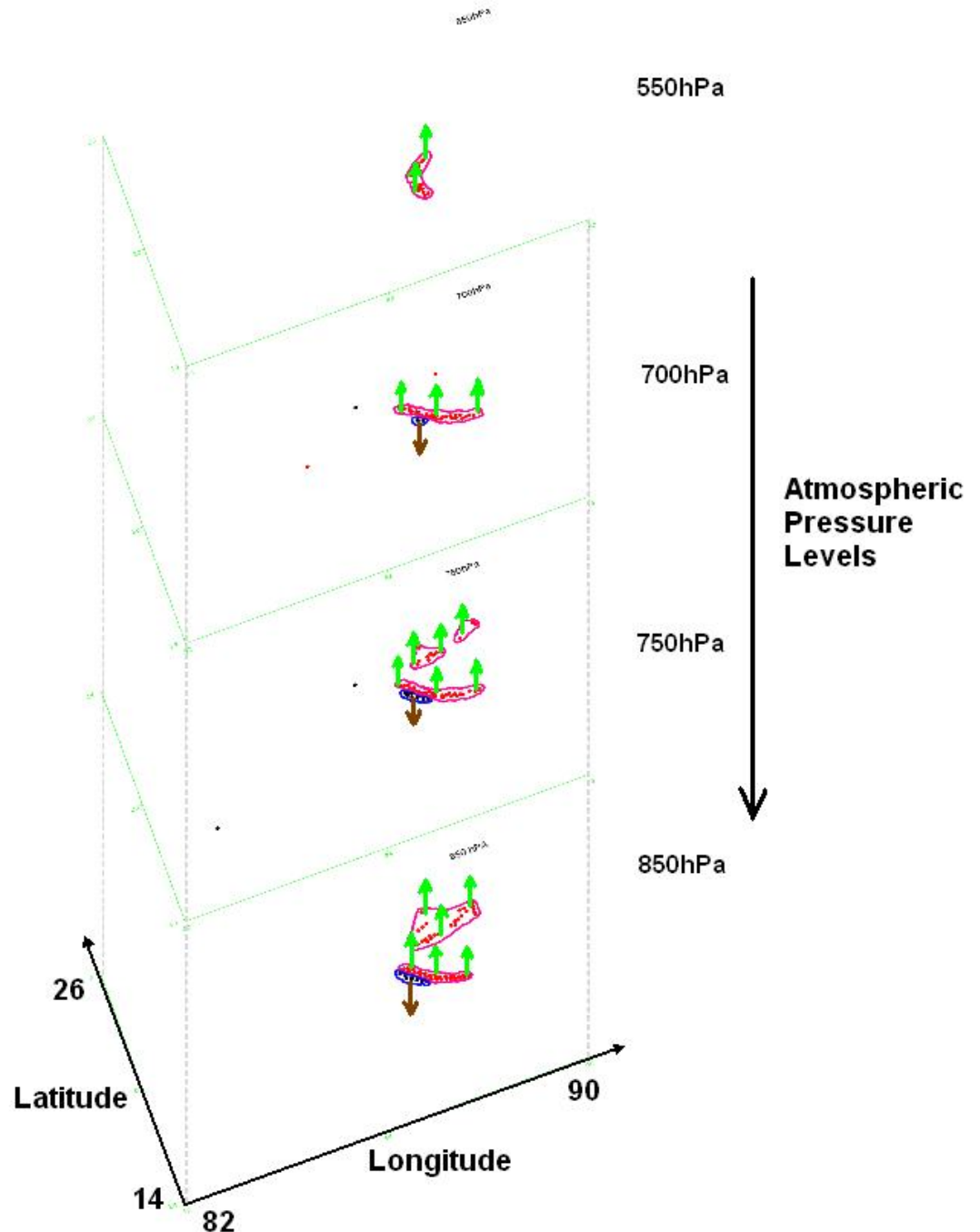


Figure 4.31 3-dimensional visualization of forecast of z-wind (arrows represent vertical motion field), valid for 0600GMT 31 March 09 (Location of tornado: 86°E, 20°N).

Within the clusters -

Red points : less than -1m/s

Blue points : more than 1m/s

The 3D view of clusters of ensemble of vertical wind component for forecast valid for 1200GMT 31 March 09 has been shown in Figure 4.32.

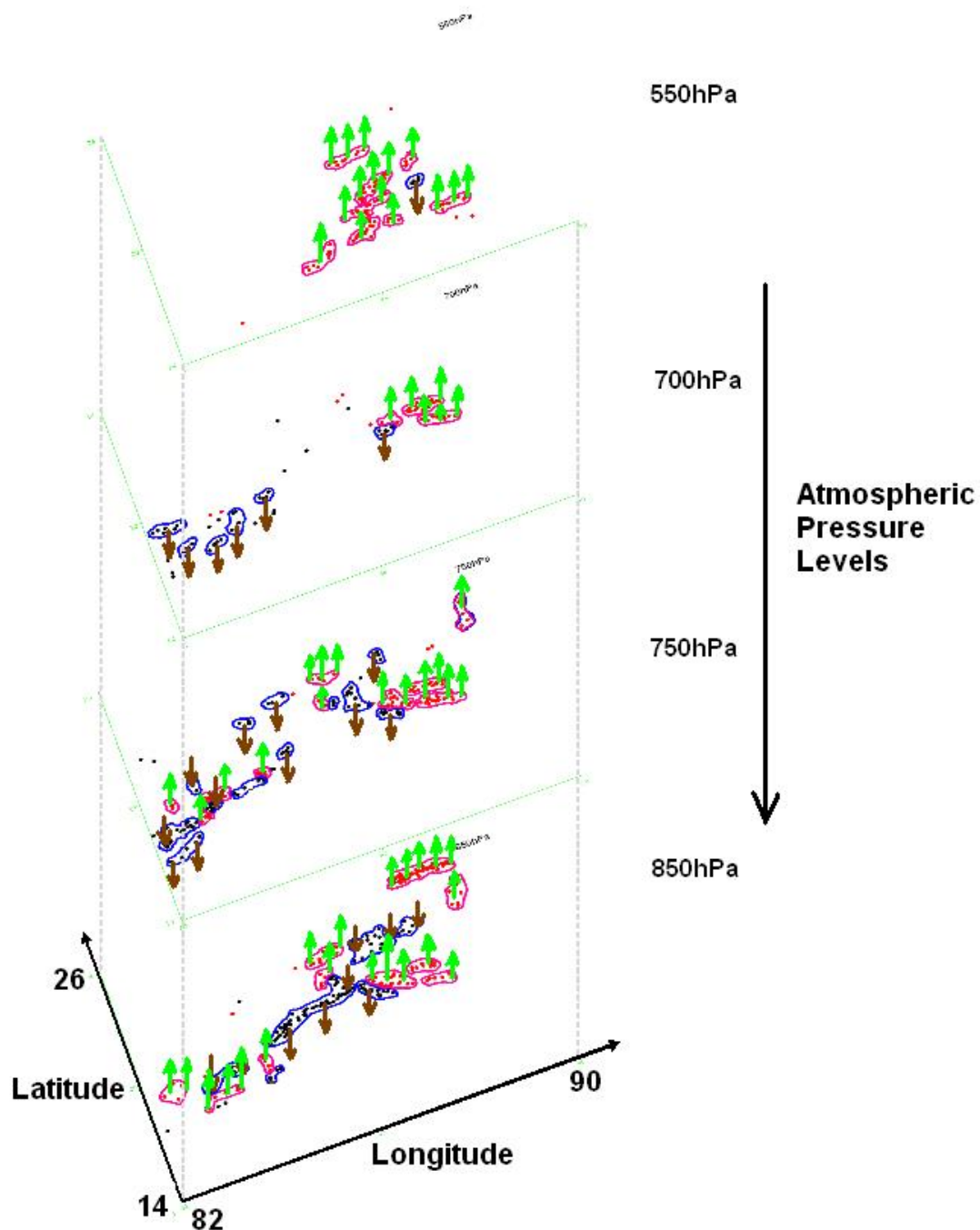


Figure 4.32 3-dimensional visualization of forecast of z-wind (arrows represent vertical motion field), valid for 1200GMT 31 March 09 (Location of tornado: 86°E, 20° N)

Within the clusters -

Red points : less than -1m/s

Blue points : more than 1m/s

4.7.6 Interpretation of clusters of z-wind component

The forecast of 0600GMT 31 March 09 indicates quite weak vertical motion field whereas the forecast of 1200GMT 31 March 09, which is more near the time of occurrence of tornado (at 1640 hrs IST the tornado hit the ground) shows presence of strong vertical motion field based on all forecasts made i.e. on 0000GMT 28 March 09, 0000GMT 29 March 09, 0000GMT 30 March 09 and 0000GMT 31 March 09. Thus this pattern of convergence is an indicative of early signal of tornado formation

4.8 Data mining for Interpretation of sub-grid scale weather system –“Cloudburst” using NWP (ECMWF) outputs

Cloudburst is a very frequent weather phenomenon that takes place in hilly as well as coastal regions of India. Few of such cases have been identified and the ECMWF model outputs for 84 hour forecast, 60 hour forecast, 36 hour forecast, 12 hour forecast at various surface levels, valid for the following dates of cloudburst in India have been collected from IMD.

1. 29 July 2009, Dhaka, Bangladesh
2. 8 August 2009, Pittorgarh distt. Of Uttarakhand
3. 18 July, 2009, Chamoli distt. Of Uttarakhand
4. 7 August 2009, Shimla, Himachal Pradesh

4.8.1 Pre-processing of data – same steps as in section 3.2.2.1

4.8.2 Technique applied

We have tried to pick up zones of updraft of air mass that is an early signal of formation of cloudburst, represented by convergence in the pre-processed dataset. Hence k-means clustering technique has been applied for feature viz. convergence. This field has positive and negative values depicting downdraft and updraft of air mass. So, the number of clusters has been taken as equal to two. The groups that are identified are exclusive so that an instance belongs to only one group. The ensemble of eight forecasts, four before time of cloudburst and four after the time of cloudburst, at each of the selected atmospheric pressure levels has been mined. The four case studies are explained in the following sections.

4.8.3 Cloudburst case under consideration- Dhaka

Date : 29 July, 2009 (between 1:00am and 7:00am)

Location : Dhaka, Bangladesh (23.5°N and 90.25°E)

Area under consideration: 2.5° X 2.5° window surrounding the location of cloudburst
i.e.21.0°N, 87.75°E to 26.0°N, 92.75°E

Forecast used for creating two ensembles:

- Forecast made on 0000GMT 25 July 09, 0000GMT 26 July 09, 0000GMT 27 July 09 and 0000GMT 28 July 09 valid for 1800GMT 28 July 09.
- Forecast made on 0000GMT 25 July 09, 0000GMT 26 July 09, 0000GMT 27 July 09 and 0000GMT 28 July 09 valid for 0000GMT 29 July 09.

Data mining:

After pre-processing as mentioned above, the clusters using k-means clustering technique have been generated. Corresponding to ensemble created for the forecast made on 0000GMT 25 July 09, 0000GMT 26 July 09, 0000GMT 27 July 09 and 0000GMT 28 July 09 valid for 1800GMT 28 July 09, two clusters of convergence and divergence have been generated by the tool. These clusters have been generated for the different atmospheric pressure levels viz. 300hPa, 400hPa, 500hPa, 700hPa, 850hPa, 925hPa, and 1000hPa. The 3-dimensional visualization of the convergence and divergence at various atmospheric pressure levels for forecast on 1800GMT 28July 09 is also being plotted, in Figure 4.33.

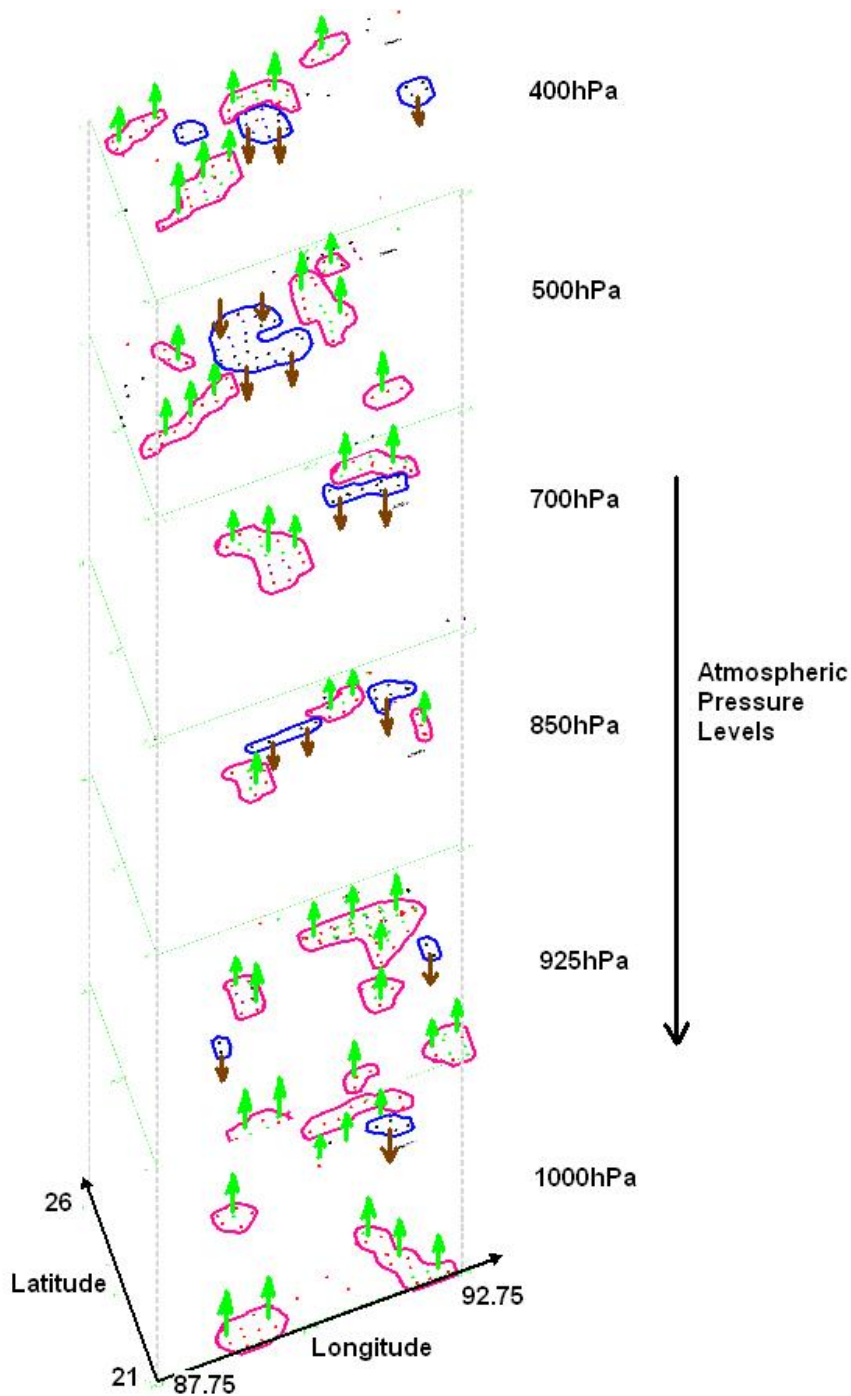


Figure 4.33 3-dimensional visualization of forecast of convergence (arrows represent vertical motion field), valid for 1800GMT 28July 09 (Location of cloudburst: 23.5°N, 90.25°E - Dhaka)

Within the clusters -

- Red points : -8×10^{-5} per sec. or -10×10^{-5} per sec. or -12×10^{-5} per sec.
- Green points : -14×10^{-5} per sec. or -16×10^{-5} per sec. or -18×10^{-5} per sec.
- Purple points : -20×10^{-5} per sec. or -22×10^{-5} per sec. or -24×10^{-5} per sec. or -26×10^{-5} per sec.
- Black points : 8×10^{-5} per sec. to 18×10^{-5} per sec.
- Blue points : 20×10^{-5} per sec. to 40×10^{-5} per sec.

Corresponding to ensemble created for the forecast made on 0000GMT 25 July 09, 0000GMT 26 July 09, 0000GMT 27 July 09 and 0000GMT 28 July 09, valid for 0000GMT 29 July 09, two clusters of convergence and divergence have been generated by the tool. These clusters have been generated for the different atmospheric pressure levels viz. 300hPa, 400hPa, 500hPa, 700hPa, 850hPa, 925hPa, and 1000hPa. The 3-dimensional visualization of the convergence

and divergence at various atmospheric pressure levels for forecast valid for 0000GMT 29July 09 is also being plotted, in Figure 4.34.

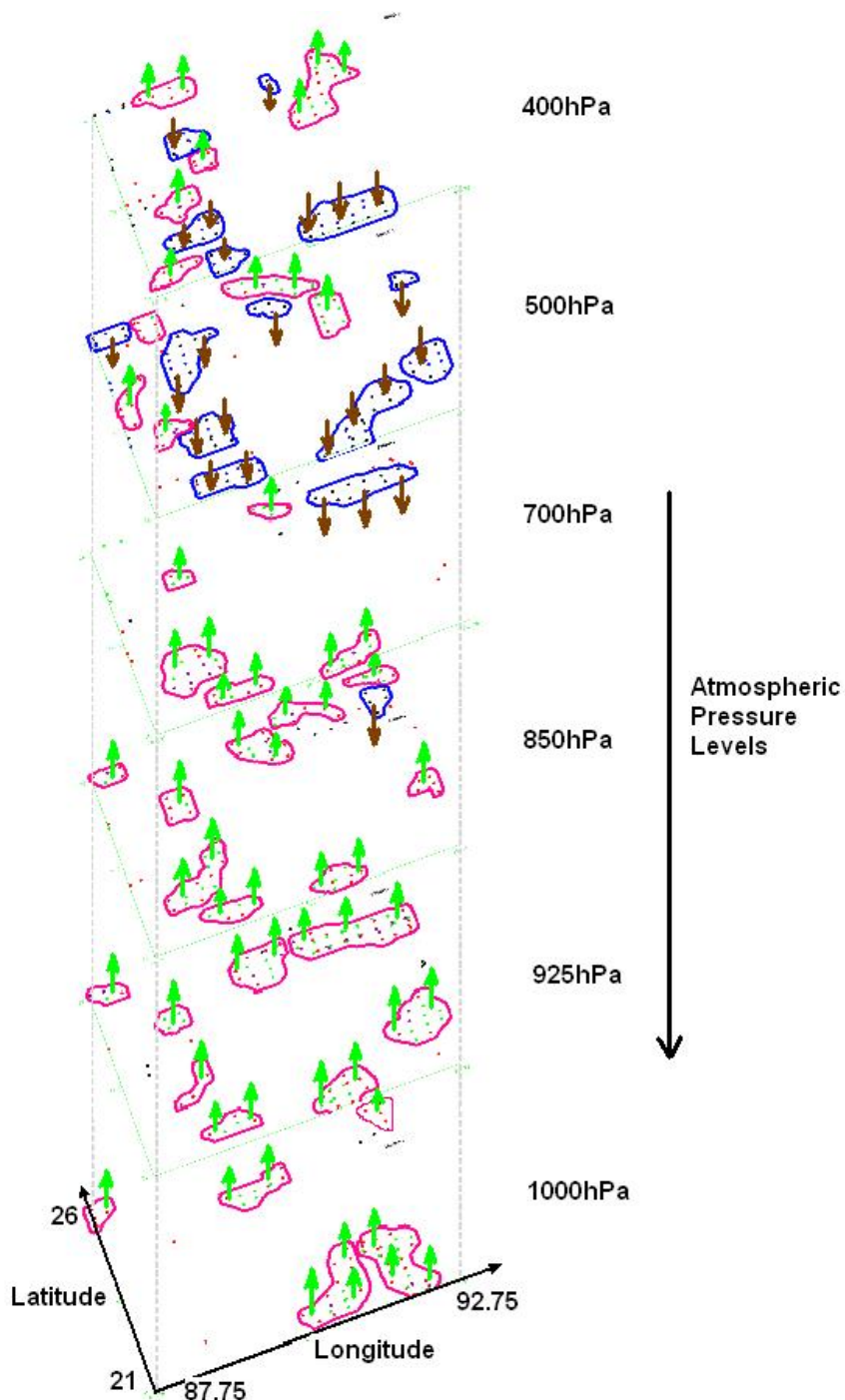


Figure 4.34 3-dimensional visualization of forecast of convergence (arrows represent vertical motion field), valid for 0000GMT 29July 09 (Location of cloudburst: 23.5°N, 90.25°E - Dhaka)

Within the clusters -

- Red points : -8×10^{-5} per sec. or -10×10^{-5} per sec. or -12×10^{-5} per sec.
- Green points : -14×10^{-5} per sec. or -16×10^{-5} per sec. or -18×10^{-5} per sec.
- Purple points : -20×10^{-5} per sec. or -22×10^{-5} per sec. or -24×10^{-5} per sec. or -26×10^{-5} per sec.
- Black points : 8×10^{-5} per sec. to 18×10^{-5} per sec.
- Blue points : 20×10^{-5} per sec. to 40×10^{-5} per sec.

4.8.4 Cloudburst case under consideration- Pittorgarh district

Date : 8 August, 2009 (2:45pm)

Location : Pittorgarh district, Uttarakhand (30°24'N, 80°19'E)

Area under consideration : 2.5° X 2.5° window surrounding the location of cloudburst
i.e. 27.5°N, 77.5°E to 32.5°N, 82.5°E

Forecast used for creating two ensembles:

- Forecast made on 0000GMT 5 Aug 09, 0000GMT 6 Aug 09, 0000GMT 7Aug 09 and 0000GMT 8 Aug 09 valid for 0600GMT 8 Aug 09.
- Forecast made on 0000GMT 5 Aug 09, 0000GMT 6 Aug 09, 0000GMT 7Aug 09 and 0000GMT 8 Aug 09 valid for 1200GMT 8 Aug 09.

Data mining:

After pre-processing as mentioned above, the clusters using k-means clustering technique have been generated. Corresponding to ensemble created for the forecast made on 0000GMT 5 Aug 09, 0000GMT 6 Aug 09, 0000GMT 7Aug 09 and 0000GMT 8 Aug 09 valid for 0600GMT 8 Aug 09, two clusters of convergence and divergence have been generated by the tool. These clusters have been generated for the different atmospheric pressure levels viz. 400hPa, 500hPa, 700hPa, 850hPa, 925hPa, and 1000hPa. The 3-dimensional visualization of the convergence and divergence at various atmospheric pressure levels for forecast valid for 0600GMT 8Aug 09 is being plotted, in Figure 4.35.

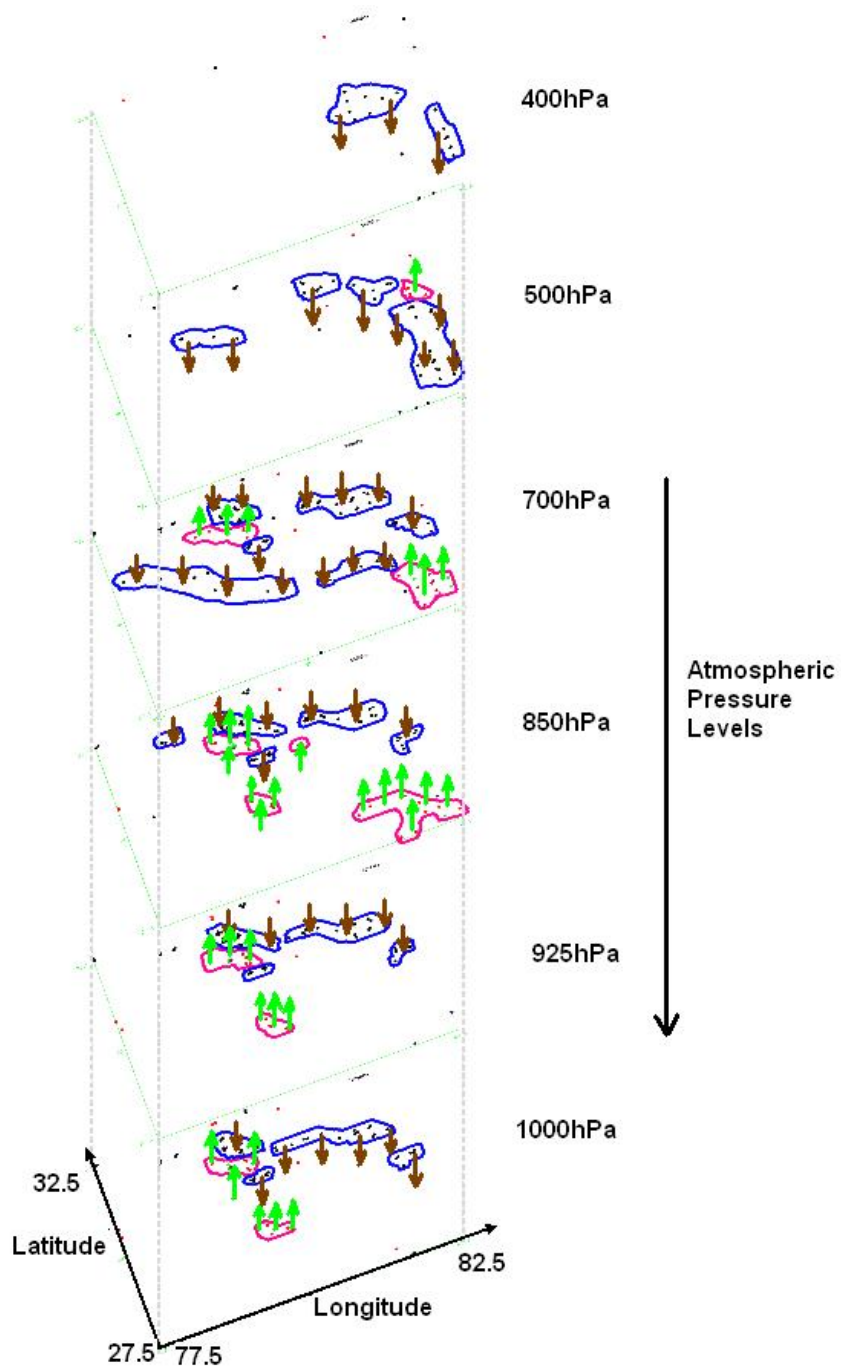


Figure 4.35 3-dimensional visualization of forecast of convergence (arrows represent vertical motion field), valid for 0600GMT 8 Aug 09 (Location of cloudburst: 30°24'N, 80°19'E - Pittorgarh)

Within the clusters -

- Red points : -8×10^{-5} per sec. or -10×10^{-5} per sec. or -12×10^{-5} per sec.
- Green points : -14×10^{-5} per sec. or -16×10^{-5} per sec. or -18×10^{-5} per sec.
- Purple points : -20×10^{-5} per sec. or -22×10^{-5} per sec. or -24×10^{-5} per sec. or -26×10^{-5} per sec.
- Black points : 8×10^{-5} per sec. to 18×10^{-5} per sec.
- Blue points : 20×10^{-5} per sec. to 40×10^{-5} per sec.

Corresponding to ensemble created for the forecast made on 0000GMT 5 Aug 09, 0000GMT 6 Aug 09, 0000GMT 7 Aug 09 and 0000GMT 8 Aug 09 valid for 1200GMT 8 Aug 09, two clusters of convergence and divergence have been generated by the tool. These clusters have been generated for the different atmospheric pressure levels viz. 400hPa, 500hPa, 700hPa,

850hPa, 925hPa, and 1000hPa. The 3-dimensional visualization of the convergence and divergence at various atmospheric pressure levels for forecast valid for 1200GMT 8 Aug 09 is also being plotted, in Figure 4.36.

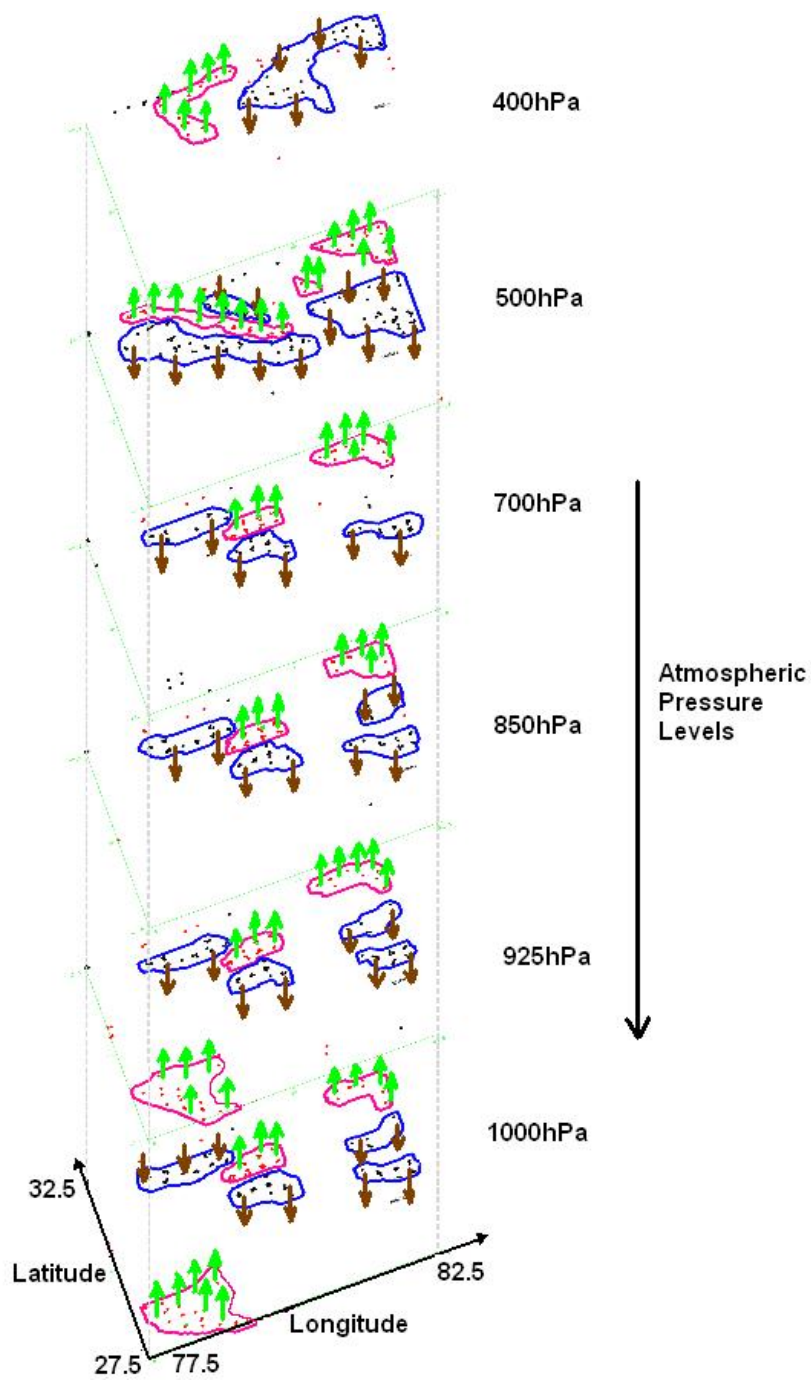


Figure 4.36 3-dimensional visualization of forecast of convergence (arrows represent vertical motion field), valid for 1200GMT 8 Aug 09 (Location of cloudburst: $30^{\circ}24'N$, $80^{\circ}19'E$ - Pittorgarh)

Within the clusters -

- Red points : -8×10^{-5} per sec. or -10×10^{-5} per sec. or -12×10^{-5} per sec.
- Green points : -14×10^{-5} per sec. or -16×10^{-5} per sec. or -18×10^{-5} per sec.
- Purple points : -20×10^{-5} per sec. or -22×10^{-5} per sec. or -24×10^{-5} per sec. or -26×10^{-5} per sec.
- Black points : 8×10^{-5} per sec. to 18×10^{-5} per sec.
- Blue points : 20×10^{-5} per sec. to 40×10^{-5} per sec.

4.8.5 Cloudburst case under consideration - Chamoli district

Date : 18 July, 2009 (night)

Location : Chamoli district, Uttarakhand (30.25°N and 79.25°E)

Area under consideration : 2.5° X 2.5° window surrounding the location of cloudburst
i.e. 27.75°N ,76.75°E to 32.75°N, 81.75°E

Forecast used for creating two ensembles:

- Forecast made on 0000GMT 15 July 09, 0000GMT 16 July 09, 0000GMT 17 July 09 and 0000GMT 18 July 09 valid for 1200GMT 18July 09.
- Forecast made on 0000GMT 15 July 09, 0000GMT 16 July 09, 0000GMT 17 July 09 and 0000GMT 18 July 09 valid for 1800GMT 18July 09.

Data mining:

After pre-processing as mentioned above, the clusters using k-means clustering technique have been generated. Corresponding to ensemble created for the forecast made on 0000GMT 15 July 09, 0000GMT 16 July 09, 0000GMT 17 July 09 and 0000GMT 18 July 09 valid for 1200GMT 18July 09, two clusters of convergence and divergence have been generated by the tool. These clusters have been generated for the different atmospheric pressure levels viz. 400hPa, 500hPa, 700hPa, 850hPa, 925hPa, and 1000hPa. The 3-dimensional visualization of the convergence and divergence at various atmospheric pressure levels for forecast valid for 1200GMT 18 July 09 is also being plotted, in Figure 4.37.

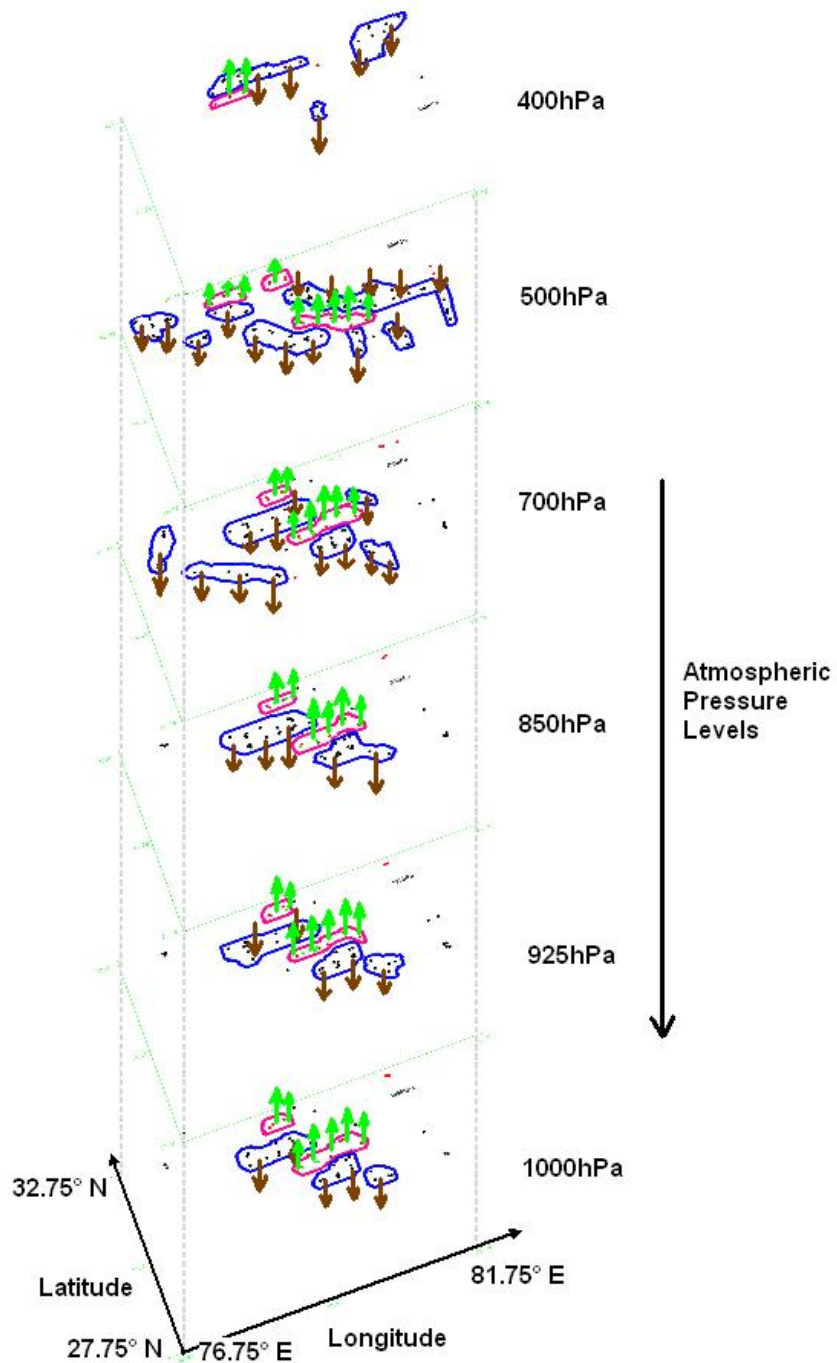


Figure 4.37 3-dimensional visualization of forecast of convergence (arrows represent vertical motion field), valid for 1200GMT 18 July 09 (Location of cloudburst: 30.25°N, 79.25°E - Chamoli)

Within the clusters -

- Red points : -8×10^{-5} per sec. or -10×10^{-5} per sec. or -12×10^{-5} per sec.
- Green points : -14×10^{-5} per sec. or -16×10^{-5} per sec. or -18×10^{-5} per sec.
- Purple points : -20×10^{-5} per sec. or -22×10^{-5} per sec. or -24×10^{-5} per sec. or -26×10^{-5} per sec.
- Black points : 8×10^{-5} per sec. to 18×10^{-5} per sec.
- Blue points : 20×10^{-5} per sec. to 40×10^{-5} per sec.

Corresponding to ensemble created for the forecast made on 0000GMT 15 July 09, 0000GMT 16 July 09, 0000GMT 17 July 09 and 0000GMT 18 July 09 valid for 1800GMT 18 July 09, two clusters of convergence and divergence have been generated by the tool. These clusters have been generated for the different atmospheric pressure levels viz. 400hPa, 500hPa, 700hPa,

850hPa, 925hPa, and 1000hPa. The 3-dimensional visualization of the convergence and divergence at various atmospheric pressure levels for forecast for 1800GMT 18July 09 is also being plotted, in Figure 4.38.

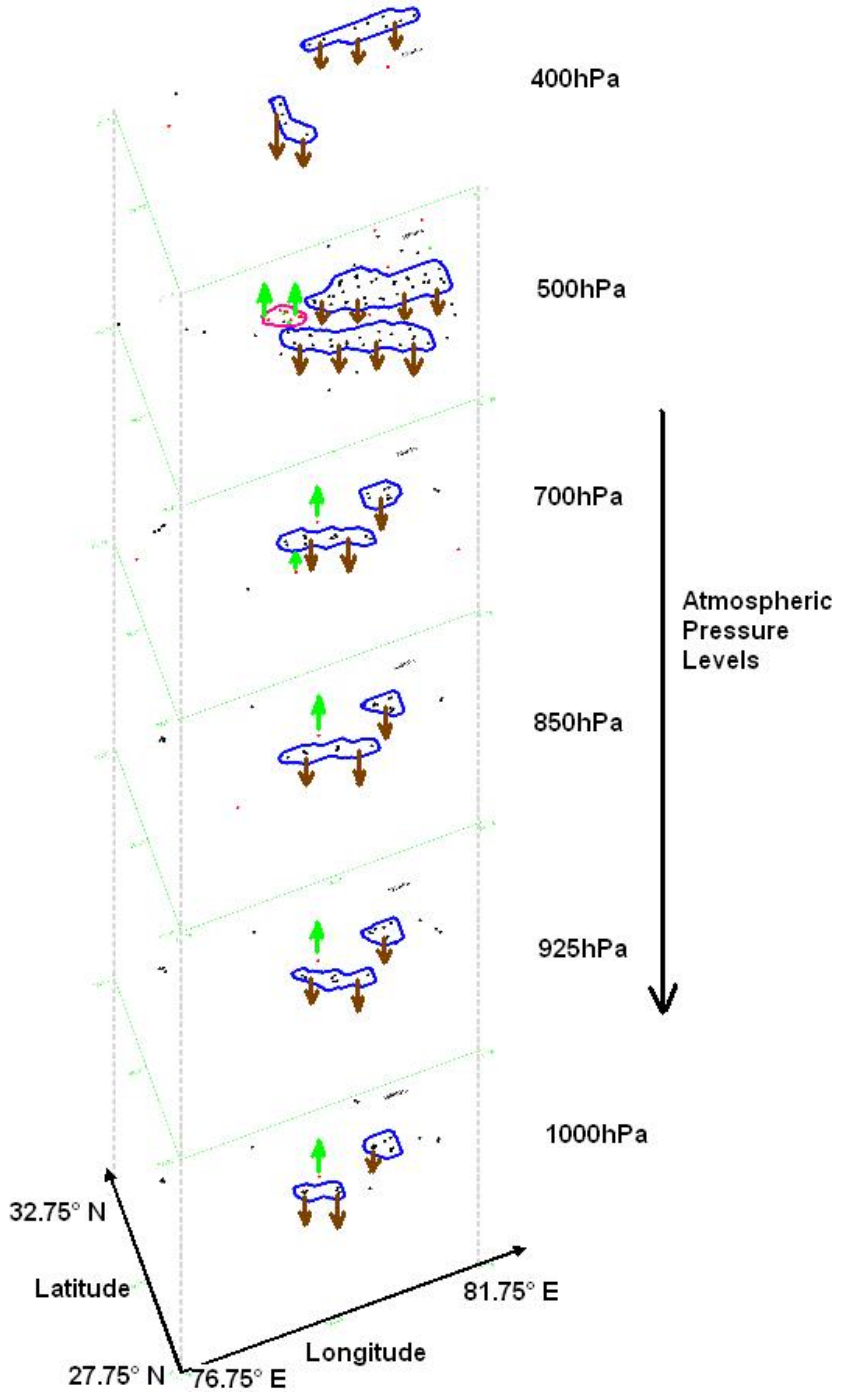


Figure 4.38 3-dimensional visualization of forecast of convergence (arrows represent vertical motion field), valid for 1800GMT 18July 09 (Location of cloudburst: 30.25°N, 79.25°E - Chamoli)

Within the clusters -

- Red points : -8×10^{-5} per sec. or -10×10^{-5} per sec. or -12×10^{-5} per sec.
- Green points : -14×10^{-5} per sec. or -16×10^{-5} per sec. or -18×10^{-5} per sec.
- Purple points : -20×10^{-5} per sec. or -22×10^{-5} per sec. or -24×10^{-5} per sec. or -26×10^{-5} per sec.
- Black points : 8×10^{-5} per sec. to 18×10^{-5} per sec.
- Blue points : 20×10^{-5} per sec. to 40×10^{-5} per sec.

4.8.6 Cloudburst case under consideration- Shimla

Date : 7 August 2009, (7:10pm)

Location : Shimla, Himachal Pradesh (31.0°N and 77.0°E)

Area under consideration : 2.5° X 2.5° window surrounding the location of cloudburst
i.e. 28.5°N ,74.5°E to 33.5°N, 79.5°E

Forecast used for creating two ensembles:

- Forecast made on 0000GMT 4 Aug 09, 0000GMT 5 Aug 09, 0000GMT 6 Aug 09 and 0000GMT 7 Aug 09 valid for 1200GMT 7 Aug 09.
- Forecast made on 0000GMT 4 Aug 09, 0000GMT 5 Aug 09, 0000GMT 6 Aug 09 and 0000GMT 7 Aug 09 valid for 1800GMT 7 Aug 09.

Data mining:

After pre-processing as mentioned above, the clusters using k-means clustering technique have been generated. Corresponding to ensemble created for the forecast made on 0000GMT 4 Aug 09, 0000GMT 5 Aug 09, 0000GMT 6 Aug 09 and 0000GMT 7 Aug 09 valid for 1200GMT 7 Aug 09, two clusters of convergence and divergence have been generated by the tool. These clusters have been generated for the different atmospheric pressure levels viz. 400hPa, 500hPa, 700hPa, 850hPa, 925hPa, and 1000hPa. The 3-dimensional visualization of the convergence and divergence at various atmospheric pressure levels for forecast on 1200GMT 7Aug 09 is also being plotted, in Figure 4.39.

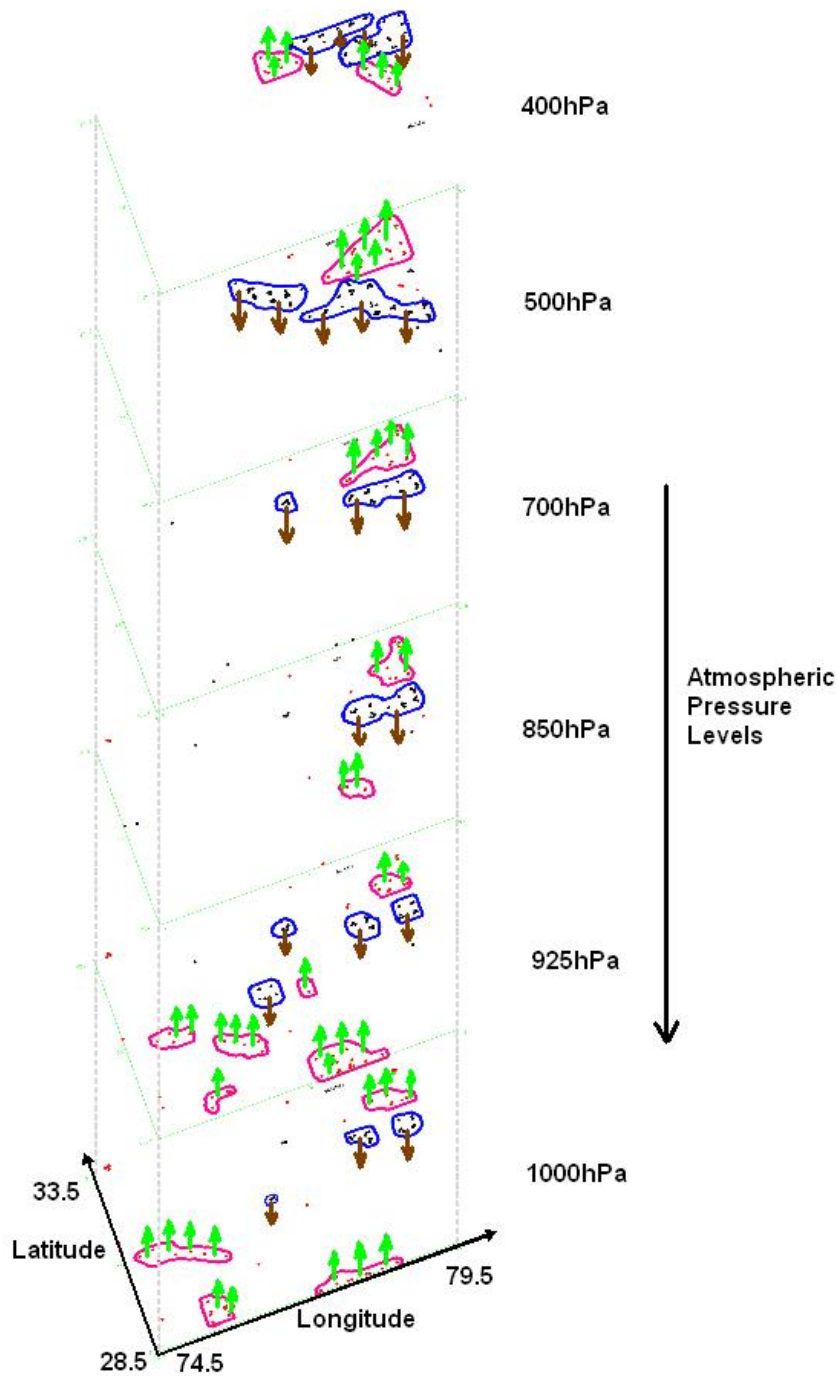


Figure 4.39. 3-dimensional visualization of forecast of convergence (arrows represent vertical motion field), valid for 1200GMT 7 Aug 09 (Location of cloudburst: 31°N, 77°E - Shimla)

Within the clusters -

- Red points : -8×10^{-5} per sec. or -10×10^{-5} per sec. or -12×10^{-5} per sec.
- Green points : -14×10^{-5} per sec. or -16×10^{-5} per sec. or -18×10^{-5} per sec.
- Purple points : -20×10^{-5} per sec. or -22×10^{-5} per sec. or -24×10^{-5} per sec. or -26×10^{-5} per sec.
- Black points : 8×10^{-5} per sec. to 18×10^{-5} per sec.
- Blue points : 20×10^{-5} per sec. to 40×10^{-5} per sec.

Corresponding to ensemble created for the forecast made on 0000GMT 4 Aug 09, 0000GMT 5 Aug 09, 0000GMT 6 Aug 09 and 0000GMT 7 Aug 09 valid for 1800GMT 7 Aug 09, two clusters of convergence and divergence have been generated by the tool. These clusters have been generated for the different atmospheric pressure levels viz. 400hPa, 500hPa, 700hPa,

850hPa, 925hPa, and 1000hPa. The 3-dimensional visualization of the convergence and divergence at various atmospheric pressure levels for forecast valid for 1800GMT 7Aug 09 is also being plotted, in Figure 4.40.

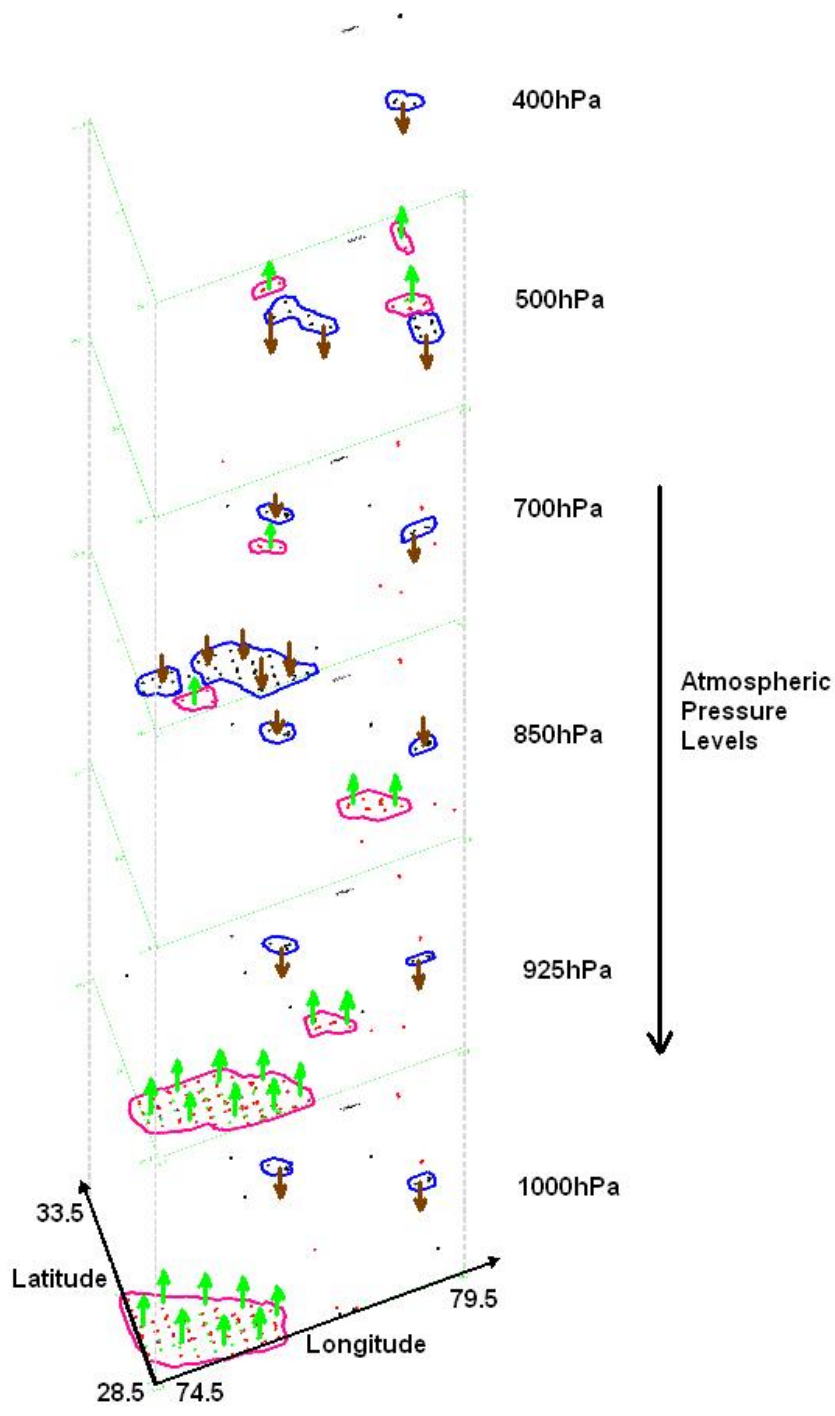


Figure 4.40 3-dimensional visualization of forecast of convergence (arrows represent vertical motion field), valid for 1800GMT 7Aug 09 (Location of cloudburst: 31°N, 77°E -Shimla)

Within the clusters -

- Red points : -8×10^{-5} per sec. or -10×10^{-5} per sec. or -12×10^{-5} per sec.
- Green points : -14×10^{-5} per sec. or -16×10^{-5} per sec. or -18×10^{-5} per sec.
- Purple points : -20×10^{-5} per sec. or -22×10^{-5} per sec. or -24×10^{-5} per sec. or -26×10^{-5} per sec.
- Black points : 8×10^{-5} per sec. to 18×10^{-5} per sec.
- Blue points : 20×10^{-5} per sec. to 40×10^{-5} per sec.

4.8.7 Interpretation of visualization of clusters of convergence

There is a very strong vertical motion field up to atmospheric pressure level 700hPa in the clusters of ensemble of forecast, shown in Figure 4.33 and Figure 4.34 which is for cloudburst on 29 July, 2009 in Dhaka. It is observed that there is an active region of convergence which is an early signal of formation of cloudburst.

For the cloudbursts cases of hilly regions *viz.* Pittorgarh on 8th Aug'09 (Figure 4.35 and Figure 4.36), Chamoli on 18July'09 (Figure 4.37 and Figure 4.38) and Shimla on 7th Aug'09 (Figure 4.39 and Figure 4.40), it is important to focus on the upper atmosphere levels only as the levels at 1000hPa and 925hPa are virtual areas because of presence of the hills. So, we have to observe the levels 850hPa and above. In these three cases, because of the orography of the area which NWP model cannot integrate well while forecasting, the vertical wind motion field is not so well formed. Although there is indication of presence of vertical motion field, but it is not as clear as in case of coastal region's cloudburst.

Chapter 5

Results and Discussion

Last 4 years, the author has to first learn a new societal significant discipline of meteorology and its various temporal/ spatial scales. It further required understanding of vast and complex datasets of Numerical Weather Prediction (NWP) output products. Serious limitations of using normal 2-D relational database management systems for retrieval and interpretation of data available in various time and space scale led the author to develop an efficient Multidimensional data model for use.

Next step was to select alternative for unstable MOS for interpretation of NWP output products in terms of weather elements. Literature survey and discussions with experts in meteorology in IMD and NCMRWF led us to the approval of DM clustering tools.

As the main aim of the research was to interpret sub-grid scale weather systems viz. tornado, cloudbursts and the associated rainfall using DM tools, in order to assess the efficacy of data mining algorithms for interpretation of weather systems at sub-grid scale, it was worthwhile to test these first on well defined synoptic scale weather systems. Fortunately for that purpose we got data for the years 1984-2003 for Low Pressure System (LPS) movement and rainfall, during monsoon period, over Indian Region. The results which gave confidence to apply these tools for sub-grid scale are given below. The variety and complexity of meteorological parameters required use of multidimensional database systems to facilitate the handling of huge datasets of NWP model outputs. The results of the experience gained in developing and using MDDM for the study of LPS and sub-grid scale weather systems viz. cloudbursts and tornado using DM tools are given below:-

5.1 Synoptic scale weather system “Low Pressure System” Movement over Indian Region

The k-means clustering technique has been used to generate the clusters of formation and dissipation of LPS. This has resulted in locating a cluster of formation of Low Pressure System on 1st day and a cluster of disappearance on 4th or 5th day of formation, on a spatio-temporal scale for the months of June-July and Aug-Sept separately, for two sets of 9years each as shown

in Figure 4.2 and Figure 4.4 (year 1984, 85, 88, 89, 90, 91, 92, 93, 94) and Figure 4.3 and Figure 4.5 (year 1995 to 2003).

There is a strong resemblance between the location of clusters of formation and disappearance in Figure. 4.2 and Figure 4.3 which is for Jun-July months for the two sets of years. There is also a strong resemblance between the location of clusters of formation and disappearance in Figure 4.4 and 4.5 which is for Aug-Sept months, for the two sets of years. Hence the favored zones of formation and disappearance of LPS on a spatio-temporal scale, over Indian region during June to September could be identified. There is also an important observation that clusters of LPS move predominantly north- west ward from the location of formation during June and July, as shown in Figure 4.6. It has also been observed that when the movement of LPS is towards North-West direction, it causes rainfall in the South-West zone as shown in Figure 4.7 to Figure 4.9. These results were discussed with meteorological community during Intromet 2009 (Pabreja, 2009) and were accepted. These findings have been explained in section 4.1.3. Mainly two noteworthy results emerge.

- (i) Mostly the Low Pressure systems form in favored regions and move north-west wards causing heavy rain in south-west sector.
- (ii) Data mining clustering tools are found useful to describe these.

5.2 Multidimensional Data Model for Meteorological Datasets

The Numerical Weather Prediction model output products are multidimensional and very huge in size. There are billions of data values that are required to be analyzed and mined. The selection and processing of the datasets was a tedious job. Hence, it was found that Multidimensional Data Model can be utilized for storage of NWP forecasts of weather variables. It was also observed that the weather variables that are important ingredients to the formation of sub-grid scale phenomenon (tornado and cloudburst) are not predicted directly by the model, so were derived and stored in the data hypercube. These are vorticity and divergence across dimensions time and location (latitude, longitude and atmospheric pressure levels) of forecast. The OLAP model with a server that is based upon a Multidimensional database has been used. The meteorological data relevant to the analysis has been pre-processed and loaded in a multidimensional database which looks like a hypercube, explained in section 4.6. This model has facilitated generation of ensembles of forecast which are input to the DM clustering tool. The efficacy of MDDM simplifies and provides speedy retrieval of the required datasets from the huge forecasts datasets.

In a 3-D cube, the gridded rainfall datasets on a scale of time provided by IMD were added along with the dimensions LPS formation and dissipation over Indian subcontinent that enabled the selection of Rainfall across any of the combination of ranges of the dimensions.

The disk space requirement is half in case of MDDM approach in comparison with relational one. There is a significant improvement in the retrieval time of facts. In the particular case of 5-D data hypercube in section 4.5, the access time of the fact “rainfall” against any combination of dimensions has been reduced to $O(\log_2 2^5)$ i.e. at the most 5 accesses which otherwise would have been 32 if relational database management system was used. The five dimensions that were used are time, gridded location, river catchment area, district and LPS.

If one considers the impact of an event like cloudburst causing a flash flood, MDDM can help in decision making with respect to taluka / district, river catchment area in a particular time. MDDM has also shown its efficiency during application of DM tools for interpretation of NWP model output products.

5.3 Interpretation of ECMWF and WRF forecast using data mining techniques for tornado forecasting

The analysis of patterns conducive to formation of sub-grid scale weather systems – tornado has been done. A real life case of Tornado over Orissa on 31st March, 2009 has been analyzed. The NWP model being used is ECMWF T-799 model which has proven skill sets and grid size of approximately 25km. This model does not forecast vertical wind directly so this has been derived indirectly using u and v wind components to generate convergence. 3-D visualization of clusters using k-means clustering technique has been done to observe the convergence at various atmospheric pressure levels. It has been found that the model can provide indication of formation of strong convergence that is an early signal of tornado as shown in Figure 4.30.

The same Real life case of Tornado has been analyzed using the higher resolution model *viz.* WRF V3.1 that produces forecast at 9km grid size. The k-means clustering technique has been used to generate clusters of vertical wind. The clusters demonstrate strong vertical wind presence at an area surrounding the location of tornado, shown in Figure 4.31 and Figure 4.32. This verifies that the WRF model considered here has a capability of forecasting formation of tornado. These case studies are explained in section 4.7. Notably the results are:-

- (i) There is formation of clusters of intense vertical updraft prior to time of occurrence of tornado in the area surrounding the location of tornado.

- (ii) The study demonstrates that a higher resolution model may be able to provide advance signal conducive to formation of tornado using DM tools.

5.4 Interpretation of ECMWF forecast using Data mining techniques for Cloudburst forecasting

Other sub-grid scale weather event is cloudburst. Four real life cases have been discussed in section 4.8 using DM k-means clustering technique. In these cases, the forecasts produced by ECMWF T-799 model have been analyzed. Three of the cloudbursts cases are from hilly areas namely Chamoli district and Pittorgarh district of Uttarakhand; and Shimla, Himachal Pradesh. One is from coastal region i.e. Dhaka, Bangladesh.

There is a very strong vertical motion field upto atmospheric pressure level 700hPa in the clusters of ensemble of forecast, shown in Figure 4.33 and Figure 4.34 which is for cloudburst on 29 July, 2009 in Dhaka. It is observed that this very large region of convergence is an early signal of formation of cloudburst. From the 3-D visualizations of the hilly regions' cloudbursts cases, Pittorgarh (Figure 4.35 and Figure 4.36), Chamoli (Figure 4.37 and Figure 4.38) and Shimla (Figure 4.39 and Figure 4.40), the observations are as follows. It is important to focus on the upper atmospheric levels only as the levels at 1000hPa and 925hPa are virtual areas because of presence of the hills. So, we have to observe the levels 850hPa and above. In these three cases, because of the orography of the area which NWP model cannot integrate well while forecasting, the vertical motion field is not so well formed so other supporting features are required to be found. Though the presence of the vertical motion field is indicated but it is not as clear as in case of coastal region's cloudburst. Significant results are as follows:-

- (i) There is presence of advance signal of formation of cloudburst that is indicated through ensemble of different temporal forecasts of convergence produced by the model.
- (ii) These patterns are clearer for coastal region in comparison to hilly areas.

5.5 Artificial Neural Network for Rainfall forecasting

Artificial Neural Networks have been used for the purpose of forecasting Rainfall based on previous year's data. Networks were trained with data of year 1989 and tested using rainfall data of the year 1990. The training has been done using three different training functions as

mentioned before: `traincgf`, `trainrp` and `trainscg`. Figure 4.12 to Figure 4.14 demonstrate the result of training with year 1989 dataset and testing with year 1990 datasets. The results are convincing and the network once trained has been tested with year 1990 datasets and the error comes out to be less than 0.005 in 5 epochs for training functions `trainscg` and `traincgf`. With `trainrp` function, it takes 35 iterations to train.

Another rainfall dataset is for the year 1991 and 1992, training with 1991 and testing with 1992. Figure 4.15 to Figure 4.17 demonstrate the result of training with year 1991 dataset and testing with year 1992 datasets. Here again, the results are convincing and the network once trained has been tested with year 1992 datasets and the error comes out to be less than 0.005 in 3 epochs for training functions `trainscg` and `traincgf`. With `trainrp` function, it takes 13 iterations to train. The notable results are as follows:-

- (i) ANN is capable of forecasting rainfall at gridded locations for current year by training the network using previous year's rainfall datasets.
- (ii) ANN has limited use and may not be suitable for interpretation of NWP output products in terms of sub-grid scale phenomenon.

Chapter 6

Conclusions

It is found that the MOS is not a theoretically stable process for interpretation of model output for the prediction of specific weather elements. Thus the MOS technique has no or limited use for the interpretation of sub-grid scale weather phenomenon (cloudburst, tornado and associated rainfall). The literature survey definitely shows that the alternative methods for MOS are necessary (Neiley and Hanson, 2004). From literature survey, it was found that though Data mining and other intelligent techniques have been used in meteorological domain but Data Mining techniques have scarcely been used for interpretation of NWP output products for sub-grid scale weather systems like cloudburst, tornado and associated rainfall.

In view of above, it was decided to explore Data mining techniques first for synoptic scale and then for sub-grid scale with the expectation that with Data Mining approach, the interpretation based predictions at sub-grid scale can be derived from the large scale feature. The author was not conversant with the different formats of the NWP model outputs, so data handling posed a challenge but was solved in a systematic manner.

Hence with the study of LPS movement, pre-processing of data of NWP model product, utilization of Multidimensional Data Model, rainfall prediction using ANN and utilization of Data Mining technique of clustering for improving the interpretation of sub-grid scale system, specifically cloudburst and tornado, the expected results were obtained and following conclusions were drawn:-

6.1 Pre-processing of NWP model output

The output products of NWP model were used to derive the required weather parameters that can be mined so as to generate patterns depicting early indication of sub-grid scale weather events viz. cloudburst and tornado. The meteorological datasets being multidimensional in nature so restructured and stored in Multidimensional Database System. These MDDM facilitated selection of required derived weather variables across a filtered time and space scale as per the time and location of sub-grid scale events.

6.2 Clustering technique for Low Pressure System study

The favored zones of formation and disappearance of LPS on a spatio-temporal scale, over Indian region during June to September for years 1984-2003 could be identified. Also the movement of LPS during June to July month during mentioned years could be located along with the favored zones of heavy rainfall. These convincing results of synoptic scale systems provided an encouragement to further look into sub-grid scale weather events viz. tornado and cloudbursts.

6.3 Patterns depicting Tornado formation

The analysis of patterns conducive to formation of sub-grid scale weather systems viz. tornado has been done. An area of $2.5^{\circ} \times 2.5^{\circ}$ surrounding the location of the event has been analyzed. The clusters (WRF and ECMWF models for tornado) demonstrate strong vertical wind presence at an area surrounding the location of tornado. This verifies that the WRF model considered here has a capability of providing an early signal for formation of tornado. Hence it has been observed that the NWP model with a finer grid is able to provide indication of tornado formation 3-4 days in advance, as the forecast made 4days in advance is providing clear indication of strong convergence.

6.4 Patterns depicting Cloudburst formation

The data mining technique of k-means clustering has also been applied to analyze four real life cases of cloudburst. The forecasts produced by ECMWF model have been analyzed and the required weather variables are derived. Three of these cloudburst cases are from hilly areas namely Chamoli district and Pittorgarh district of Uttarakhand; and Shimla, Himachal Pradesh. One is from coastal region i.e. Dhaka, Bangladesh. In the cloudburst zone, all forecasts converge to give intense vertical motion field. This has been seen in a very comprehensible manner for the cloudburst on the coastal region i.e. Dhaka case as shown in Figure 4.33 and Figure 4.34. But the same could not be observed in the cloudbursts which are of hilly zones as illustrated in the 3-D visualizations of the hilly regions' cloudbursts cases, Pittorgarh (Figure 4.35 and Figure 4.36), Chamoli (Figure 4.37 and Figure 4.38) and Shimla (Figure 4.39 and Figure 4.40).

It is observed that because of the orography of the area, the clusters of ensemble of forecasts of convergence are not strong so other supporting features are required to be found. Also the clusters at atmospheric pressure levels 1000hPa, 925hPa do not contribute because of hilly areas. But, in contrast to this, for the coastal area, the pattern of convergence is able to indicate an early signal of occurrence of cloudburst. Hence, the NWP model is suitable for the coastal areas to pick up early signal of formation of patterns leading to cloudburst. With this

approach, there is a high probability of detection of these severe weather events well in advance. So, the present state of MOS needs to be upgraded with intelligent systems as has been shown in this study.

6.5 Rainfall Forecasting

Artificial Neural Networks have been used for the purpose of forecasting Rainfall based on previous year's data. Networks were trained with data of previous year and tested using rainfall data of the following year. It is concluded that ANN has demonstrated promising results and is very suitable for solving the problem of rainfall forecasting but ANN can not be applied for interpretation of formation of sub-grid scale weather phenomenon.

The study demonstrates interesting, useful and fairly clear signals of formation of sub-grid scale weather phenomenon. The forecast of these systems is extremely important for the society. The weather events like tornadoes and cloudbursts are disastrous and threatening to life. This study has clearly brought out that Data Mining techniques when applied rigorously can help in providing advance information for forecast of sub-grid phenomenon.

Chapter 7

Contributions, Limitations and Future Scope

7.1 Contribution

Based on the study of synoptic and sub-grid scale systems, following contributions are made which are useful for meteorological community.

1. An efficient data model for storage of Multidimensional data of NWP model output products is demonstrated. Hence archive of model outputs for current and past cases of sub-grid scale weather events can be maintained.
2. Validation of movement of LPS over Indian subcontinent using k-means clustering technique which is accepted by meteorological community.
3. Derivation of sub-grid scale weather systems from NWP model output products is demonstrated. These weather systems are not directly derivable from the NWP model output products because their scale is much smaller than the scale represented by models. In the normal conventional method also these extreme weather events of cloudburst and tornado are predicted on a short time scale through certain significant radar imageries. Even the radar imageries provide only the probability of occurrence, not surety of occurrence and that too a few minutes in advance by observation of hook like image. By creating ensembles of forecasts for the time of occurrence of the extreme weather events, and the cluster generation of these, it was noticed that significant signals become available well in advance to predict the probability of formation of extreme weather events.

Such signals are not possible through normal MOS technique. Despite the number of cases being small, the study has generated sufficient evidence to show that data mining clustering technique can be used effectively for possible prediction of such extreme weather systems through interpretation of NWP output products.

This study has demonstrated that Intelligent systems can be good alternative for unstable MOS. This study is an effort towards providing timely and actionable information of these events using data mining techniques in supplement with NWP models that can be a great benefit to society.

7.2 Limitation

The format of data output produced by NWP models is not in a form that can be handled by any data mining software directly. It involved a lot of steps and efforts for final selection of data for mining. To mention, approximately 60 million data values are to be managed for just one case of tornado / cloudburst.

In order to facilitate storage, retrieval and analysis of such huge meteorological datasets, multidimensional data model has been used and this approach that is very rarely been used as of now in this domain, proved to be very efficient in terms of storage, retrieval, selection and analysis. The entire procedure from data preprocessing to data mining has been set up and can now be applied on other future cases also. Hence, the data processing problem was solved.

In case of the sub-grid scale events viz. tornado and cloudburst, their occurrence is not very frequent over Indian subcontinent and when they occur, it is in remote areas. The number of cases which could be got is only a few. Moreover, the data availability was only from year 2009 onwards as far as the NWP models of India and data availability at IMD and NCMRWF are concerned. Hence only four cases of cloudbursts from ECMWF forecast could be analyzed and one case of tornado with WRF and ECMWF forecasts could be evaluated. So the results are based on these limited cases which have shown a positive impact on improved forecasting.

7.3 Future Scope

It is further proposed to approach Ministry of Earth Sciences for project on integration of data mining techniques with the NWP models. This would provide the detection of sub-grid scale events automatically from the output products of NWP models. Based on observation of patterns conducive to formation of cloudbursts, the forecast of convergence and vorticity can be further applied as input to an ANN that can forecast rainfall. This would help provide better forecast accuracy. Even the NASA Goddard Data Assimilation Office (DAO) have plans to avail itself of the data mining tools that currently exist, with the adaptation to 3-D numerical model forecast output (Atlas, Graves and Emmitt,2004). The authors have mentioned that as the

Numerical models developed and run at the NASA Goddard DAO continue to increase both in spatial and temporal resolution, the analysis of the model output files becomes more challenging. This is where data mining techniques complements the state of the forecast. Every year computer simulations run faster and faster, with a finer grid. It is hoped that some day we will have a real-time computer simulation of a severe weather outbreak that runs along with the outbreak itself with real-world information (Grazulis, 2001). This study can be extended for other regions like United States of America where the frequency of tornado is more.

It is also proposed to apply other intelligent techniques like hybrid harmony search algorithms to search for the optimal association between the convergence/ vorticity and the forecasted rainfall or other weather variables downscaled to sub-grid level. This can testify the forecasting of sub-grid scale system with support of other forecasted weather variables.

References

- (Abraham, Philip and Mahanti, 2004) Abraham A., Philip N.S., Mahanti P.K. Soft Computing Models for Weather Forecasting, *International Journal of Applied Science and Computations*, USA, 2004, Vol.11, No.3, 106-117.
- (Agee and Zurn-Birkhimer, 1998) Agee E., Zurn-Birkhimer S. Variations in USA tornado occurrences during El Niño and La Niña. Preprints, Proceedings of the 19th Conference on Severe Local Storms, American Meteorological Society, 1998, 287-290.
Available from:
<http://ams.confex.com/ams/pdfpapers/115322.pdf>
- (Antonio et al., 2002) Antonio S., Cano R., Sordo C., Jose M., Bayesian Networks for Probabilistic Weather Prediction, Proceedings of the 15th European Conference on Artificial Intelligence, IOS Press, 2002, 695 – 700.
- (Atlas, Graves and Emmitt, 2004) Atlas B., Graves S., Emmitt D. The Data Mining of 3-D Numerical Model Forecast Output, NASA Goddard Space Flight Center, NASA SISIM Intelligent Systems Project, 2004
- (Barnes, 1968) Barnes S.L. On the source of thunderstorm rotation, NSSL Technical Memo ERLTM-NSSL, no. 38, 1968
- (Barnes, 1970) Barnes S.L., Some aspects of a severe right moving thunderstorm deduced from Mesonetwork Rawinsonde Observations, *Journal of Atmospheric Sciences*. 1970, 27: 634-638.
- (Barry and Scott, a technical memorandum) Barry K. C., Scott M. S. National Weather Service Office Melbourne, Florida, A WSR-88D Approach to Waterspout Forecasting, A technical Memorandum
Available from:
http://www.srh.noaa.gov/media/mlb/pdfs/SR_TechMemo156.pdf

- (Bellone, Hughes and Guttorp, 2000) Bellone E., Hughes J.P., Guttorp P. A hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts, *Climate research*, 2000, 15: 1–15.
- (Bell, Anand and Shapcott, 1994) Bell D.A., Anand S.S., Shapcott C.M. Database mining in Spatial databases. Proceedings of International workshop on Spatio-temporal databases, Benicassim, Spain, 1994 (dates are not available)
Available from:
<http://citeseer.ist.psu.edu/cache/papers/cs/2717/http://zSzzSzwww.di.uoa.grzSz~rouvaszSzresearchzSzstdb94.pdf/database-mining-in-spatial.pdf>
- (An Oracle White Paper, 2006) Benefits of a Multi-dimensional Model, An Oracle White Paper, May 2006
- (Berson and Smith, 2005) Berson A., Smith S.J. Data Warehousing, Data Mining and OLAP, Tata McGraw-Hill Publishing Company Limited, New Delhi, 2005
- (Bove, 1998) Bove, M. C. Impacts of ENSO on United States tornadic activity. Preprints, 19th Conference on Severe Local Storms, AMS, 1998. 313-316.
- (Browning and Landry, 1963) Browning K.A., Landry C.R. Airflow within a tornadic storm. Pre-print 10th Weather Radar Conference, *American Meteorological Society* 1963, 116-122.
- (Cavazos, 2000) Cavazos T. Using self-organization maps to investigate extreme climate event, *Journal of Climate*. 2000, 13: 1718-1732.
- (Charles et al., 1993) Charles A., Doswell III, Weiss S.J., Johns R.H. Tornado Forecasting: A Review. In: *The Tornado: Its Structure, Dynamics, Prediction, and Hazards* (C. Church et al., Eds.), Geophysical Monograph 79, Amer. Geophys. Union, 1993, 557-571.
- (Chattopadhyay, 2007)
Chattopadhyay S., *Multilayered feed forward Artificial Neural Network model to predict the average summer-monsoon rainfall in India*, *Journal Acta Geophysica*. 2007, Volume 55, Number 3, 369-382.

(Chattopadhyay and Chattopadhyay, 2007) Chattopadhyay S., Chattopadhyay M., A soft computing technique in rainfall forecasting, Proceedings of the International conference on IT, HIT, March 2007, 523-526

Available from:

<http://arxiv.org/ftp/nlin/papers/0703/0703042.pdf>

(Chow and Cho, 1997) Chow T.W., Cho S.Y. Development of a Recurrent Sigma-Pi Neural Network Rainfall Forecasting System in Hong Kong, *Neural Computation & Applications*, 1997, 5: 66-75.

Available from:

<http://www.springerlink.com/content/q16436171w6lp013/>

(Colliat, 1996) Colliat G. OLAP, Relational, and Multidimensional Database Systems, SIGMOD Record, 1996, Vol. 25, No. 3: 64-69.

Available from:

http://infolab.usc.edu/csci599/Fall2002/paper/i2_p064.pdf

(Collins and Tissot, 2008)

Collins W., Tissot P. Use of an artificial neural network to forecast thunderstorm location, Proceedings of the Fifth Conference on Artificial Intelligence Applications to Environmental Science, San Antonio, TX, 2008 Jan 13-18. Published in Journal of AMS.

Available from:

<http://ams.confex.com/ams/pdfpapers/132577.pdf>

(Cofiño, 2003) Cofiño A.S., Gutiérrez J.M., Jakubiak B., Melonek M. Implementation of data mining techniques for meteorological applications. In: Realizing Teracomputing (W.Zwiefelhofer W., Kreitz N., Eds.), World scientific, 2003, 215-240.

(CyRDAS, 2004) Cyber infrastructure for the atmospheric sciences in the 21st century. Boulder, CO:National Center for Atmospheric Research(NCAR); Ad Hoc Committee for Cyber infrastructure Research, Development and Education in the Atmospheric Sciences (CyRDAS). CyRDAS, 2004

(Das, 2005) Das S., *Mountain weather forecasting using MM5 modelling system*, Current Science, Vol. 88, No. 6, 2005 March

Available from:

<http://www.ias.ac.in/currsci/mar252005/899.pdf>

(Das, Arshit, and Moncrief, 2006) Das S., Arshit R., Moncrief M.W. Simulation of a Himalayan cloudburst event. *Journal of Earth System Science*. 115, No. 3, June 2006, 299-313

(Datta, 1992) Datta R.K. Computer as a Window to Deterministic Weather Forecasting. Proceedings of Eighteenth Session of Indian Science Congress Association (ISCA), Calcutta, India, 1992-93, 5-15

(David-Jones, 1985) David-Jones R.P., Tornado Dynamics. In: Thunderstorms: Morphology and Dynamics, 2nd Edition. (E. Kessler Ed.), Norman, University of Oklahoma Press, 1985, 197-236.

(Droegemeier et al., 2004) Droegemeier K.K, Chandrasekar V., Clark R., Gannon D., Graves S., Joseph E., Ramamurthy M., Wilhelmson R., Brewster K., Domenico B., Leyton T., Morris V., Murray D., Plale B., Ramachandran R., Reed D., Rushing J., Weber D., Wilson A., Xue M., Yalda S. LEAD: Linked Environments For Atmospheric Discovery (Lead): A Cyberinfrastructure For Mesoscale Meteorology Research And Education, a report, 2004

Available from:

<http://lead.ou.edu>.

(Dubes and Jain, 1988) Dubes R.C., Jain A.K. *Algorithms for Clustering Data*. Prentice Hall, 1988.

(Ertöz, Steinbach and Kumar, 2001) Ertöz L., Steinbach M., Kumar V. Finding Topics in Collections of Documents: A Shared Nearest Neighbor Approach. Proceedings of Text Mine'01, First SIAM International Conference on Data Mining, Chicago, IL, USA, 2001.

(Ertöz, Steinbach and Kumar, 2002) Ertöz L., Steinbach M., Kumar V. A New Shared Nearest Neighbor Clustering Algorithm and its Applications. In Workshop on Clustering High Dimensional Data and its Applications, SIAM Data Mining, Arlington, VA, USA, 2002.

- (Ertöz, Steinbach and Kumar, 2003) Ertöz L., Steinbach M., Kumar V. Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data. In Proceedings of the 3rd SIAM International Conference on Data Mining, San Francisco, CA, USA, 2003.
- (Enke and Spekat, 1997) Enke W., Spekat A., Downscaling climate model outputs into local and regional weather elements by classification and regression, *Climate Research*. 1997, 8:195–207.
- (Fayyad, Piatetsky-Shapiro and Smyth, 1996) Fayyad U.M., Piatetsky-Shapiro G., Smyth P. From Data mining to knowledge discovery in databases. In: Advances in Knowledge Discovery and Data Mining (Fayyad U.M., Piatetsky-Shapiro G., Smyth P., Uthurusamy R., Eds.), AAAI / MIT Press, Menlo Park, CA 1996, 37-54
- (Fayyad and Smyth, 1993) Fayyad U.M., Smyth P. Image Database Exploration: Progress and Challenges. Proceedings of Workshop on KDD, Washington, DC, 1993, July 11-12, 14-27.
- (Foster, 1964) Foster D.S. *Relationship among tornadoes, vorticity acceleration and air mass stability*, Monthly Weather Review, American Meteorological Society, 1964.
Available from:
[http://journals.ametsoc.org/doi/abs/10.1175/1520-0493\(1964\)092%3C0339%3ARATVAA%3E2.3.CO%3B2](http://journals.ametsoc.org/doi/abs/10.1175/1520-0493(1964)092%3C0339%3ARATVAA%3E2.3.CO%3B2)
- (Fujita, 1970) Fujita T.T. The Lubbock Tornadoes: A study of suction spots, *Weatherwise*. 1970, 23: 160-173.
- (Fujita, 1981) Fujita T.T., Wakimoto R.M. Five scales of airflow associated with a series of downbursts on 16 July 1980. *Monthly Weather Review* 109:1438-56.
- (Fuzzy logic from Stanford Encyclopedia, 2010) Fuzzy Logic, *Stanford Encyclopedia of Philosophy*. Stanford University, First published Tue Sep 3, 2002; substantive revision Wed Aug 4, 2010
Available from:
<http://plato.stanford.edu/entries/logic-fuzzy/>

- (Gardner and Dorling, 1998) Gardner M.W., Dorling S.R. Artificial Neural Networks (the Multilayer Perceptron) - A Review of Applications in the Atmospheric Sciences, *Journal of Applied Meteorology*. 1998, 39: 147–159.
- (Grazulis, 1993) Grazulis, T.P. Significant Tornadoes 1680-1991, the Tornado Project of Environmental Films. St. Johnsbury, VT. 1993.
- (Grazulis, 2001) Grazulis T.P., The Tornado- Nature's ultimate windstorm, University of Oklahoma Press, Norman, 2001
- (Guo, Dai and Lin, 2004) Guo Z., Dai X., Lin H. Application of association rule in disaster weather forecasting, The international Association of Chinese Professionals in Geographic Information Science, 2004.
- (Hall,1998) Hall T., Precipitation Forecasting Using a Neural Network, *Weather and Forecasting*, Journal of AMS. 1998, 14:338-345.
Available from:
<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.133.9290&rep=rep1&type=pdf>
- (Hall, Brooks and Doswell, 1999) Hall T., Brooks H.E., Doswell C.A. Precipitation Forecasting Using a Neural Network, *Weather and Forecasting*. 1999, 14 : 338-345.
Available from:
<http://journals.ametsoc.org/doi/abs/10.1175/1520-0434%281999%29014%3C0338%3APFUANN%3E2.0.CO%3B2>
- (Han and Kamber, 2006) Han J., Kamber M., Data Mining Concepts and techniques, Morgan Kaufmann Publisher, 2006
- (Han et al. 2001) Han J., Kamber M., Anthony K.H. Tung. Spatial clustering methods in data mining: a survey. In :Geographic Data Mining (Han J., Miller H.J, Eds.), Taylor and Francis, London. 2001, 188-217
- (Han and Kamber, 2001) Han J., Kamber M. Mining Complex Types of Data. In: Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, San Francisco, USA. 2001, 395-450.

(Harinath and Carroll, 2009) Harinath S., Carroll M., Professional Microsoft SQL Server Analysis Services with MDX, Wiley India Pvt. Ltd., 2009.

(Hayati and Mohebi, 2007) Hayati M., Mohebi Z. Temperature forecasting based on neural network approach, *World Applied Sciences journal*. 2007, 2(6) : 613-620.

(Held, 1981) Held G. Comparison of radar observations of a devastating hailstorm and a cloudburst at Jan Smuts Airport. In: Cloud dynamics (Agee E.M., Asal T., Eds.), Proceedings of a Symposium held at the Third General Assembly of IAMAP, Hamburg, West Germany, 17-18 August, 1981, 273-284

(Heidorn, The Weather doctor, online book) Heidorn K.C., The weather doctor, Exploring the Science and Poetry of Our Weather and Atmosphere (online book)
Available from:
<http://www.islandnet.com/~see/weather/doctor.htm>

(Henderson, 2009) Henderson T. *Now we are ready for the next cloudburst; radar can predict flooding danger*, The Journal (Newcastle, England): 2009 July - Free Online Library
Available from:
<http://www.thefreelibrary.com/ /print/PrintArticle.aspx?id=203637660>

(Huanga and Zhaob, 2000) Huanga X., Zhaob F. Relation-based aggregation: finding objects in large spatial datasets, USA Intelligent Data Analysis. 2000, 4: 129-147.
Available from:
<http://portal.acm.org/citation.cfm?id=1294188>

(Hughes, 1976) Hughes H. *Model output statistics forecast guidance*. United States Air Force Environmental Technical Applications Center. pp. 1–16.

(Huiming, 1996) Huiming Y. *Application of meteorological satellites to rainstorm research and forecast in China*, Institute of Astronautics Information of China Aerospace Corporation, 1996
Available from:
<http://www.space.cetin.net.cn/docs/HTM-E/003.HTM>

(Hunt, 1998) Hunt J.C.R. *Lewis fry Richardson And his contributions to Mathematics, meteorology,And models of conflict*, Annu. Rev. Fluid Mech. 1998. 30:xiii–xxxvi
Available from:

[http://www.cpom.org/people/jcrh/AnnRevFluMech\(30\)LFR.pdf](http://www.cpom.org/people/jcrh/AnnRevFluMech(30)LFR.pdf)

(Ivakhnenko, 1971) Ivakhnenko A.G., Polynomial Theory of Complex Systems. IEEE Transactions on Systems, Man, and Cybernetics, Oct 1971,vol. 1, issue 4: 364-378.
Availablefrom: -

<http://ieeexplore.ieee.org/iel5/21/4308307/04308320.pdf?tp=&arnumber=4308320&isnumber=4308307>

(Jayanta, 2004) Jayanta B., Sudarshan A., Trivedi D., Santhanam M.S., *Weather Data Mining using Independent Component Analysis*. Journal of Machine Learning Research, 2004 Dec, 5:239-253

Available from:

<http://portal.acm.org/citation.cfm?id=1005341>

(Jaye, 2006) Jaye S.M., Determining the Likelihood of Severe Weather Based On Model output, 23rd Conference on Severe Local Storms, Nov 5-10, 2006.

Available from:

http://ams.confex.com/ams/23SLS/techprogram/paper_115467.htm

(Jena et al. 2009) Jena R.K., Aqel M.M., Srivastava P., Mahanti P.K. Soft Computing Methodologies in Bioinformatics, European Journal of Scientific Research, 2009, Vol.26 No.2:189-203.

Available from:

<http://www.eurojournals.com/ejsr.htm>

(Johnson et al., 1998) Johnson J. T., Mackeen P. L., Witt A., Mitchell E. D., Stumpf G. J., Eilts M. D., Thomas K. W., 1998: The storm cell identification and tracking algorithm: An enhanced wsr-88d algorithm. Weather and Forecasting (AMS online Journal), June 1998, vol. 13, Issue 2: 263–276

Available from:

<http://ams.allenpress.com/archive/1520-0434/13/2/pdf/i1520-0434-13-2-263.pdf>

(Kjærulff and Madsen, 2005) Kjærulff U.B., Madsen A.L. Probabilistic Networks — An Introduction to Bayesian Networks and Influence Diagrams, Aalborg University, Denmark. 10 May 2005 (online book)

Available from:

<http://www.cs.aau.dk/~uk/papers/pgm-book-I-05.pdf>

(Koperski, Han and Adhikary, 1998) Koperski K., Han J., Adhikary J. *Mining Knowledge in Geographical Data*, Communications of the ACM, 1998 March, Vol.26, No.1., 65-74.

Available from:

<http://citeseer.ist.psu.edu/127714.html>

(Koska, 2005) Koska B., Neural networks and Fuzzy systems – A dynamical systems approach to Machine Intelligence, Prentice hall of India, 2005.

(Kumar et al. 2001) Kumar V., Steinbach M., Tan P.N., Potter C., Klooster S., Torregrosa A. Finding Spatio-temporal Patterns in earth science data. 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, 2001, Aug 26-29, 1-12

Available from:

<http://www.sigkdd.org/kdd2001/Workshops/tsk.pdf>

(Kumar et al. 2004) Kumar V., Steinbach M., Zhang P., Shekhar S., Tan P.N., Potter C., Klooster S., Discovery of changes from the Global Carbon Cycle and climate system using data mining. Fourth annual Earth Science Technology Conference (ESTC), Palo Alto, CA, USA, 2004, Jun 22-24, 1-6

Available from:

<http://esto.nasa.gov/conferences/estc2004/papers/b9p2.pdf>

(Lemke and Müller, 2003) Lemke F., Müller J.A., *Self organizing data mining*, Systems Analysis Modelling Simulation, Feb. 2003 , Vol. 43, No. 2 : 231–240

Available from:

<http://www.knowledgeminer.net/pdf/sodm.pdf>

(Li et al., 2008) Li X., Plale B., Vijayakumar N., Ramachandran R., Graves S., Conover H. Real-time Storm Detection and Weather Forecast Activation through Data Mining and Events Processing, *Earth Science Informatics*, 2008, 49-57.

Available from:

<http://www.cs.indiana.edu/~plale/papers/EarthSciInformatics-preprint.pdf>

(Lin, Lin and Chen, 2008) Lin K., Lin J., Chen B. Study on short-range Precipitation Forecasting Method based on Genetic Algorithm Neural Network, Proceedings of the 7th World congress on Intelligent control and Automation, China, 2008

(Liu and Lee, 1999) Liu J.N.K., Lee R.S.T. Rainfall Forecasting from Multiple Point Sources Using Neural Networks, *IEEE* 1999 , 3: 429-434.

Available from:

<http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=00823243>

(Liu and George, 2006) Liu Z., George R. Mining Weather Data using Fuzzy Cluster Analysis. In: *Fuzzy Modeling with Spatial Information for Geographic Problems*(Petry F.E., Robinson V.B, Cobb M.A., Eds), Springer Berlin Heidelberg, Germany. 2006, 105-119

(Liu and George, 2005) Liu Z., George R. Mining weather data using fuzzy cluster analysis. In: *Fuzzy modeling with spatial information for geographic problems.* (Petry F.E., Vincent B. Robinson, Cobb M.A., Eds.), Springer Berlin Heidelberg, 2005, 105-119.

(Lynch, 2007) Lynch P. *The origins of computer weather prediction and climate modeling*, *Journal of Computational Physics*, ScienceDirect, doi:10.1016/j.jcp.2007.02.034

(Mahabir, Hicks and Robinson, 2003) Mahabir C., Hicks F.E., Robinson F. A. Application of fuzzy logic to forecast seasonal runoff, *Hydrological Processes*, *Wiley InterScience*, 2003, 17:3749–3762.

(Markowski, 2002) Markowski P. M. Surface thermodynamic characteristics of hook echoes and rear-flank downdrafts. Part I: A review, *Monthly Weather Review*, 2002, 130, 852–876.

(Mesinger, 2002) Mesinger F. NCEP Regional Reanalysis. AMS Symposium on Observations, Data Assimilation, and Probabilistic Prediction, 2002, J59-J63.

Available from:

<http://ams.confex.com/ams/pdfpapers/57744.pdf>

(McGovern et al. 2007) McGovern A. , Rosendahl D.H., Kruger A. , Beaton M.G., School of Computer Science, University of Oklahoma, Kelvin K. Droegemeier, School of Meteorology, University of Oklahoma, Brown R.A., NOAA, National Severe Storms Laboratory. Anticipating the formation of tornadoes through data mining. Fifth Conference on Artificial Intelligence Applications to Environmental Sciences (sponsored by AMS), Beacon Street Boston, MA, Jan 14-18, 2007

Available from:

<http://ams.confex.com/ams/pdfpapers/117344.pdf>

(McGregor, Walsh and Katzfey, 1993) McGregor J.L., Walsh K. J., Katzfey J. Climate simulations for Tasmania. Proceedings of the Fourth International Conference on Southern Hemisphere Meteorological and Oceanography, American Meteorological Society, 1993, 514-515.

(Mooley and Shukla, 1987) Mooley, D. A., Shukla J., Characteristics of the westward moving summer monsoon low pressure systems over the Indian region and their relationship with the monsoon rainfall, Department of Meteorology, University of Maryland, College Park, MD, USA, Centre for Ocean-Land-Atmosphere interactions, a report, 1987

(Müller, 2003) Muller J.A. *GMDH-based knowledge extraction from data*. Journal of Control Systems and Computer. 2003, vol. 2. (page numbers not available)

Available from:

<http://www.gmdh.net/articles/usim/Mueller.pdf>

(Murtha, 1995) Murtha J., Applications of fuzzy logic in operational meteorology, scientific services and professional development newsletter, Canadian Forces weather service, 1995, 42-54.

NDFD GRIB2 decoder program of NOAA from Internet

Available from:

www.nws.noaa.gov/mdl/degrib/download.php

(Neiley and Hanson, 2004) Neiley P.P., Hanson K.A. Are model output statistics still needed?
84th AMS Annual Meeting (session 6) , Seattle, Washington, 2004, Jan 11-15, 1-5
Available from:
<http://ams.confex.com/ams/pdfpapers/73333.pdf>

(Ng and Han, 1994) Ng R., Han J. Efficient and Effective clustering method for spatial data mining. Proceedings of International Conference on Very Large Databases, Santiago de Chile, Chile, 1994, September 12-15, 144-155
Available from:
<http://www.cs.sfu.ca/CourseCentral/459/han/papers/ng94.pdf>

(Oprea and Bell, 2009) Oprea I.C., Bell A. Meteorological environment of a tornado outbreak in Southern Romania, *National Hazards Earth System Science*, 2009, Vol. 9, 609-622

(Onwubolu et al. 2007) Onwubolu G.C. , Garimella S., Ramachandran V., Buadromo V., Abraham A. Self-organizing Data Mining for Weather Forecasting. Proceedings of IADIS European Conference Data Mining, Lisbon, Portugal, 2007, July 5-7, 81-88.
Available from:
http://www.softcomputing.net/ecdm07_1.pdf

(Orlanski, 1975) Orlanski I. A rational subdivision of scales of atmospheric processes, *Bulletin of American Meteorological Society*. 56, 1975, 527–530.

(Pabreja, 2005) Pabreja K. Data mining for spatio-temporal datasets with special reference to Interpretation of weather patterns, presented at International Brainstorming meeting on Modeling and Prediction over Indian Monsoon Region, National Centre of Medium Range Weather Forecast (NCMRWF), Noida, India, 2005, Feb 1-2.
Available from:
http://www.ncmrwf.gov.in/bs_meeting/kavita.pdf

(Pabreja, 2010a) Pabreja K. Application of Multidimensional Databases of Rainfall and Low Pressure Systems on OLAP-Based Model, Proceedings of the Second International Conference on Computer Modeling and Simulation, Sanya, China, 2010, Jan 22-24, 4:249-253.
Available from:

<http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5421470>

(Pabreja, 2010b) Pabreja K. Mapping of spatio-temporal relational databases onto a multidimensional data hypercube, Proceedings of Einblick – Research Paper Competition held during **Confluence 2010** organized by Amity University in association with EMC data storage systems (India) Pvt. Ltd., Noida, UP, India, 2010, Jan. 22-23, 127-133.

Paper was selected as the **BEST** paper and awarded the **FIRST** prize.

Available as news on EMC data storage systems (India) Pvt. Ltd website

<https://community.emc.com/thread/101383>

(Pabreja, 2010c) Pabreja K. Mapping of spatio-temporal weather relational databases onto a multidimensional data hypercube. In: Knowledge Management. (Lathar A.S., Saini A.K., Dhingra S., Eds.), Macmillan Publishers India Ltd., 2010, 253-265

(Pabreja, 2010d) Pabreja K. Pattern data warehouse for analysis of cloudburst events using OLAP-based data hypercube for storage of Numerical Weather Prediction Model outputs, *International Journal of Computational Intelligence and Research*. 2010, 6: 511-516.

(Pabreja, 2010e) Pabreja K. Identification of weather system from Numerical Weather Prediction output data using Data Mining. Proceedings of the 4th International Conference on Advanced Computing & Communication Technologies sponsored by IEEE, Oct 30, 2010, 630-635.

(Pabreja, 2009) Pabreja K. Data Mining – a tool in support of Analysis of rainfall on spatial and temporal scale associated with Low Pressure Systems formation, Proceedings of International Conference on Challenges and Opportunities in Agrometeorology (INTROMET 2009) (Under publication with Springerlink)

(Pal, et al., 2006) Pal N.R., Pal S., Das J., Lakshmanan V., Fuzzy Rule–Based Approach for Detection of Bounded Weak-Echo Regions in Radar Images, *Journal of Applied Meteorology and Climatology*, American Meteorological Society, 2006

Available from:

<http://journals.ametsoc.org/doi/pdf/10.1175/JAM2408.1>

(Paras et al., 2007) Paras, Mathur S., Kumar A., Chandra M. A Feature Based Neural Network Model for Weather Forecasting. *World Academy of Science, Engineering and Technology*. 2007, 34:66-73.

Available from:

<http://www.waset.org/journals/waset/v34/v34-13.pdf>

(Pedersen and Jensen, 2001) Pedersen T.B., Jensen C.S, *Multidimensional Database Technology*, Journal of IEEE, 2001, Volume 34, Issue 12

Available from:

<http://portal.acm.org/citation.cfm?id=621859>

(Premium Weather service) Premium Weather, The ultimate personal weather service

<http://www.tornadochaser.net/tornado.html>

(Rajeevan and Bhate, 2009) Rajeevan M., Bhate J. *A high resolution daily gridded rainfall dataset (1971–2005) for mesoscale meteorological studies*, Current Science, vol. 96, no. 4, 25, 2009 Feb.

Available from:

<http://www.ias.ac.in/currsci/feb252009/558.pdf>

(Riordan and Hansen, 2002) Riordan D., Hansen B.K. *A fuzzy case-based system for weather prediction*, International journal of engineering intelligent systems for electrical engineering and communications. 2002, ISSN 0969-1170, Vol. 10, 139–146

Available from:

http://collaboration.cmc.ec.gc.ca/science/rpn/publications/pdf/Riordan_Hansen_2002.pdf

(Roiger and Geatz, 2003) Roiger R., Geatz M., *Data Mining: A Tutorial-based Primer*. Addison Wesley, USA; 2003.

(Sen and Oztopal, 2001) Sen Z., Oztopal A. Genetic algorithms for the classification and prediction of precipitation occurrence, *Hydrological Sciences journal*, 2001, 46(2): 255-266.

(Shekhar et al. 2004) Shekhar S., Zhang P., Huang Y., Vatsavai R.R. Trends in Spatial Data Mining. In : *Data Mining: Next Generation Challenges and Future Directions* (Kargupta

H., Joshi A., Sivakumar K., Yesha Y., Eds.) , AAAI/MIT Press, Oct. 2004 (pg. no. are not available)

Available from:

http://www.spatial.cs.umn.edu/paper_ps/dmchap.pdf

(Short, 2005) Short N.M., *The Remote Sensing Tutorial* [CD-ROM]. Federation of American Scientists, Washington , 2005

Available from:

http://www.fas.org/irp/imint/docs/rst/Sect14/Sect14_1d.html

(Sikka, 2006) Sikka D.R., A study on the Monsoon Low Pressure Systems over the Indian Region and their Relationship with Drought and Excess Monsoon Seasonal Rainfall, May 2006.

(Singh, Ganju and Singh, 2005) Singh D., Ganju A., Singh A. *Weather prediction using nearest-neighbour model*. Current Science, April 2005, vol. 88, no. 8, 25: 1283-1289.

Available from:

<http://www.ias.ac.in/currsci/apr252005/1283.pdf>

(Singh and Roy,2002) Singh R.B., Roy S.S. Climate variability and hydrological extremes in a Himalayan catchment. Proceedings of the ERB and Northern European FRIEND Project 5 conference, Slovakia, 2002

(Sivanandam and Deepa, 2007) Sivanandam S.N., Deepa S.N. Introduction to Genetic Algorithms (available online), Springer, 2007.

(Skamarock et al. online) Skamarock W.C., Klemp J.B., Dudhia J., Gill J.O., Barker D.M., Duda M.G., Huang X., Wang W., Powers J.G. , A Description of the Advanced Research WRF Version 3

Available from:

http://www.mmm.ucar.edu/wrf/users/docs/arw_v3.pdf

(Smith) Smith S.E., Cloudburst, Available from:

<http://www.wisegeek.com/what-is-a-cloudburst.htm>

(Stojanova, Panov, and Koblar, 2006) Stojanova D, Panov P, Koblar A. *Learning to predict Forest Fires with different data mining techniques*, Proceedings of the Conference on Data mining and Data Warehousing (SiKDD 2006), Ljubljana, Slovenia, 2006

Available from:-

<http://kt.ijs.si/Dunja/SiKDD2006/Papers/Stojanova.pdf>

(Storch and Zwiers, 1999) Storch H.V., Zwiers F.W. *Statistical Analysis in Climate Research*. Cambridge University Press, 1999.

(Stolorz and Nakamura, 1995) Stolorz P., Nakamura H., *Fast Spatio-temporal Data Mining of Large Geophysical Datasets*. Proceedings of the First International Conference on Knowledge Discovery and Data Mining, AAAI Press, Montreal, Canada, 1995, Aug 20-21, 300-305.

Available from:

<http://citeseer.ist.psu.edu/stolorz95fast.html>

(Talman, 1930) Talman C.F., *Popular mechanics*, Hearst Magazines, Dec 1930

(Report by IMD, 2009) Tornado over Orissa on 31st March 2009 –A preliminary report, Indian Meteorological Department, 2009

(Tsalgalidis and Evangelidis, 2010) Tsalgalidis E., Evangelidis G. *The effect of Training Set selection in Meteorological Data Mining*. Proceedings of the 14th Panhellenic Conference on Informatics, 2010

Available from:

<http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=05600362>

(Uppala, Dee and Kobayashi, 2008) Uppala S., Dee D., Kobayashi S. Simmons A. *Evolution of reanalysis at ECMWF*, Proceedings of the Workshop by World Climate Research Programme, France, 2008

Available from:

http://wcrp.ipsl.jussieu.fr/Workshops/Reanalysis2008/Documents/V1-102_ea.pdf

- (Vatuiu and Popeanga, 2007) Vatuiu T., Popeanga, *Overview of Oracle OLAP and using SQL for manipulate Multidimensional data*, Annals of the University of Petrosani, Economics, 7, 2007, 355-362.
- (Wang, Li and Li, 2006) Wang F., Li H., Li R., Data mining with independent component analysis. Proceedings of the 6th World congress on intelligent control and automation, Dalian, China, 2006, Jun21-23, 6043-6047.
- Available from:
<http://ieeexplore.ieee.org/iel5/11210/36092/01714240.pdf?tp=&arnumber=1714240&isn umber=36092>
- (Ward, 1972) Ward N.B. The exploration of Certain features of Tornado dynamics using a laboratory model, *Journal of Atmospheric Sciences*. 1972, 29: 1194-1204.
- (Weka 3, downloaded from Internet) Weka 3- Data mining with open source machine learning software .
- Available from:
<http://www.cs.waikato.ac.nz/ml/weka/>
- (Witten and Frank, 2005) Witten I.H, Frank E., Data Mining Practical Machine Learning Tools and Techniques, second edition, Morgan Kaufmann Publishers, 2005
- (Wong and Yip, 2005a) Wong K.Y., Yip C.L. *An intelligent tropical cyclone eye fix system using motion field analysis*. Proceedings of the 17th ICTAI, Hong Kong, Nov. 2005, 652–656.
- (Wong and Yip, 2005b) Wong K.Y., Yip C.L. *Tropical cyclone eye fix using genetic algorithm with temporal information*. Proceedings of the 9th KES, LNAI-3681, Australia, Sept. 2005, 854–860.
- (Wong et al. 2004) Wong K.Y., Yip C.L., Li P.W. Tsang W.W. *Automatic template matching method for tropical cyclone eye fix*. Proceedings of the 17th ICPR, 2004, volume 3, 650-653.

(Xue et al. 2009) Xue S., Yang M., Li C., Nie J. Meteorological Prediction Using Support Vector Regression with Genetic Algorithms Proceedings of the 1st International Conference on Information Science and Engineering (ICISE2009), Nanjing, Jiangsu China, 2009, 4931-4935.

Available from:

http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5455025

(Yan, Lap and Wah, 2006) Yan W.K., Lap Y.C., Wah L.P. *Identifying Weather Systems from Numerical Weather Prediction Data*, Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), 2006 IEEE

Available from:

<http://www.computer.org/portal/web/csd/doi/10.1109/ICPR.2006.677>

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

(Zurada, 1992) Zurada J M., Introduction to Artificial Neural Systems, West Publishing, 1992.

Appendix A

.ctl file describing the .grd file

```
DSET D:\rf\rf1988.grd
TITLE 0.5 degranalyzed normal grids
UNDEF -999.0
XDEF 69 LINEAR 66.5 0.5
YDEF 65 LINEAR 6.5 0.5
ZDEF 1 linear 1 1
TDEF 366 LINEAR 1jan1988 1DY
VARS 1
rf 0 99 GRIDDED RAINFALL
ENDVARS
```

Appendix B

Program in FORTRAN to convert .grd file (input.grd) to .dat file (output.dat)

```
PROGRAM READ
PARAMETER(ISIZ=35,JSIZ=32)
DIMENSION RF(ISIZ,JSIZ)
OPEN(7,FILE='input.grd',
+ FORM='UNFORMATTED',
ACCESS='DIRECT',RECL=ISIZ*JSIZ*4,
+ STATUS='OLD')
OPEN(10,FILE='output.DAT',+FORM='FORMATTED', STATUS='UNKNOWN')
55 FORMAT(20F8.1)
C TAKE NDAY=366 FOR LEAP YEARS NDAY=365
DO 101 IDAY=1,NDAY
READ(7,REC=IDAY)((RF(I,J),I=1,ISIZ),J=1,JSIZ)
WRITE(10,55)((RF(I,J),I=1,ISIZ),J=1,JSIZ)
101 CONTINUE
STOP
```

Appendix C

Program in Visual Basic to organize gridded rainfall datasets into tabular format

```
Dim db As Connection
Dim rs As ADODB.Recordset

Private Sub Command1_Click()
Dim k, x As Double
Dim i, j As Long
Dim no As Long
Dim cdt, dt As Date
Dim fg As Boolean
Dim p As Integer
Set rs = New ADODB.Recordset
rs.Open "Select * from rainfall1989 where field1>0 and field1<=122", db
fg = false
dt = #6/1/1989#
i = 1
For p = 1 To 130
    j = 2
    If (fg = True And rs!field2 = 38.5) Then
        dt = DateAdd("w", 1, dt)
    End If

    For x = 66.5 To 90 Step 0.5
        db.Execute "insert into table1989 values(" & i & "," & rs!field1 & ",#" & dt & "#," &
rs!field2 & "," & x & "," & rs.Fields(j) & ")"
        i = i + 1
        j = j + 1
    Next
    fg = True
rs.MoveNext
Next p
MsgBox "Done"
End
End Sub

Private Sub Form_Load()
    Set db = New Connection
    db.Open "PROVIDER=Microsoft.Jet.OLEDB.4.0;Data Source=d:\db2.MDB;"
End Sub
```

Appendix D

Code in Visual Basic for calculation of vorticity and divergence using vertical and horizontal wind components of forecast

```
Dim db As Connection
Dim rs1 As ADODB.Recordset
Private Function findv(lo As Double, la As Double) As Double
rs1.Open "Select * from wind where long=" & lo & " and lat=" & la, db
findv = rs1!v
rs1.Close
End Function
Private Function findu(lo As Double, la As Double) As Double
rs1.Open "Select * from wind where long=" & lo & " and lat=" & la, db
findu = rs1!u
rs1.Close
End Function
Private Function findvd(lo As Double, la As Double) As Double
rs1.Open "Select * from wind where long=" & lo & " and lat=" & la, db
findvd = rs1!v
rs1.Close
End Function
Private Function findud(lo As Double, la As Double) As Double
rs1.Open "Select * from wind where long=" & lo & " and lat=" & la, db
findud = rs1!u
rs1.Close
End Function
Private Sub Command1_Click()
Dim k As Double, x As Double, a As Double, b As Double
Dim v1 As Double, v2 As Double, v3 As Double, va As Double, u1 As Double, u2 As Double,
u3 As Double
Dim v1d As Double, v2d As Double, v3d As Double, vad As Double, u1d As Double, u2d As
Double, u3d As Double
For a = 17.5 To 22.5 Step 0.25
For b = 83.5 To 88.5 Step 0.25
    v1 = findv(b + 0.25, a)
    v2 = findv(b - 0.25, a)
    u1 = findu(b, a + 0.25)
    u2 = findu(b, a - 0.25)
    v3 = (v1 - v2) * 2
    u3 = (u1 - u2) * 2
    va = v3 - u3
    u1d = findud(b + 0.25, a)
    u2d = findud(b - 0.25, a)
    v1d = findvd(b, a + 0.25)
    v2d = findvd(b, a - 0.25)
    v3d = (v1d - v2d) * 2
    u3d = (u1d - u2d) * 2
    vad = u3d + v3d
db.Execute "update wind set vor=" & va & ", div=" & vad & " where long=" & b & " and lat=" & a
Next
Next
```

```
Next
MsgBox "Done"
End Sub
Private Sub Form_Load()
Set db = New Connection
db.Open"PROVIDER=Microsoft.Jet.OLEDB.4.0;
Datasource=e:\tproj\F31FOR3112\vu1000.mdb;"
Set rs1 = New ADODB.Recordset
End Sub
```

Appendix E

Contents of.ctl file of WRF forecast

```
dset ^20090331_00.dat
undef 1.e30
title OUTPUT FROM WRF V3.1 MODEL
xdef 149 levels
76.70129395
76.78641510
76.87153625
76.95664978
77.04177094
77.12688446
77.21200562
77.29711914
77.38224030
77.46735382
77.55247498
77.63758850
77.72270966
77.80782318
77.89294434
77.97805786
78.06317902
78.14830017
78.23341370
78.31853485
78.40364838
78.48876953
78.57388306
78.65900421
78.74411774
78.82923889
78.91435242
78.99947357
79.08458710
79.16970825
79.25482941
79.33994293
79.42506409
79.51017761
79.59529877
79.68041229
79.76553345
79.85064697
79.93576813
80.02088165
```


80.10600281
80.19111633
80.27623749
80.36135101
80.44647217
80.53159332
80.61670685
80.70182800
80.78694153
80.87206268
80.95717621
81.04229736
81.12741089
81.21253204
81.29764557
81.38276672
81.46788025
81.55300140
81.63811493
81.72323608
81.80835724
81.89347076
81.97859192
82.06370544
82.14882660
82.23394012
82.31906128
82.40417480
82.48929596
82.57440948
82.65953064
82.74464417
82.82976532
82.91487885
83.00000000
83.08512115
83.17023468
83.25535583
83.34046936
83.42559052
83.51070404
83.59582520
83.68093872
83.76605988
83.85117340
83.93629456
84.02140808
84.10652924
84.19164276
84.27676392
84.36188507
84.44699860
84.53211975
84.61723328
84.70235443
84.78746796
84.87258911
84.95770264
85.04282379
85.12793732
85.21305847
85.29817200
85.38329315
85.46840668

85.55352783
85.63864899
85.72376251
85.80888367
85.89399719
85.97911835
86.06423187
86.14935303
86.23446655
86.31958771
86.40470123
86.48982239
86.57493591
86.66005707
86.74517059
86.83029175
86.91541290
87.00052643
87.08564758
87.17076111
87.25588226
87.34099579
87.42611694
87.51123047
87.59635162
87.68146515
87.76658630
87.85169983
87.93682098
88.02194214
88.10705566
88.19217682
88.27729034
88.36241150
88.44752502
88.53264618
88.61775970
88.70288086
88.78799438
88.87311554
88.95822906
89.04335022
89.12846375
89.21358490
89.29870605
ydef 139 levels
14.39208412
14.47451305
14.55691624
14.63928413
14.72162151
14.80393314
14.88620758
14.96845055
15.05066586
15.13284969
15.21500015
15.29711723
15.37920094
15.46125603
15.54327774
15.62526417
15.70722103
15.78914261

15.87103367
15.95288849
16.03471375
16.11650085
16.19825554
16.27997971
16.36166573
16.44331932
16.52494049
16.60652161
16.68807411
16.76958847
16.85106659
16.93251228
17.01392555
17.09529877
17.17663765
17.25793839
17.33920860
17.42043686
17.50163269
17.58279228
17.66391754
17.74500465
17.82605362
17.90706444
17.98804092
18.06897736
18.14987946
18.23074532
18.31157112
18.39236069
18.47311020
18.55382156
18.63449860
18.71513176
18.79573250
18.87628937
18.95681000
19.03729057
19.11773872
19.19814110
19.27850533
19.35882950
19.43911362
19.51936150
19.59956360
19.67972946
19.75985909
19.83994102
19.91998672
19.99999046
20.07995605
20.15987778
20.23976135
20.31960487
20.39940643
20.47916222
20.55887794
20.63855553
20.71819115
20.79778099
20.87733269
20.95684052

21.03630638
21.11573029
21.19511032
21.27444839
21.35374260
21.43299866
21.51220703
21.59137535
21.67049789
21.74957848
21.82861710
21.90760612
21.98655319
22.06546021
22.14432144
22.22313881
22.30191040
22.38064003
22.45932579
22.53796387
22.61655998
22.69510651
22.77361298
22.85206985
22.93048859
23.00885582
23.08717918
23.16545486
23.24368668
23.32187080
23.40001297
23.47810936
23.55615234
23.63415909
23.71211243
23.79001999
23.86788177
23.94569778
24.02346611
24.10118484
24.17886162
24.25648499
24.33406639
24.41159821
24.48908234
24.56651688
24.64390755
24.72124672
24.79854012
24.87578773
24.95298386
25.03013039
25.10722923
25.18428230
25.26128197
25.33823586
25.41514015
zdef 12 levels
1000.00000
950.00000
850.00000
750.00000
700.00000
550.00000

```

500.00000
250.00000
200.00000
150.00000
100.00000
50.00000
tdef      1 linear 00Z31MAR2009      360MN
VARS      14
U          12  0  x-wind component (m s-1)
V          12  0  y-wind component (m s-1)
W          12  0  z-wind component (m s-1)
PSFC       1  0  SFC PRESSURE (Pa)
U10        1  0  U at 10 M (m s-1)
V10        1  0  V at 10 M (m s-1)
SST        1  0  SEA SURFACE TEMPERATURE (K)
RAINC      1  0  ACCUMULATED TOTAL CUMULUS PRECIPITATION (mm)
RAINNC     1  0  ACCUMULATED TOTAL GRID SCALE PRECIPITATION (mm)
tc         12  0  Temperature (C)
rh         12  0  Relative Humidity (%)
ws10       1  0  Wind Speed at 10 M (m s-1)
wd10       1  0  Wind Direction at 10 M (Degrees)
slp        1  0  Sea Level Pressure (hPa)
ENDVARS
@ global String comment TITLE = OUTPUT FROM WRF V3.1 MODEL
@ global String comment SIMULATION_START_DATE = 2009-03-31_00:00:00
@ global String comment GRIDTYPE = C
@ global String comment MMINLU = USGS
@ global String comment WEST-EAST_GRID_DIMENSION = 150
@ global String comment SOUTH-NORTH_GRID_DIMENSION = 140
@ global String comment BOTTOM-TOP_GRID_DIMENSION = 38
@ global String comment MAP_PROJ = 3
@ global String comment DX = 9000.00
@ global String comment DY = 9000.00
@ global String comment CEN_LAT = 20.00
@ global String comment CEN_LON = 83.00
@ global String comment TRUELAT1 = 18.00
@ global String comment TRUELAT2 = 18.00
@ global String comment MOAD_CEN_LAT = 20.00
@ global String comment STAND_LON = 83.00
@ global String comment DIFF_OPT = 1
@ global String comment KM_OPT = 4
@ global String comment DAMP_OPT = 0
@ global String comment KHDIF = 0.00
@ global String comment KVDIF = 0.00
@ global String comment MP_PHYSICS = 3
@ global String comment RA_LW_PHYSICS = 1
@ global String comment RA_SW_PHYSICS = 1
@ global String comment SF_SFCLAY_PHYSICS = 1
@ global String comment SF_SURFACE_PHYSICS = 2
@ global String comment BL_PBL_PHYSICS = 1
@ global String comment CU_PHYSICS = 2
@ global String comment SURFACE_INPUT_SOURCE = 1
@ global String comment SST_UPDATE = 0
@ global String comment GRID_FDDA = 0
@ global String comment FEEDBACK = 1
@ global String comment SMOOTH_OPTION = 0
@ global String comment W_DAMPING = 0
@ global String comment OBS_NUDGE_OPT = 0
@ global String comment GRID_ID = 1
@ global String comment PARENT_ID = 0
@ global String comment I_PARENT_START = 1
@ global String comment J_PARENT_START = 1
@ global String comment PARENT_GRID_RATIO = 1
@ global String comment DT = 45.00

```

```

@ global String comment ISWATER =    16
@ global String comment ISICE =     24
@ global String comment ISURBAN =    1
@ global String comment ISOILWATER =  14

```

Appendix F

Program in C plus plus to convert the wrf forecast file (.dat file) to a .txt file

```

#include <stdio.h>
#include <conio.h>
#include <iostream.h>
#include <fstream.h>
struct rec
{
    float a[149];
};
int ReadFloat(FILE *fptr,float *n)
{
    unsigned char *cptr,tmp;
    if (fread(n,4,1,fptr) != 1)
        return(0);
    cptr = (unsigned char *)n;
    tmp = cptr[0];
    cptr[0] = cptr[3];
    cptr[3] =tmp;
    tmp = cptr[1];
    cptr[1] = cptr[2];
    cptr[2] = tmp;
    return(1);
}

void main()
{
    int i,j,k;
    float x[149], fl;
    FILE *f, *ft;
    char arr[4];
    struct rec r;
    ofstream writefile;
    writefile.open("31_06.txt" );
    clrscr();
    f=fopen("31_06.dat","rb");
    for (k=1;k<=69;k++)
    {cout<<"\nk= " <<k<<"\n";
        for (i=0;i<139;i++)
        { for (j=0;j<149; j++)
            {
                ReadFloat(f,&fl);
                x[j]=fl;
                writefile<<x[j]<<" ";
            }
        }
    }
}

```

```

        writefile<<"\n";
    }
}
fclose(f);
writefile.close();
cout<<"complete\n";
getch();
}

```

Appendix G

Implementation of point in polygon algorithm to update fact_rainfall table

Option Base 1

Dim db As Connection

Dim rs1 As ADODB.Recordset

Dim rs2 As ADODB.Recordset

Dim rid As Integer

Dim districtid() As Integer

Dim districtno As Integer

Dim vertx() As Double

Dim verty() As Double

Dim vertex As Integer

Private Sub Command1_Click()

For rid = 1 To districtno

Set rs1 = New ADODB.Recordset

ReDim vertx(1)

ReDim verty(1)

vertex = 0

If districtid(rid) <> 0 Then

rs1.Open "Select * from districtdetails where districtid=" & districtid(rid), db

While rs1.EOF = False

vertex = vertex + 1

ReDim Preserve vertx(vertex)

ReDim Preserve verty(vertex)

vertx(vertex) = rs1!Long

verty(vertex) = rs1!lat

rs1.MoveNext

Wend

Call pnpoly(vertex, vertx, verty)

End If

Next

End Sub

Private Sub Form_Load()

```

Set db = New Connection
db.Open "PROVIDER=Microsoft.Jet.OLEDB.4.0;Data Source=e:\association.mdb;"
Set rs1 = New ADODB.Recordset
rs1.Open "Select * from district", db
i = 1
rs1.MoveFirst
While rs1.EOF = False
ReDim Preserve districtid(i)
districtid(i) = rs1!districtid
districtno = i
i = i + 1
rs1.MoveNext
Wend
Text1.Text = districtno
rs1.Close
db.Execute "update fact_rainfall set districtid=NULL "

```

```
End Sub
```

```

Private Sub pnpoly(vertex As Integer, vertx() As Double, verty() As Double)
Set rs2 = New ADODB.Recordset
rs2.Open "Select * from locationnew", db
While rs2.EOF = False
ans = checkpt(vertex, vertx, verty, rs2!Long, rs2!lat)
If ans <> 0 Then
db.Execute "update fact_rainfall set districtid=" & rid & " where loc_key=" & rs2!loc_key
End If
rs2.MoveNext
Wend
MsgBox "ok"
End Sub

```

```

Private Function checkpt(vertex As Integer, vertx() As Double, verty() As Double, testx As
Double, testy As Double) As Integer
Dim i As Integer, j As Integer, c As Integer
c = 0
j = vertex
For i = 1 To vertex
If ((verty(i) > testy) <> (verty(j) > testy)) And (testx < (vertx(j) - vertx(i)) * (testy - verty(i)) /
(verty(j) - verty(i)) + vertx(i)) Then
If c = 0 Then
c = 1
Else
c = 0
End If
End If
j = i
Next
checkpt = c
End Function

```


List of Publications

Publications Made out of this Thesis Work are given below:-

Journal Publications (International Refereed) - 05

1. Pabreja K., Datta R.K. A data warehousing and data mining approach for analysis and forecast of cloudburst events using OLAP-based data hypercube, *International Journal of Data Analysis Techniques and Strategies, Inderscience Publishers*, 2012 - Vol. 4, No.1, pp. 57 – 82.
2. Pabreja K. Pattern data warehouse for analysis of cloudburst events using OLAP-based data hypercube for storage of Numerical Weather Prediction Model outputs. *International Journal of Computational Intelligence Research*, (ISSN 0973-1873) 2010, 6: 511-516.
3. Pabreja K. , Clustering technique to interpret Numerical Weather Prediction output products for forecast of Cloudburst, *International Journal of Computer Science and Information Technologies (IJCSIT)*, (ISSN: 0975–9646), 2012, Vol. 3 (1), pp.2996 – 2999.
4. Pabreja K., An Adaptive Neuro-Fuzzy Inference System based on Vorticity and Divergence for Rainfall forecasting , *International Journal of Computer Science and Information Security (IJCSIS)*, (ISSN: 1947-5500), 2011, Vol. 9, No. 12, pp.45-52.
5. Pabreja K. Multi dimensional data model of Biological datasets for Analysis and Prediction of Cancer. *GAMS Journal of Mathematics and Mathematical Biosciences (GAMSJMMB)*, published by Taylor and Francis, Co-published with Anamaya Publishers (accepted)

Chapter in book (International Refereed) - 01

1. Pabreja K., Data mining based on Neural Networks for Gridded Rainfall Forecasting, In : **Business Intelligence**, ISBN 979-953-307-595-1, published by **InTech - Open Access Publisher**, 97-108.

Conference Publications

International (Refereed) - 02

1. Pabreja K. Application of Multidimensional databases of Rainfall and Low Pressure Systems on OLAP-based model. The 2nd International Conference on Computer modeling and simulation (ICCMS 2010) organized by the International Association of Computer Science and Information Technology on January 22-24, 2010, Sanya, China Conference proceeding have been published by the Conference Publishing Service, which is available at **IEEE Xplore** and **CSDL**, and indexed by **EI Compendex** and **Thomson ISI**, 2010, 4:249-253.
2. Pabreja K. Data Mining – a tool in support of Analysis of rainfall on spatial and temporal scale associated with Low Pressure Systems formation. In: Challenges and Opportunities in Agrometeorology (ISBN : 978-3-642-19359-0), (Attri S.D., Rathore L.S., Sivakumar M.V.K., Dash S.K., Eds.), **Springer Heidelberg Dordrecht London New York**, 2011. (as one of the selected papers presented in Intromet 2009, International Conference organized by India Meteorological Society and co-sponsored by India Meteorological Department, Ministry of Earth Sciences, Ministry of Agriculture, Department of Space, Ministry of Environment and Forests, Department of Science and Technology, Indian Council of Agricultural Research and World Meteorological Organization, New Delhi, India, 2009, Feb 23-25), 287-297.

International - 03

1. Pabreja K. Data mining - a tool in support of weather interpretation, poster presented at International conference organized by SEARCC, CSI and CII at Chennai Trade Centre, Chennai, India, 2008, Sept 11-13.
2. Pabreja K. Multi dimensional data model for analysis of biological, climatological and atmospheric data. Presented at 1st IFIP International Conference on Bioinformatics organized by Sardar Vallabhbhai National Institute of Technology, Surat, India, 2010, March 25-28.
3. Pabreja K. Identification of weather system from Numerical Weather Prediction output data using Data Mining. Proceedings of the 4th International Conference on Advanced Computing & Communication Technologies (ISBN 93-80697-28-7), sponsored by IEEE, 2010, Oct 30, 630-635.

National - 04

1. Pabreja K. Mapping of spatio-temporal weather relational databases onto a multidimensional data hypercube. In: **Knowledge Management** (ISBN 023-032-937-3) (Lathar A.S., Saini A.K., Dhingra S., Eds.), **Macmillan Publishers India Ltd.**, 2010, 253-264.
2. Pabreja K. Mapping of spatio-temporal relational databases onto a multidimensional data hypercube. Proceedings of Einblick – Research Paper Competition held during Confluence 2010 organized by Amity University in association with EMC data storage systems (India) Pvt. Ltd., Noida, India, 2010, Jan 22-23, 127-133.
Paper was selected as the **BEST** paper and was awarded the **FIRST** prize.
3. Pabreja K. Artificial Neural Networks for Rainfall Forecasting. Presented at Fifteenth Annual National Conference of Gwalior Academy of Mathematical Sciences (GAMS) on theme “Mathematics and Development of ICT (including allied applications)”, New Delhi, India, 2010, Dec 12-14.
4. Pabreja K. Data Mining – a tool in support of Interpretation of Low Pressure System movement over Indian Region. Proceedings of the 3rd National Conference INDIACom-2009 Computing For Nation Development (ISSN 0973-7529 and ISBN 978-81-904526-6-3), organized by BVICAM, Computer Society of India, Institution of Electronics and Telecommunication Engineers and Guru Gobind Singh Indraprastha University, New Delhi, India, 2009, Feb 26 – 27, 631-634.

Biography of the candidate – **Kavita Kapoor**

Ms. Kavita holds more than 17 years of experience with Educational institution and Industry. She is currently Assistant Professor - Computer Society, Maharaja Surajmal Institute, an affiliate of GGS Indraprastha University. She had been the Head of Department - Computer Science from Aug'00 to Mar'05. She has teaching experience of over a decade and has taught subjects like Object Oriented Programming using C++, Database Management Systems, Systems Programming, Data mining and Data warehousing, Data structures, Computer Architecture, Knowledge Management in new economy, Mobile Computing. She has worked for more than 5 years with Indian as well as USA MNC. These companies include Rockwell International Overseas Corp. (Rockwell was one of the fortune 500 companies and world leader in Electronics control and Communication with annual revenues in excess of US\$ 10 billions), Parekh Microelectronics (I) Ltd., HCL Hewlett Packard Ltd. and Shyam Telecom Ltd. (Research and Development Unit).

She is M.S.(Software Systems) from BITS, Pilani; AMIETE (eq. B.E. (Electronics and Telecommunication Engg.)) from IETE. She was awarded **Merit Certificate** from **Govt. of India** for outstanding performance in D.S.S.E., CBSE in 1988.

She holds membership of many professional bodies *viz.* Senior Member of Computer Society of India, Member of Institute of Electronics and Telecommunication Engineers, Member of Indian Meteorological Society and Member of IACSIT, Singapore.

She has designed and developed Workbooks and textbooks for the **ICT Project, Punjab** undertaken by Educational Consultants India Ltd.

She has contributed **five papers in International Journals** and **one in an International publisher's Book**. She has presented her research work in **five International and four National conferences** of repute. Her paper "Mapping of spatio-temporal relational databases onto a multidimensional data hypercube" presented at Einblick – Research Paper Competition held during Confluence 2010 organized by Amity University in association with EMC data storage systems (India) Pvt. Ltd. on January 22-23, 2010 was selected as the **Best paper** and awarded the **FIRST prize**.

Biography of the supervisor- **Dr. Rattan K. Datta**

Dr. Datta is currently Honorary Research Director of Computer Society of India. He is CEO and Director, MERIT, an IT educational institute at Delhi. He is President – GAMS (Gwalior Academy of Mathematical Sciences). He is Vice chairman, TC5 of IFIP (International federation of information processing) a technical body of UNESCO.

Dr. Datta is first class M.Sc (Hons) in Physics from Punjab University and Ph.D from IIT Delhi , besides other professional qualifications. Dr Datta is fellow of Computer Society of India, Institution of Electronics and Telecommunication Engineers, India Meteorological Society, Telematic forum and member of Indian Science Congress (ISCA).

Dr. Datta is former National president of computer society of India (CSI), Indian Meteorological Society and IT section of ISCA. He was Adjunct professor in computer science & engineering at Delhi University, GGS IndraPrastha University and SBBS Institute of engineering & technology.

Dr Rattan K. Datta was Advisor, Department of Science & Technology, Government of India, project Coordinator & founder head of National Centre for Medium Range Weather Forecasting , a center involved in development of global weather prediction models for weather 3-10 days in advance and acquisition , installation & management of the first supercomputing facility in India.

In 1970, Dr.Datta was selected for UN fellowship to study in USA and Japan. In 1973, he was invited to USSR under the program of exchange of experts. During 1976-78, he was invited to work as United Nation expert on data processing and meteorological adviser in East Africa.

During Monsoon Experiment (MONEX-79) as a part of First global GARP Experiment (FGGE) conducted under the aegis of World Meteorological Organization, he was the chief scientist from India to coordinate the scientific experimentation. He was later Director, International Monsoon Data Management Centre.

He has traveled widely; specifically he was delivering a course on monsoon in University of Miami from 1980 to 1992 every alternate year. He has also lectured at ICTP Trieste and was also invited by the “Pontificas Vatican Academy” in Sept, 1986. His lecture has been published as a monogram of the academy.

His research interests include Numerical Weather Prediction, Application of Intelligent systems for interpretation like Data mining. He has contributed over 100 research papers besides a number of popular articles in various magazines and news papers. He was awarded a gold medal by the Govt. of India in 1975 for his research contributions during 1972-1973. Dr Datta has written two and edited three books.

The International Biographical Centre, Cambridge, UK, nominated Dr Datta as the International man for the 1992-93. On 9th December 2008, the 20th Anniversary of NCMRWF, Dr. Datta was felicitated by Ministry of Earth Sciences Govt. of India for his Outstanding Research. Dr Datta was awarded the “Sidha Sewa Puraskar” as eminent Scientist on behalf of Swami Hardas Foundation on the occasion of Glorious World Day in March 2010. Dr. Datta was felicitated by Chancellor, Lingaya’s University for his “Exceptional Contributions and Leadership in promoting Computer Technology and Education in India” in November 2010 during International Conference on Reliability, Infocom Technology and Optimization.