Chapter-2

Data and Methodology

2.1 Introduction

To study the prediction of Indian summer monsoon rainfall on smaller space-time scales, we have used long period (1951-2012) data of meteorological parameters (temperature, pressure, rainfall, zonal wind, geo-potential height etc). Seasonal values for all meteorological parameters, except Indian rainfall, are computed by averaging the respective parameters for corresponding months as follows:

- Winter : December-January-February
- Spring : March-April-May
- Summer : June-July-August
- Autumn : September-October-November.

Monsoon seasonal rainfall is the rainfall accumulation over June to September.

2.2 Datasets used in this study

2.2.1 Indian summer monsoon rainfall (ISMR)

Monthly rainfall data for monsoon season (June through September) for India as a whole, its homogeneous regions and sub-divisions (Figure-2.1) have been taken from the web site of Indian Institute of Tropical Meteorology, Pune, www.tropmet.res.in. The seasonal rainfall series is prepared by adding the rainfall
from June to September. The percentage departures from long term (1871-2012) mean are calculated and have been used in further analysis.

Figure-2.1 depicts thirty six sub-divisions over India as given by India Meteorological Department (IMD). The abbreviations for sub-divisions are given in brackets. The sub-divisions in the hilly regions (number 2, 12, 15 and 16), Andaman and Nicobar islands (no 1) and the Lakshdweep islands (no 36) are not considered in this study. The figure also depicts five homogeneous regions of India as given by Indian Institute of Tropical Meteorology (IITM).
2.2.2 Temperature, geo-potential height and zonal wind

The NCEP/NCAR Reanalysis gridded 2.5° x 2.5° long/lat global temperature (Kelvin), geo-potential height (meter), zonal wind (ms⁻¹) data for all pre-monsoon months and seasons (January, February, March, April, May, Winter, Spring) at surface, 850-hPa, 500-hPa, 200-hPa for 1948-2012 have been taken from http://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html. The data have been interpolated on 5° x 5° lat/long.

2.2.3 Effective Strength Index (ESI)

Monthly North Atlantic Oscillation (NAO) and Southern Oscillation (SO) data for the period 1951-2012 have been taken from www.cpc.ncep.noaa.gov. Effective Strength Index (ESI) is defined as the algebraic difference between monthly indices of NAO and SO. The anomalies from the annual mean have been calculated for each month and these anomaly series are then divided by the standard deviation. These series are called as effective strength index (ESI) series of respective month. ESI tendency from winter to spring is defined as the difference April minus January ESI values.

2.2.4 Sea Surface Temperature (SST)

SST anomaly data over Nino1+2 (0-10S; 90W-80W), Nino3 (5N-5S; 150W-90W), Nino4 (5N-5S; 160E-150W), Nino3.4 (5N-5S; 170-120W) and North Atlantic Ocean (5-20N; 60-30W) for the period 1951-2012 have been taken from the website http://www.cpc.ncep.noaa.gov/data/indices/.

2.2.5 Frequency of El Nino and La Nina events

The El Nino (warm) and La Nina (cool) events in the tropical Pacific are identified by using Oceanic Nino Index (ONI). It is the running 3-months mean SST anomaly for the Nino 3.4 region (i.e., 5°N-5°S, 120°-170°W). El Nino (La Nina) events are identified if SST anomalies for 5 consecutive months are at or above (below) the +0.5° (-0.5°) C. The threshold is further broken down into Weak (with a 0.5 to 0.9
SST anomaly), Moderate (1.0 to 1.4) and Strong (≥ 1.5) events. Frequencies of El Nino and La Nina years, for the period 1951-2012, have been taken from the website http://ggweather.com/enso/oni.htm.

2.2.6 Arctic Oscillation (AO)
The Arctic Oscillation (AO) is a large scale mode of climate variability, also referred to as the Northern Hemisphere annular mode. It is a climate pattern characterized by winds circulating counterclockwise around the Arctic at around 55°N latitude. During positive AO, a ring of strong winds circulating around the North Pole which acts to confine colder air across Polar Regions. This belt of winds becomes weaker and more distorted in the negative phase of the AO, which allows an easier southward penetration of colder, arctic air-masses and increased storminess into the mid-latitudes. Monthly Arctic Oscillation (AO) data for the period 1951-2012 have been taken from the website http://www.cpc.ncep.noaa.gov/products/precip/CWlink/daily_ao_index/monthly.a o.index.b50.current.ascii.table.
The indices are constructed by projecting monthly mean 1000-hPa height anomalies onto the leading EOF mode. The time series is normalized by the standard deviation of the monthly index (1979-2000 base period).

2.3 Methodologies
The main goal of the study is to develop equations for LRF of ISMR to improve the prediction skill. Here the multiple regression equation has been developed to predict ISMR. For any regression equation, sufficient number of predictors is required. In search of new predictors at various pressure levels, we try to get homogeneous clusters for the meteorological parameter fields. To identify the clusters the method of Shared Nearest Neighbor has been used.
2.3.1. Shared Nearest Neighbor (SNN)

Cluster analysis has been used to group the data points into useful or meaningful groups (clusters). Clustering has a long history and a large number of clustering techniques have been developed. For high dimensional data like meteorological data, traditional clustering methods like K-means algorithm and agglomerative hierarchical technique do not perform well because the data contain outliers and clusters of different sizes, shapes and densities. In high dimensional datasets, the traditional Euclidean notion of density, which is the number of points per unit volume, is meaningless. Therefore traditional density based method cannot be used to identify core points in high density regions. An alternative to this is to define the similarity (closeness) between a pair of points in terms of their shared nearest neighbors. The technique was first developed by Ertoz et al (2003) to find the clusters of different sizes, shapes, and densities in noisy, high dimensional data. Steinbach et al (2003) have used SNN to discover new climate indices using meteorological parameters like SST, SLP precipitation. Boriah et al (2004) have applied this SNN technique to ocean temperatures and used the clusters to predict land temperatures. The algorithm based on these ideas eliminates noise (low-density points) and builds clusters by associating non-noise points with representative or core points (high-density points). This approach handles many problems like finding clusters in the presence of noise and outliers, finding clusters in data that have clusters of different shapes, sizes and density, etc. In SNN, similarity is confirmed by the common nearest neighbours.

The steps involved in the SNN clustering algorithm are:

1. Preprocessing of data: Seasonality is removed by computing anomalies. Any trend or autocorrelation is removed from time series at each grid if it exists.
2. To compute the similarity matrix. The correlation coefficient (CC) between the time series over two grid points, for 1951-2012, is the typical measure of similarity.
3. To sparsify the similarity matrix by keeping only its $k$ strongest links. Here $k$
is called the neighbourhood list size. It is the most important factor as it adjusts the focus of the clusters. If $k$ is too small, even a uniform cluster will be broken up into pieces due to local variations in the similarity, and the algorithm will tend to find many small, but tight clusters. On the other hand, if $k$ is too large, then the algorithm will tend to find only a few large, well-separated clusters and small local variations in similarity will not have an impact. In the SNN, a point can be similar to at most $k$ other points.

4. To construct the SNN list from the sparsified similarity matrix. At this point, we could apply a similarity threshold and find the connected components to obtain the clusters (Jarvis–Patrick algorithm). For each grid point, a neighbor list containing grid points which show significant (at 1% level) CC is prepared. Jarvis and Patrick (1973) suggested that a link is created between points $P$ and $Q$ if and only if both $P$ and $Q$ have each other in their closest $k$ nearest neighbor lists, where $k$ is the nearest neighbor size (here $k = 100$).

Let $i, j$ be two points. Then strength of the link between $i$ and $j$ is calculated as $\text{Str}(i, j) = \sum (k + 1 - m)(k + 1 - n)$, where $k$ is the nearest neighbor list size, and $m$ and $n$ are positions of SNNs in the lists of $i$ and $j$.

5. To find the SNN density of each point. Here we consider the sum of link strengths for every point. The points having high total link strength will become candidates for representative points, while those having very low total link strength become candidates for noise points.

6. To find the core points. These are the points having SNN density greater than the threshold value. This value is to be decided by trial and error. (Here we have taken it to be 100).

7. To form clusters from the core points by averaging the grids in the cluster.
2.3.2 Cross-validation scheme

It is a common experience that any regression equation developed for prediction performs well on the dataset from which it is obtained (training or dependent dataset) but may perform poorly on test dataset (independent dataset). The skill of prediction equation depends upon its performance on independent dataset. For this purpose the cross-validation scheme is applied.

In cross-validation procedure, the data are divided into two disjoint sets: one is treated as dependent sample for developing (or “training”) the model and the other is the independent sample for verifying (or “testing”) the estimated forecasts. Suppose the total independent sample is partitioned into $K$ mutually exclusive sets of equal size; let the $k$th partition be $I_k$. Any one of these partitions can be considered the independent set, while the complement of this is considered as the dependent set ($D_k$). The quality of the resulting forecasts can be measured by the mean-square forecast errors and anomaly correlation coefficients in the independent set.

The distinction between dependent and independent error variances becomes clear for small sample sizes $N$. For large $N$, the two error estimates approach the same value. If $P$ is number of predictors then as $P/N$ increases, the expected error in the dependent data tends to be negligible while the expected error in the independent set approaches infinity. Davis (1976) derived similar results in the context of autoregressive models for moderately large $N$. Therefore for a good predictive model, minimum number of predictors should be used to achieve reasonably good forecasts in independent set. Delsole and Shukla (2002) proposed cross-validation scheme to select best linear prediction model for ISMR. Cross-validation procedure first screens out all models that are likely to perform poorly on independent datasets then the prediction error of each model is compared with those of all other models to determine whether the difference in error variance exceeds some threshold of significance. Sahai et al. (2002) used this scheme for
selecting the best predictor set. Cross-validation scheme involves following steps for selecting best predictor set for predictions in independent datasets.

1. Model development data (N number of years) is divided into two mutually exclusive sets namely independent set consisting of one year and dependent set consisting of remaining N-1 years, which is also known as “leave one out” method. The data are for model development period (1951-2012) are divided in N (=62) mutually exclusive dependent and independent sets in which each independent set consists of one year and remaining N-1 years are in dependent set.

2. For each predictor we get N independent predicted values along with the observed ones. Values of root mean square error (RMSE) are calculated for each predictor and the predictor having minimum RMSE value is selected as first predictor.

3. After selecting the first predictor, trial regression equations are again constructed using first selected predictor in combination with the remaining P-1 predictors. The combination of two predictors having minimum RMSE value is selected and this procedure is repeated for all P predictors.

4. Minimum RMSE value is plotted against number of predictors and the set of predictors associated with number of predictors having least RMSE value is considered as best predictor set.

### 2.4 Summary

The various datasets used in this study are explained in terms of its source, data length, grid size etc. In order to understand the strong collective force of various meteorological parameters on ISMR, first the cluster regions of these meteorological parameters at different pressure levels are identified using SNN technique. The set of parameters showing best prediction of ISMR is obtained by selecting the cluster
parameters by cross validation procedure. The details of SNN and cross validation methods are elaborated in this chapter.