Chapter 4

Prediction of Indian summer monsoon rainfall using Nino Indices: a neural network approach

4.1 Introduction

The problem of seasonal climate prediction is one whereby the average behaviour of the weather, or its component, over the span of a season is simulated. Seasonal forecasting using GCMs is relatively recent and the atmospheric models used are similar in nature to what is used in weather forecasting. The very nature of the atmosphere makes it unpredictable beyond a few days, or a week. The reason for the interest in seasonal prediction comes from the fact that the ocean, being ‘inert’, imparts some memory to the atmosphere. The simplest picture is that of atmosphere driving the ocean (Hasselmann, 1976). However, in the tropics, the other direction is also feasible. Although the atmosphere has much high frequency components, there are low-frequency variations, which result from the ocean-atmosphere interaction and the processes resulting therein such as the El Niño and the La Nina. The slower changes in the sea surfaces temperature result in imparting a degree of predictability to the atmosphere (Anderson, 2008). The atmospheric predictability is higher in the tropics than in the mid-latitudes owing to the stronger coupling with the ocean in the tropical region. This factor has resulted in the tropics acquiring more attention from the research community than the seasonal prediction in the extra-tropics. Both, the GCMs and the statistical prediction techniques have been rigorously used for the seasonal climate prediction of the Indian subcontinent.
Seasonal climate prediction is important for ground level users in the fields such as agriculture, energy, finance and insurance, fisheries, food security, health, water resource management and transport.

The relationship between the tropical central eastern pacific SST and Asian monsoon has been extensively studied by researchers since Walker (1923, 1924) defined the Southern Oscillation, which is closely related to the zonal circulation over the equator. It was further observed that Walker circulation is strongly coupled with underlying oceans, especially the SST in the tropical central eastern Pacific (Bjerknes, 1969). Studies have been carried out to investigate the physical mechanisms responsible for the ENSO-monsoon association (Rasmusson and Carpenter, 1983; Webster and Yang, 1992; Ju and Slingo, 1995; Kirtman and Shukla, 1997).

The relationship between El Niño events and Indian monsoon has been studied by many researchers (Shukla and Paolino, 1983; Shukla and Mooley, 1987; Sikka, 1980; Barnett, 1983; Mooley and Parthasarathy, 1984; Krishnamurthy and Goswami, 2000).

These studies reveal that Indian summer monsoon is weaker (stronger) than the normal before (after) the peak of an El Nino in winter, and that the relationship is opposite for the monsoon and La Nina. It is also observed that the monsoon circulation over southern Asia is generally weaker (stronger) than normal during El Nino (La Nina) summers (Webster and Yang, 1992; Lau and Yang, 1997).

It is recently pointed out that El Niño events with the warmest SST anomalies in the central Pacific are more effective in focusing drought, as they are produced by subsidence over India than events with the warmest SST anomalies in the eastern equatorial Pacific (Krishna Kumar et. al., 1999).
With the success of the back propagation rule *(Bishop, 1995)* for finding the derivatives, the ANN became a popular tool in artificial intelligence, robotics, and other allied fields *(Crick, 1989)*. Comparison of ANN and traditional techniques of forecasts is now a popular area of research.

ANN has been widely used to the study of various Pacific Ocean phenomena *(Tangang et. al., 1997; Hsieh and Tang, 1998; Tang et. al., 2000; Tang et. al., 2001; Wu et. al., 2006)*. However, the utility of the model in identifying connections between the Pacific Ocean phenomenon and the ISMR has not been investigated.

Recently, different indices associated with different regions in the tropics and extra-tropics at different levels of the atmosphere have been combined objectively to predict the ISMR using the ANN model *(Sahai et. al., 2003)* and it was observed that the model predicted reasonably well.

In the present study two statistical models, regression and artificial neural networks have been employed for predicting the Indian Summer Monsoon Rainfall (ISMR) using lagged relationship between the ISMRI and the various combinations of Niño indices.

This chapter is arranged in four sections; section 4.2 describes the theoretical concepts of artificial neural networks. Section 4.3 presents the data and method of prediction. Results of the application of models and the discussion of results are presented in section 4.4. Concluding remarks are mentioned in section 4.5.

**4.2. Artificial Neural Networks**

Artificial Neural networks (ANNs) are massively parallel, distributed processing systems representing a new computational technology built on the analogy to the human information processing system *(Palit and Popovik, 2005)*.
The major tasks that a neural network performs are classification, auto-association and time-series prediction (Masters, 1993). These are specialized forms of function approximation. Some of the traditional techniques to perform function approximation are splines, rational approximations, continued fractions. These techniques are particularly suited for the approximation of “well-behaved” functions with low frequency sinusoidal components in the Fourier domain. On the other hand, ANNs are useful for the approximation of functions that do not, in general, follow a set pattern and contain high frequency harmonics. ANNs use the “recall” methods employed by the human brain for mapping real world data (Zurada, 2002).

The mathematical model of Artificial Neural Network (ANN) consists of simple processing units called neurons (Haykin, 2002) interconnected by abstract links called “weights”. Feed-forward neural networks are a special type of ANN models, which consists of a layered assembly of neurons. Each neuron in a layer is connected to all the neurons in the next layer. The input layer receives the input patterns and the output layer gives the approximated function. Each neuron in the hidden layer receives the weighted outputs from all the neurons in the input layer. The total amount of input received is called the “activation”. This activation value is scaled down by a “transfer function”, which is a characteristic function of all the neurons except the input neurons. The output neuron receives this scaled value, again multiplied by the corresponding weight. The output neuron has its own transfer function which finally gives the network output.

There are essentially three steps in the optimization by ANN. In the “training” phase, a part of the entire dataset is used to develop the model (Haykin, 2002). In this phase, the free parameters on the connections between neurons are “optimized” by minimizing the least squared errors between the ANN output and the observed quantity for the corresponding input. The next step is called the “validation” (Haykin,
in which the error is calculated on a second set of data which forms disjoint set with the set used in the training phase. In this phase, however, optimization of the weights does not take place. This is a necessary step to ensure that the network does not end up “over fitting” the training data (Haykin, 2002). This technique is called the “hold out method” (Haykin, 2002; Bishop, 1995). If the set available for training is not too large, it is useful to take different subsets as validation sets, so that the same data is sometimes used in learning sets and sometimes for validation sets. This is called “cross-validation.”

The last step is the “test” phase in which the model is applied on yet another set, which is disjoint to the two sets of data used in the training and validation phases, called the “test” set. It is essentially a hind cast that is done during the test phase. If the performance of the model on the test set is acceptable, the model is said to be developed and can be used on future data.

4.3 Data and Method of prediction

The observed rainfall data used is the grided rainfall data (1° x 1° resolution) from India Meteorological Department (IMD) for the period 1951-2003 based on 1803 station (Rajeevan et. al., 2006). The time series corresponding to 1951-2003 is used for our analysis. The Indian Summer Monsoon Rainfall Index (ISMRI) has been constructed by calculating anomaly series and averaging over the summer season i.e. June-July-August-September (JJAS).

The extended reconstructed SST (ERSST, version 2.0, 2° X 2° resolutions) has been used in the present study (Smith and Reynolds, 2004). This data set contains global record of monthly SST from 1871 to 2003. But, due to the uncertainties and data scarcity in the earlier data set of 19th century the temporal coverage has been considered only from 1950 to 2003.
Table 4.1 show the areas, the SST of which is used as various Niño indices (Niño-1+2, Niño-3, Niño-3.4 and Niño-4) (Trenberth, 1997; Trenberth and Stepaniak, 2001). The actual indices have been calculated as the area-averaged sea surface temperature anomalies (°C) for the specified region.

<table>
<thead>
<tr>
<th>Niño Region</th>
<th>Range Longitude</th>
<th>Range Latitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>1+2</td>
<td>90°W – 80°W</td>
<td>10°S – 0°</td>
</tr>
<tr>
<td>3</td>
<td>150°W – 90°W</td>
<td>5°S – 5°N</td>
</tr>
<tr>
<td>3.4</td>
<td>170°W – 120°W</td>
<td>5°S – 5°N</td>
</tr>
<tr>
<td>4</td>
<td>160°E – 150°W</td>
<td>5°S – 5°N</td>
</tr>
</tbody>
</table>

El Niño has been quantified in terms of simple indices as corresponding to times when (SST) anomalies in the Niño-3 region exceeded 0.5 °C or when those in the Niño-3.4 region exceeded 0.4 °C, which are evidently enough to produce perceptible impacts in Pacific rim countries (Trenberth, 1997).

4.3.1 Principal Components Analysis

Principal components analysis (PCA) is widely used for multivariate data analysis (Wilks, 1995). PCA is particularly suited to the application wherein the data records are few compared to model complexity. ANN models are such models where the number of free parameters is large. If the training set is not large enough to accommodate for the reasonable optimization of those parameters, the model tends to over-fit the training population. In the present study, there are 5 predictors, which imply 5 neurons in the input layer. The output layer has only one neuron. Considering a small network with 3 neurons in the hidden layer, the number of weights will be 18.
The number of data points that are available (40) is insufficient to optimize 23 free parameters without the danger of over-fitting. PCA can thus be done to reduce the number of predictors, and hence the number of input neurons. This will considerably reduce the number of free parameters in the network and avoid the danger of over-fitting. The exact number of free parameters is obtained after fixing the number of neurons in the hidden layer which is done by suitable error analyses as described in the next section.

In the present study the entire data set of 53 points is divided into two parts comprising 40 and 13 points. The first 40 points correspond to the training data of the ANN. It is to be noted that each point lies in a five-dimensional space corresponding to the five predictors. The PCA has been done on the training data only. The Eigen vector corresponding to the largest Eigen value of the covariance matrix of the predictors is determined. It is found that the first principal component explains 77% of the variance. This PC can thus be used to represent the data in one-dimensional space. The remaining 13 points (called the test points in the ANN case) are still in 5-dimensional space. We have projected these 13 test points one at a time on the determined Eigen vector. This gives us the test points in one-dimensional space. These projected points are then used to predict the next points using the ANN.

4.3.2 ANN model for the present study

Multi-layer feed-forward neural network with error back propagation algorithm and delta-learning rule (Zurada, 2002) has been used for the prediction of ISMR with the first PC as the predictor. The model consists of three layers, the input layer having one neuron, the output layer having one neuron and the hidden having three neurons. We have taken only one hidden layer as it has been established that a single hidden
layer is sufficient to approximate any non-linear function to arbitrary accuracy (Cybenko, 1989).

The number of neurons in the hidden layer is a factor that significantly affects the performance of the model. Too few neurons in the hidden layer will starve the network of the necessary resources needed to solve the problem and too many neurons will force the network to learn the noise present in the training data set (Haykin, 2002) instead of capturing the underlying deterministic relationship. For deciding the number of neurons in the hidden layer, an objective criterion, based on minimum prediction error for the validation set, was followed. Since the number of samples in the population is small, cross-validation technique has been used. The data set was divided into three sets viz estimation set (35 patterns), cross-validation set (5 patterns) and the test set (13 patterns). Note that the training data is the first PC as determined above. Projecting the test points on the first PC of the training set forms the test data.

For a given number of neuron n, after training the network on the estimation subset, predictions were made for the validation set, and Root Mean Square Error (RMSE) was computed over this set. As the number of neurons was increased, RMSE decreased, but started increasing as n exceeds three. We took different subsets for validation sets, using the same data sometimes in learning sets and sometimes for validation sets. The optimum number of neurons remained the same. This means that the number of free parameters, if PC1 is used as predictor, is eight.

The sigmoid function, defined below has been for the transfer function of all the neurons (Haykin, 2002)

\[
g(a) = \frac{1}{1 + e^{-a}}
\]  

(4.1)
Where \( a \) is the activation at the neuron.

To protect the output of a neuron from being driven to saturation, the input to it, i.e. the activation, must not be too small or too large. The input is, thus, often normalized in the range (0, 1) (Palit and Popovik, 2005). In the present study, however, the data is normalized in the range (0.2, 0.8) using the following normalization equation:

\[
x_n = \frac{(x - x_{\text{min}})}{(x_{\text{max}} - x_{\text{min}})} * 0.6 + 0.2
\]

(4.2)

where \( x \) is the original data value, \( x_{\text{max}} \) is the maximum data value, \( x_{\text{min}} \) is the minimum data value and \( x_n \) is the normalized value of the data.

### 4.3.3 Regression Models for the present study

Different regression models (Wilks, 1995) have been constructed for predicting the ISMR with different predictors. Following regression equations summarise these models:

\[
ismr_i = -1.12 + 0.35 * nino3_{t-4}
\]

(4.3)

\[
ismr_i = -7.09 + 0.55 * nino3_{t-5}
\]

(4.4)

\[
ismr_i = -4.16 + 0.44 * nino3.4_{t-4}
\]

(4.5)

\[
ismr_i = -9.40 + 0.62 * nino3.4_{t-5}
\]

(4.6)

\[
ismr_i = -11.09 + 0.67 * nino4_{t-5}
\]

(4.7)

\[
ismr_i = 1.96 + 0.61 * nino3_{t-4} - 1.0 * nino3_{t-5} - 0.45 * nino3.4_{t-4} + 2.18 * nino3.4_{t-5} - 1.10 * nino4_{t-5}
\]

(4.8)
\[ ismr_t = 7.72 + 0.18 \times PC_i \]  \hspace{1cm} (4.9)

where

\[ ismr_t = ISMR \text{ index at time } t \]

\[ nino3_{t-i} = \text{nino3 index } i \text{ seasons earlier} \]

\[ nino3.4_{t-i} = \text{nino3.4 index } i \text{ seasons earlier} \]

\[ nino4_{t-i} = \text{nino4 index } i \text{ seasons earlier} \]

\[ PC_i = \text{First principal component of all the above indices with respective lags} \]

Overall, seven regression models have been developed. Five models for individual predictors, one multiple regression model for all the predictors combined and finally a regression model using the first principal component of the predictors as the only predictor of ISMR. 40 patterns were used in the training set for the determination of coefficients and 13 were used for the evaluation of model on unseen patterns (test set). The skills of these regression models have been evaluated using the root mean square error (RMSE) and correlation coefficient (r) with the observed precipitation. The standard deviation (std) of the observed and predicted precipitation have also been analysed.

**4.4 Results and Discussion**

Fig. 4.1 (a-d) shows the correlation coefficients of the Niño indices with ISMRI. The negative numbers in brackets on the x-axis indicate that the concerned SST is recorded prior to the onset of monsoon. The positive numbers indicate that the SST is
recorded after the onset of monsoon and hence these cases are not used for the correlation analysis.

**Fig. 4.1 a**

**Fig. 4.1 b**
Figure 4.1(a-e) Correlation coefficient ($r$) between India Summer Monsoon Rainfall Index (ISMRI) and (a) Niño 1+2 index, (b) Niño 3 index, (c) Niño 3.4 index and (d) Niño 4 index with Nino indices lagging by 1-8 season(s).
The following predictors have been found to have a correlation with the ISMRI for period 1951-1990 with a confidence level of 99%:

Niño3 index with a lag of 4 seasons \( (r = 0.35) \),
Niño 3 index with a lag of 5 seasons \( (r = 0.43) \),
Niño 3.4 index with a lag of 4 seasons \( (r = 0.37) \),
Niño 3.4 index with a lag of 5 seasons \( (r = 0.48) \) and
Niño 4 index with a lag of 5 seasons \( (r = 0.39) \).

Hence these have been chosen as the predictors. Niño 1+2 index did not show a confidence level of greater than 95% for any of the lag values. Fig. 4.2 (a-e) shows the plots of various Niño indices varying with the ISMRI. The plot of PC1 of the predictors varying with the ISMRI is shown in Fig. 4.2 (f).
Fig. 4.2 b

Fig. 4.2 c
Figure 4.2(a-f) Time series of observed ISMRI normalized anomaly (solid line) and (a) Niño – 3 index with a lag of 4 seasons, (b) Niño – 3 index with a lag of 5 seasons, (c) Niño – 3.4 index with a lag of 4 seasons, (d) Niño – 3.4 index with a lag of 5 seasons, (e) Niño – 4 index with a lag of 5 seasons (dashed line) and (f) PC1 of all the indices.

As described in section 4.3.3 (equations 4.3 – 4.9), seven regression models have been developed for predicting the ISMR from the predictors mentioned above. The ANN model was trained using the PC1 of all the predictors as the input and the ISMRI anomaly as the output.

Fig. 4.3 (a-c) shows the comparison of model output ISMRI with observed ISMRI in training and test case, Fig. 4.3(a) for multiple regression model, Fig. 4.3(b) for PC regression model and Fig. 4.3(c) for ANN model. The difference of ISMRI between model output and observed data are shown with the Fig. 4.4(a) for multiple regression model, Fig. 4.4(b) for PC regression model and Fig. 4.4(c) for ANN model.
It may be noted that there were two El-Nino years in the period taken for the test cases viz, 1991 and 1997 and two La-Nina years viz, 1995 and 1998. It can be seen from Fig. 4.4(a-c) that the ANN model best predicted the 1991 precipitation and 1997 precipitation was predicted by both multiple linear regression model and PC1 regression model. The ANN model did not capture the 1997 seasonal precipitation. For the La-Nina years the 1995 precipitation was best predicted by the ANN model and the 1998 precipitation was best predicted by ANN model, although the multiple regression model also predicted with comparable skill. Fig. 4.5 shows the plot of observed precipitation anomaly and the Fig. 4.5(a) that with multiple linear regression predicted precipitation anomaly for training, multiple linear regression predicted precipitation anomaly for test (Fig. 4.5(b)), PC1 predicted precipitation anomaly for training (Fig. 4.5(c)), PC1 predicted precipitation anomaly for test [Fig. 4.5(d)], ANN predicted precipitation anomaly for training (Fig. 4.5(e)), ANN predicted precipitation anomaly for test cases (Fig. 4.5(f)). From Fig 4.5 (b,d,f), it is noted that in test case the multiple linear regression model have same sign as observed for 10 year out of 13 years whereas for PCA regression model and ANN model, it is 11 years.

![Graph showing observed and predicted rainfall anomalies](image)

**Fig. 4.3 a**
Figure 4.3(a-c) Comparison of model output ISMRI with observed ISMRI in training and test case for 4.3(a) Multiple regression model, 4.3(b) PC regression model and 4.3(c) ANN model.
Fig. 4.4 a

Fig. 4.4 b
Figure 4.4(a-c) The difference of seasonal mean rainfall between model output and observed data for 4.4(a) multiple regression model, 4.4(b) PC regression model and 4.4(c) ANN model.
Fig. 4.5 b

Fig. 4.5 c
Fig. 4.5 d

Fig. 4.5 e
Figure 4.5 (a-f) Comparison of model output ISMRI anomaly with observed ISMRI anomaly for (a) Multiple regression training, (b) Multiple regression test, (c) PC regression training, (f) PC regression test cases, (e) ANN training and (f) ANN test.

Table 4.2 summarises the prediction skills of these models both for the training and test cases. It can be seen (table 4.2) that for none of the regression models, the RMS error for test case is smaller than the standard deviation of the observed data (0.56) implying that none of these predicts better than the mean. The best model, according to this criterion is the multiple regression model (sixth row). The test case correlation coefficient is also better for the multiple regression model.
Table 4.2 Performance of various regression models

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Training</th>
<th></th>
<th>Test</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$r$</td>
<td>RMSE</td>
<td>Std</td>
<td>$r$</td>
</tr>
<tr>
<td>Niño3 (lag=4)</td>
<td>0.35</td>
<td>0.72</td>
<td>0.28</td>
<td>0.55</td>
</tr>
<tr>
<td>Niño3 (lag=5)</td>
<td>0.44</td>
<td>0.69</td>
<td>0.34</td>
<td>0.29</td>
</tr>
<tr>
<td>Niño3.4 (lag=4)</td>
<td>0.37</td>
<td>0.71</td>
<td>0.29</td>
<td>0.61</td>
</tr>
<tr>
<td>Niño3.4 (lag=5)</td>
<td>0.48</td>
<td>0.68</td>
<td>0.37</td>
<td>0.57</td>
</tr>
<tr>
<td>Niño4 (lag=5)</td>
<td>0.40</td>
<td>0.71</td>
<td>0.31</td>
<td>0.62</td>
</tr>
<tr>
<td>All the above indices</td>
<td>0.52</td>
<td>0.65</td>
<td>0.41</td>
<td>0.64</td>
</tr>
<tr>
<td>PC1 of all the above indices</td>
<td>0.47</td>
<td>0.68</td>
<td>0.36</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Standard deviation of the ISMRI for the period 1951-1990 (period used for training) is 0.79 and that for the period 1991-2003 (period used for test) is 0.56.

Table 4.3 shows the performance of the ANN model based on correlation with the observed precipitation and the RMS errors of the model for the training and test cases. The correlation coefficient of the ANN model output has been found to be 0.53 and 0.66 for the training and test cases, respectively. The RMS error of the ANN model for the training case is 0.65 and that for the test case is 0.40. It can be seen that the RMS error of ANN model for test case is significantly smaller than the standard deviation of the observed data showing that the prediction is better than the mean prediction. Also, the ANN model has smaller training and test case RMS errors than the best regression model (sixth model) as can be seen from table 4.2.
It can be seen that the test case standard deviation of the ANN prediction is 0.42 (table 4.3) as compared to the observed standard deviation. Standard deviation of Nino3.4 (lag = 5 seasons) and the multiple linear regression model for test case is 0.41 (table 4.2). Thus the variance explained by the ANN model is slightly higher than explained by the linear regression model with Nino3.4 (lag = 5 seasons) as predictor and the multiple linear regression model.

Thus, individual seasonal precipitation prediction accuracy of the ANN model and the multiple linear regression model are comparable as far as El-Nino and La-Nina years are concerned. However, the overall prediction skill of the ANN model is better than the multiple regression model.

It is to be noted that the predictions from the ANN model have been made using only the first PC of the predictors and that from the best regression model have been made using the multiple regression of all the predictors.

The fact that the first PC of the Niño indices correlated with ISMIRI anomaly is successfully used for the prediction of the latter is an important result in the sense that it highlights the fact that spatial variability of the SST values in the above Niño regions is as important as the actual SST values in the individual regions. This relationship,
between the spatial variability of the SST in the above Niño regions and the ISMRI anomaly, is complex and cannot be discovered by the linear regression model, which is established by the fact that performance of the multivariate regression model when all the individual indices are used as predictors is better than when the first PC of the individual models is used. ANN model is successful in capturing this non-linear feature of the variability as is clear from the results discussed above.

4.5 Conclusion

Correlation analysis of Niño 1+2, Niño 3, Niño 3.4, and Niño 4 indices with the Indian summer monsoon rainfall index (ISMRI) reveal that there is significant lag correlation between them. Five predictors have been selected based on this analysis. Statistical models of regression and artificial neural networks have been used to study the predictability of the ISMR using these indices as predictors. Analysis of the outputs of the regression model reveals that the multiple regression model, using all five indices as the predictors, is the best among its class. However, the prediction skills remain inferior to those of the mean prediction, as is observed by comparing the Root Mean Square error and standard deviation of the observed precipitation. A three-layer feed-forward artificial neural network model has been used for the prediction of ISMR using the first principal component of the five predictors obtained by the correlation analysis. It is observed that the ANN model gives significantly better results than the regression models and the prediction has found to be better than the mean prediction. It is concluded that the relationship between the Niño indices described above and the ISMR is essentially non-linear.

Weather patterns over specific areas of the globe as well as SST’s in temperate oceans are now routinely predicted with some confidence based on tropical sea surface temperature (SST) anomalies associated with El Niño. The lead lag
relationships, arrived at in the present study, can be employed in the study of climate change impact driven by global climate variability which is frequently communicated over large distances by atmospheric tele-connections due to forcing in Atmosphere-Ocean coupling owing to large oceanic system memory that are modified by the dominant local and regional processes.