

INTERCOMPARISON OF DATA MINING, NEURAL NETWORKS AND THEIR HYBRID COMBINATION FOR RAINFALL PREDICTION

**Balakrishnan Athiyaman¹, Raghavendra S. Mupparthy², Rajendra Sahu³,
Manoj Dash⁴, S Ramkumar⁵**

^{1,2} National Centre for Medium Range Weather Forecasting, Noida, (India)

^{3,4}ABV-Indian Institute of Information Technology and Management, Morena Link Road, Gwalior, (India)

⁵Regional Integrated Multi-Hazard Early Warning System (RIMES),
RIMES Project, IIT, Chennai(India)

ABSTRACT

Weather plays a major role in all spheres of life on this planet, in general, and particularly human activities, ever since man decided to tackle it by predicting. Weather forecasting is a prediction of few key variables like pressure, temperature, wind speeds and rainfall. Rainfall or precipitation is the most important variable or phenomenon or entity in weather forecasting that has a direct bearing on all activities. Rainfall forecasting for the Indian subcontinent, in particular, is a challenging task because of the unique geographical position of India with a huge land mass in north and oceans to the south, west and east of it (India is the largest peninsula in the world). The land heats up fast and cools down fast too, while the oceans cannot and have their inertia, causing differential heating in a non-linearly coupled system. Further, the elevation of the Himalayan mountain range in the north further contributes to the topographic boundary conditions, thus contributing to the overall unpredictability of rainfall pattern.

The present work is an attempt to develop and evaluate four common statistical techniques like: Data Mining (DM) , Neural Networks (NN) and Combination of Data Mining and Neural Network Hybrid system using observations from Automatic Weather Stations (AWS) collected from National Climatic Data Centre (NCDC) for the period 2000-2014. The three statistical models were developed using WEKA &Java and MATLAB, respectively.

Past works have commonly used five critical meteorological parameters like: Temperature, Dew Point, Mean Sea-level Pressure (MSLP), Wind speed, Humidity to correlate it with precipitation. In the present study, the data and the techniques were evaluated for four selected cities in India located in four different geographical regions of the country. The results were quite interesting and the rainfall prediction made through Neural Network for all four regions were reasonably accurate compared to the other models.

I. INTRODUCTION

India is primarily an agriculture based economy and an agrarian society for which weather plays an important role in determining the success or the failure of agriculture and also agro-related enterprises. According to Government of India's Economic Survey 2012, agriculture and its related sectors contribute to 17% to the national gross domestic product (GDP), while employing 51% of the total national workforce. The country is heavily dependent on the agricultural usage and particularly on rains for harvesting. Nearly 65% of the total population in rural area are dependent on agricultural related activities, and a significant percentage of rural population is indirectly dependent on the throughput from the cultivable land. An important point to note is that the amount of cultivable land too is a variable of rainfall, as only a small portion of cultivatable area is today under canal irrigation and most areas are rain-fed. Hence a reliable and accurate rainfall prediction would help farmers to plan their agriculture activities well in advance and takes precaution of heavy or scanty rainfall. An accurate rainfall prediction (from weekly to seasonal time scales) would also help many utility departments like: power, civil aviation, shipping, fisheries, space program etc. to plan their activities well in advance.

To help the numerical weather forecasting agencies and as part of wider National and World Meteorological Organization (WMO) goals, India has been continuously upgrading its automatic weather stations (AWS) both qualitatively and quantitatively. As of today, there are a total of 550 AWSs operating in the country of which 127 are exclusively Agro-met stations (Ranalkar, et al. 2010), along with 1350 automatic rain gauges (ARGs) stations.

Nearly 80% of the total annual rainfall in the country is due to the Indian Summer Monsoon (ISM) or the Southwest monsoon, which is a major part of the global circulation feature called the Asian Summer Monsoon (ASM). There is a large variability in the rainfall during the ISM due to the forcing from sea surface temperature (SST) of primarily Indian Ocean and internal dynamics of the ISM dictated and modulated by the Madden-Julian Oscillation (MJO) and the monsoon intra-seasonal oscillations (ISO) (Ajaya Mohan and Goswami 2003). Due to this complex nature of the ISM, India Meteorological Department (IMD) has been using statistical/empirical methods to provide long range forecasts (LRF) of the rainfall due to ISM since their first introduction by Blanford (1884). Since then many researchers have proposed various statistical techniques, of which major works are: Autoregressive Integrated Moving Average (ARIMA) [(Thapliyal 1981); and (Tektas 2010)]; multiple linear regression [Shukla and Mooley, 1987]; power regression model [Gowrikar et al., 1991; and Rajeevan et al., 2004]; and artificial neural networks [Goswami and Srividya, 1996; Goswami et al., 1999; and Sahai et al., 2000]; singular value decomposition (SVD) and principal component analysis (PCA) [Mohanty et al., 2013].

The present paper presents results from the application of two empirical techniques for forecasting rainfall: data mining (DM) and neural networks (NN) to the observation data obtained from AWS. A hybrid combination of DM and NN (called as Hybrid from now) was tested in an attempt to improve the accuracy. The following section briefly summarizes the relevant literature survey; while the third section describes the study area. The fourth section presents the adopted methodology while the fifth section presents the results. The final section discusses these results and presents a conclusion on the selected techniques.

II. LITERATURE SURVEY

Author	A Brief Description of the study	Used Technique/Methods
Sharma, (2015)	The author implemented the algorithm using MATLAB and evaluated three algorithms with the Mean Square Error as the metric. They observe the Back propagation algorithm to be superior to other two algorithms. They also observed that the performance of the Neural Networks to fall with increase of number of neurons.	Back Propagation algorithm Cascaded back propagation Layer current network
Taksande (2014)	The author has adopted a ANN and GA based hybrid model for rainfall forecasting. Using data of temperature, air pressure, rainfall, humidity and wind speed, they observed that the neural network perform the best in the group. They also developed Hidden Markov Mode 1 (HMM) based GA and compares the predictability with data mining ANN models. They observed that the result from HMM based GA outperformed those from the ANN models.	Artificial Neural Network (ANN), Classification and Regression Tree (CRT), Support Vector Machine (SVM), K-Nearest Neighbour, Genetic Algorithm (GA)
Indrabayu (2013)	The author proposed a hybrid approach, SVM-Fuzzy that integrate the Support Vector Machine with Fuzzy Logic Methods. With predictor variables such as temperature, wind speed, humidity and rainfall, observe the proposed SVM-Fuzzy approach to outperform the Neuro-Fuzzy approach. The accuracy of the SVM-Fuzzy approach are found to be superior to the Neuro-Fuzzy approach	(i) Neural Network,(NN), MATLAB tool (ii) SVM-Fuzzy method, NN-Fuzzy method
Hemachandra (2013)	The author suggested the Hybrid model of Neuro-Fuzzy for short-term load forecasting. The Mean Error from the Neuro fuzzy model is found to be superior compare to multiple linear regression method.	Regression Method Fuzzy Logic Approach Neural Network Approach Hybrid (Neuro-fuzzy approach)
Rao (2013)	Developed an ESVR model. They have used SVM technique for future weather prediction. The outcome of this study provided a forecast of current weather conditions using the EVSR model. For enhancing the accuracy they are sending the predicted values again in the self organized maps.	Linear regression Support Vector Machines.
Paltasingh (2012)	They finds the Temperature and Rainfall are two important factors that affect crop yields. Study also explain that using multiple regression analysis and direct use of meteorological	Multiple Regression Analysis Aridity Index

	factors to measure weather impact on crop yield. They advocate the incorporation of 'aridity index' variable in the regression model to simplify the econometric analysis and also found to improve results.	
--	--	--

III. STUDY AREA

The present study focuses on observation data from India. The selected study sites are from four distinct geographic locations, with each city experiencing rainfall in different monsoon phases and exhibiting different climatic zones. The selected cities are: Patiala, Kolkata, Mumbai and Chennai, located in Northern, Eastern, Western and Southern regions of the country, respectively.

The city of Patiala is located in Punjab at about 30 °N latitude and according Köppen-Geiger climate classification (KGCCS; [Peel et al., 2007]), it is classified as *BSh*, which corresponds to *Steppe* or *semi-arid climate*. It experiences very harsh summers with temperatures touching 40 °C and pleasant winters with an average temperature of 8 °C. The average annual rainfall is about 754 mm, with most of the rains occurring in July, August and September months, i.e., SW monsoon period.

Kolkata is located at about 22 °N latitude and according to KGCCS; it is classified as *Aw*, which corresponds to *Tropical Wet and Dry Climate* or *Tropical Savanna*. The temperature ranges from a record high of 43.7 °C in June to a record low of 6.7 °C in January. The average annual rainfall is about 1735 mm, with rainfall over 100 mm occurring between May and October [IMD].

Mumbai is located at latitude of approximately 19 °N and according to KGCCS, Mumbai's climate is classified as *Am*, which corresponds to *tropical wet climate* or *tropical monsoon and trade wind littoral climate*. The coastal and tropical nature of the city modulates the temperatures, hence the mean maximum summer temperature is about 32 °C with a mean minimum winter temperature is about 30 °C, while the record maximum and minimum temperatures are 42.2 °C in April and 7.4 °C in January, respectively [IMD]. Mumbai city experiences lot of rainfall with an annual average of 2258 mm, with more than 100 mm between May and October.

The city of Chennai is located at approximately 13 °N latitude and according to KGCCS, it is classified as *Aw*. Chennai is located on the "thermal equator", with a record maximum and minimum of 45 °C in May and 13.9 °C in January, respectively. The mean annual rainfall is about 1400 mm, with peak rain season during the retreating monsoon phase (NE monsoon) during November. This is in marked contrast with the other three selected stations.

IV. METHODOLOGY

4.1 Criteria for selection of the critical parameters:

From the recent works 13 parameters were found to influence rainfall amount, of which top five independent variables (parameters) were considered based on reported correlation. They are: Temperature, Dew Point, Mean Sea Level Pressure (MSLP), Wind Speed (WS) and Humidity.

Data collection and preparation: Fifteen years of daily data (2000-2014) have been collected from NCDC and the data were quality controlled for missing data. The data was collected for the above mentioned four cities of India.

4.2 Data Mining

Data mining is one of the important exploratory tools to mine every large volume of data. The weather prediction researchers are mostly using this tool to predict the rainfall. The accuracy of the data mining prediction is comparable with the other model predictions.

The quality controlled observation data required to be discretized to facilitate the application of data mining predictive and cluster algorithm. In this process the data shall be classified and converted from numeric to nominal. An open source data mining tool called the Waikato Environment for Knowledge Analysis (WEKA) was used to classify the data. The fifteen-year data was segregated into two parts training (13 years) and testing (2 year). The classification algorithm will classify and make cluster with respect to rainfall as a prognostic variable. WEKA has several *a priori* algorithms to generate associative rules based on the observed data. A priori rules (Predictive rules) have been commonly applied in Business Analytics (BA) to determine the consumer behaviour based on their recent purchases. With the application of these techniques, few of the topmost rules are as follows:

- Temperature[°C]='(20.06-30.04)' DewPoint[°C]='(20.06-30.04)' MSLP[hPa]='(996-1008)' WindSpeed[knots]='(4-8)' Humidity[%]='(55-70)' 229 ==> Rainfall[mm]='(0-50)' 102 acc:(0.4752)
- Temperature[°C]='(20.06-30.04)' DewPoint[°C]='(20.06-30.04)' MSLP[hPa]='(996-1008)' Humidity[%]='(55-70)' 286 ==> Rainfall[mm]='(0-50)' 127 acc:(0.46681)
- Temperature[°C]='(20.06-30.04)' MSLP[hPa]='(996-1008)' WindSpeed[knots]='(8-12)' Humidity[%]='(70-85)' 26 ==> Rainfall[mm]='(50-100)' 3 acc:(0.14494)
- Temperature[°C]='(20.06-30.04)' DewPoint[°C]='(20.06-30.04)' MSLP[hPa]='(1008-inf)' Humidity[%]='(55-70)' 269 ==> Rainfall[mm]='(0)' 256 acc:(0.9510)

The interpretation of these rules is as follows: for example in the first rule, when the temperature is between 20.06°C to 30.04°C and when the Dew Point is between 20.06°C to 30.04°C; and when the MSLP is between 996 hpa and 1008, and when Wind Speed between 4 knots and 8 knots and Humidity is between 55% and 70% then the predicted rainfall would be maximum of 50 mm with an accuracy of 47%.

An search engine was developed to extract the rule that suggest rainfall with high accuracy and explained with most parameters. The search engine algorithm basically takes the observation from the validation/test data to

predict the rainfall. The search algorithm sorts the obtained list of rules based on the number of parameters and the level of accuracy. For example, if two rules match, then the one with more parameter was considered as top, followed by the accuracy. The predicted rainfall was compared with the actual observation to derive RMSE values presented in Table 1.2

19-5-2010 -Three Parameter Match Rule No. :-191. Temperature[°C]='(20.06-30.04)' DewPoint[°C]='(20.06-30.04)' MSLP[hPa]='(996-1008)' 916 ==> Rainfall[mm]='(50-100)' 23 acc:(0.01427)

Three Para Are :-Temperature = 24.83 MSLP = 999.8 DewPoint = 23.17

23-4-2011 - Three Parameter Match Rule No :- 80. Temperature[°C]='(20.06-30.04)' DewPoint[°C]='(20.06-30.04)' MSLP[hPa]='(996-1008)' WindSpeed[knots]='(4-8)' Humidity[%]='(55-70)' 229 ==> Rainfall[mm]='(0-50)' 102 acc:(0.4752)

Three Para Are :-Temperature = 27.39 DewPoint = 24.67 WindSpeed = 5.0

17-10-2012 – Three Parameter Match Rule No. :- 80. Temperature [°C]='(20.06-30.04)' DewPoint [°C]='(20.06-30.04)' MSLP[hPa]='(996-1008)' WindSpeed[m/s]='(4-8)' Humidity[Degree]='(55-70)' 229 ==> Rainfall [mm]='(0-50)' 102 acc:(0.4752)

Three Para Are :-Temperature = 25.89 DewPoint = 23.33 WindSpeed = 4.6

15-02-2013 - Three Parameter Match Rule No 25. Temperature[°C]='(20.06-30.04)' DewPoint[°C]='(20.06-30.04)' WindSpeed[knots]='(0-4)' Humidity[%]='(55-70)' 123 ==> Rainfall[mm]='(0)' 107 acc:(0.85891)

Three Para Are :-Temperature = 26.33 DewPoint = 23.61 WindSpeed = 3.4

18-10-2014 - Three Parameter Match Rule No 208. Temperature[°C]='(20.06-30.04)' DewPoint[°C]='(20.06-30.04)' WindSpeed[knots]='(0-4)' 1255 ==> Rainfall[mm]='(100-150)' 12 acc:(0.00642)

Three Para Are:-Temperature = 25.56 DewPoint = 24.67 WindSpeed = 2.4

The Predictive algorithm predicted results are shown in the Table 1.2. Using the generated rules, the predicted rainfall was estimated and compared with actual rainfall to calculate the RMSE for the model for each station. Comparing these results with previously tested regression model, the Data mining results gives the better results compare to regression results

Data mining predictive algorithm predicts the rules with various combinations of the selcted (05) parameters and there by using the rule search engine, it predictsthe rainfall with the determined accuracy of the rule. It is observed that with a combination of three parameter of temperature,dew point and wind speed , the accuracy have been found to be 0.8589 and minimum accuracy level with a combination of three parameter Temperature, Dewpoint and Wind Speed increases by 0.00642.

4.3 Neural Network

Neural Network is nonlinear model that is easy to use and understand compared to statistical methods. ANN is non-parametric model while most of statistical methods are parametric model that need higher background of statistic. ANN with Back propagation (BP) learning algorithm is widely used in solving various classifications

and forecasting problems. Even though BP convergence is slow but it is guaranteed. However, ANN is black box learning approach, cannot interpret relationship between input and output and cannot deal with uncertainties. To overcome this several approaches have been combined with ANN such as feature selection and etc.

Based on the literature review, Neural Network models handle big volumes of data. Most of the authors used Neural Network methods for analysing the rainfall data. Considering the outcome of the literature review the researcher has evaluated the rainfall prediction data using the Neural Network method. The Neural Network learns the input-output relationship through the training process. The learning process in the Neural Network is an interactive procedure in which its connection weights are adapted through the presentation of a set of input-output training example pairs. The "Feed-forward back-prob" and "Traingdx" techniques will be used during the Neural Network designing and implementation. The present work made use of MSE performance function as a metric. The number of Network Neurons used in this evaluation is ten. The test results are De-Normalized and compared with the observed data and calculated the difference and MSE. The same procedure adopted all four geographical locations.

4.4 Data Mining and Neural Network Hybrid Combination

The prepared data set (station wise) and horizon-wisewere used in the Hybrid Data Mining and Neural Network Analysis. The data have been segregated into two parts: (i) Training set and (ii) Test sets. Neural Network Model is designed to train along with data mining rainfall output data and tested with a test data set. The Neural Network learns the input-output relationship through the training process. The learning process in the Neural Network is an interactive procedure in which its connection weights are adapted through the presentation of a set of input-output training example pairs. The Network technique used "Feed-forward of input data and back-propagation of errors" and training function used hyper sigmoid.

V. RESULTS AND ANALYSIS

5.1 Result of Data Mining Approach

Data mining predictive algorithm predicts the rules with various combinations of the selected (05) parameters and there by using the rule predict rainfall with the determined accuracy of the rule. It is observed that with a combination of four parameter of temperature,dew point, wind speed and humidity, the accuracy have been found to be 0.462 with a combination of three parameters: Dew point, MSLP and wind speed the accuracy of prediction increase to 0.994.

The Predictive algorithm predicted results are shown in the Table 1.3 Using the generated rules, the predicted rainfall was estimated and compared with actual rainfall to calculate the RMSE for the model for each station. Comparing these results with previously tested regression model, the Data mining results gives the better results compare to regression results.

Table No: 1.2 Data Mining Results

Stations	1day	3days	7days	14days	28days
Chennai	7.7	12.04	14	16.14	13.9
Kolkata	6.5	12.54	23.8	11.56	11.82
Mumbai	6.8	54.28	14.4	15.43	17.16
Patiala	7.14	9.71	10.9	11.17	12.04
All India	11.02	9.48	10.04	10.07	10.41

5.2 Result of Neural Network Approach

The Neural Network Model derived from the training data were applied to the test data to evaluate their performances based on RMSE. They were also tested for short, medium and long-range forecasts. The values are tabulated in Table 5.8, and it shows that the best (lowest) RMSE values were obtained for Patiala station for 3-days forecast, while the maximum error was noticed in the case of Kolkata for 28-day range forecast. Overall, Neural Network models show bad performance for Kolkata station, with an average RMSE value of 21.12 mm for all the temporal ranges; while they seem to perform better for Chennai station with an average RMSE value of 8.59mm, followed by Patiala station with RMSE value of 10.81mm.

Table No 1.3 Neural Network Prediction Results using test data

Station	1day	3days	7days	14days	28days
Chennai	9.8	10.78	8.14	7.46	6.78
Kolkata	8.56	9.9	20.7	27.87	38.59
Mumbai	11.1	7.6	14.19	21.56	36.71
Patiala	8.58	4.78	8.7	12.62	19.4
AllIndia	16.48	12.4	18.06	22.9	30.65

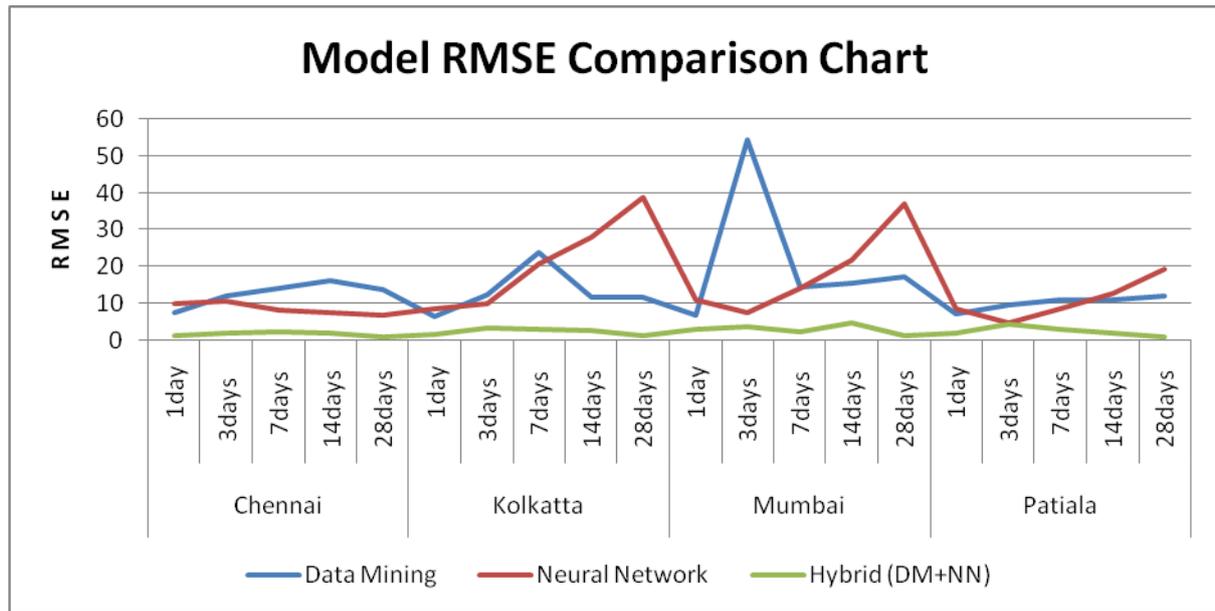
5.3 Result Analysis of Data Mining and Neural Network:

The values are tabulated in Table 1.4, and it shows the outcome of a hybrid combination of Data Mining and Neural Network System used to forecast rainfall for all stations. The best (lowest) RMSE values were predicted through model (1.23 mm) for Chennai station, while the maximum error was predicted through Hybrid Data Mining and Neural network model (4.92 mm) for Mumbai station. In Overall, Mumbai station predicted very high error forecast through this hybrid model with an average RMSE value of 3.12 mm for all the temporal ranges; All India station predicts with an average accuracy level of (2.36 mm). The Data Mining and Neural Network System model perform better with an average RMSE value of 1.26 mm for 28-day forecast, followed by 1-day forecast with RMSE value 2.08 mm compare to all other models.

The Artificial Neural Network (ANN) models were selected through a process of iterations for the number of nodes, and the Neural Network parameters to provide higher accuracies for the training data. The five station data sets with different horizons were normalised for ingesting into hybrid data mining and neural network analysis algorithm. Each training data set, (station-wise) yielded separate neural network architecture. The best fitted neural architecture was then used with the respective test data sets (two years of data for each station) for model accuracy.

Table No 1.4 Hybrid Neural Network and Data Mining system predicted results

Station	1 Day	3 Days	7 Days	14 Days	28 Days
Chennai	1.58	2.2	2.61	2.06	1.23
Kolkata	1.92	3.48	3.15	2.91	1.42
Mumbai	3.06	3.9	2.41	4.92	1.32
Patiala	2.23	4.57	3.21	2.26	1.14
All India	1.64	2.68	2.45	3.87	1.20



VI. CONCLUSION

The present study has examined three different models namely Data mining, NeuralNetwork and their Hybrid combination using fifteen years daily data for four different geographical locations. Six parameters were used in this study namely Temperature, Dew Point, MSLP, Wind Speed, Humidity and Rainfall. Out of three model prediction the Neural Network and Data Mining Hybrid model predict the rainfall with high accuracy.

REFERENCES

- [1] Agrawal, R., & Srikant R. (1994). Fast algorithms for mining association rules. Proceedings of the 20th International Conference on Very Large Data Bases, Santiago, Chile, 478-499.
- [2] Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. ACM SIGMOD Record, 22(2), 207-216. <http://dx.doi.org/10.1145/170036.170072>
- [3] Aggarwal, R., & Kumar, R. (2013). A Comprehensive Review of Numerical Weather Prediction Models. International Journal of Computer Applications, 74(18).
- [4] Asadi, S., Shahrabi, J., Abbaszadeh, P., & Tabanmehr, S. (2013). A new hybrid artificial neural networks for rainfall-runoff process modeling. Neurocomputing, 121, 470-480.
- [5] Ganti, R. (2014). Monthly monsoon rainfall forecasting using artificial neural networks.
- [6] Goyal, M. K. (2013). Monthly rainfall prediction using wavelet regression and neural network: an analysis of 1901-2002 data, Assam, India. Theoretical and Applied Climatology, 118(1-2), 25-34.
- [7] Guhathakurta, P. (2008). Long lead monsoon rainfall prediction for meteorological sub-divisions of India using deterministic artificial neural network model. Meteorol Atmos Phys, 101(1-2), 93-108.

- [8] Kumar, A., Pai, D., Singh, J., Singh, R., & Sikka, D. (2012). Statistical Models for Long-range Forecasting of Southwest Monsoon Rainfall over India Using Step Wise Regression and Neural Network. ACS, 02(03), 322-336.
- [9] Kumar, D., Pandey, A., Sharma, N., & Flügel, W. (2015). Modeling Suspended Sediment Using Artificial Neural Networks and TRMM-3B42 Version 7 Rainfall Dataset. Journal of Hydrologic Engineering, 20(6), C4014007. doi:10.1061/(asce)he.1943-5584.0001082
- [10] Luk., K.C. Ball., J.E. & Sharma., A. (2000). A Study of Optimal Model Lag And Spatial Inputs To Artificial Neural Network for Rainfall Forecasting. ELSEVIER Journal of Hydrology 227, 56-65.
- [11] Sarkar, B. K. (2013). Local area rainfall prediction using hybrid approach. International Journal of Innovative Computing and Applications, 5(4), 213.
- [12] Sharma, A., & Nijhawan, G., (2015). Rainfall Prediction Using Neural Network. International Journal of Computer Science Trends and Technologies, 3(3), 65- 69.
- [13] Taksande, A., Khandait, S.P., & Katkar, M. (2014). Rainfall Forecasting Using Artificial Neural Network: A Data Mining Aroach International. Journal Of Engineering Sciences & Research Technology, ISSN: 2277-9655 .2018-2020
- [14] Indrabayu., Harun N., Pallu M., & Achmad A., (2013). A New Approach of Expert System for Rainfall Prediction Based on Data Series. International Journal of Engineering Research and Applications (IJERA), 3(2), 1805-1809.
- [15] Hemachandra, S., & Sathyananayana, R.V.S. (2013). Electric Load Forecasting using Neuro-Fuzzy Systems. Global Research Analysis, 2(5), 2277 - 8160.
- [16] Rao, G.K. , Richard., M (2013). Artificial neural networks in temporal and spatial variability studies and prediction of rainfall. ISH Journal of Hydraulic Engineering,20(1), 1-6.
- [17] Paltasingh, K., R., A., Goyari, P., & Mishra., R., K., (2015). Measuring Weather Impact on Crop Yield Using Aridity Index: Evidence from Odisha. 25(2), 205- 216.