2015 IEEE International Conference on Research in Computational Intelligence and Communication Networks
(ICRCICN)

Bayesian Network Model for Monthly Rainfall Forecast

Ashutosh Sharma
Department of Civil Engineering
Indian Institute of Technology Guwahati
Assam, India
ashutoshsharmaiit@gmail.com

Manish Kumar Goyal
Department of Civil Engineering
Indian Institute of Technology Guwahati
Assam, India
mkgoyal@iitg.ernet.in

***Abstract*— This paper aims at rainfall forecasting which has been one of the most challenging problems around the world. Rainfall forecasting has importance in different areas including scientific research, agriculture etc. A Bayesian network model is proposed in this paper for forecasting monthly rainfall at 21 stations in Assam, India. Bayesian network or belief network (BN) is a probabilistic graphical model which shows conditional probabilities between different variables/nodes. Rainfall at a station is taken as a variable for this model and dependencies between rainfalls at different station is shown by BN. Rainfall dependencies between different stations is found using K2 algorithm which finds BN based on a greedy search algorithm. Five local and global atmospheric parameters which include Temperature, Relative Humidity, Wind Speed, Cloud Cover and Southern Oscillation Index (SOI) are used as evidences for this model. Conditional probabilities between stations and atmospheric parameters are calculated using Maximum Likelihood Parameter Estimation (MLE). Monthly data of 20 years from 1981 to 2000 for all the parameters is used for this study which was taken from different sources. Bayesian model runs on discretized data so for this study we have taken into account three discretized values for each variable based on their distribution. Thirteen different combinations of five atmospheric parameters are studied which gives a comparison of the efficacy of different parameters in rainfall prediction. Standard data ratio 70:30 is taken for training and testing of model. Efficiency of the model predictions is presented in the form of percentage of correct predictions for every case. Efficiency is found to be above 85 percent for most of the cases. This model can serve well for prediction of monthly rainfall. Similar model can be developed for daily data also.**

Keywords*— *Bayesian network, Bayesian classification, rainfall forecast.

I. INTRODUCTION

Rainfall is one of the most important components of the hydrologic cycle and affects surface water resources. Prediction of rainfall has been one of the most challenging problems as it depends on many local and global parameters. This study is based on Assam region where agriculture plays a very important role in the economy and is the principal occupation of the rural people who constitute nearly 90% of the total population. [1] Flooding occurs in many parts of the state every year causing huge losses in term of economy and human lives. The importance of rainfall prediction can be attributed by the fact that it is directly linked with the prediction, planning and mitigation of extreme events like droughts and floods.

We have used a Bayesian Network (BN) model for prediction of rainfall. BN is a probabilistic graphical model that represents a set of random variables in the form of nodes and their conditional dependencies by directed arrows. [2, 3] Bayesian networks offer many advantages for data modelling in case of missing data entries, avoiding over-fitting of data, combining prior knowledge and data.[4] Many techniques are available for finding the Bayesian network structure and the calculating the conditional probabilities. [3, 5] Bayesian networks have wide range of application in many fields of science and engineering including medical diagnostic. [6]. Reference [7] presented a Bayesian network integrating model for the various processes involved in eutrophication. Some studies are available for the use of Bayesian network for prediction of rainfall. Reference [8] used a Bayesian network model for weather prediction and combined this model with numerical atmospheric prediction for future forecasts. Reference [9] showed a Bayesian network based model for weather forecast and stochastic weather generation for a network of 100 stations. Reference [10] used seven different atmospheric parameters for Bayesian network based model based on two classifications and achieved accuracy above 90 percent.

Assam is situated in Brahmaputra basin in northeast of India. A study on the long term trends of rainfall, temperature and rainy days for various basins and sub basins in India shows a decreasing trends for rainfall and rainy days, where as an increasing trend is found for temperature for Brahmaputra basin.[11] Reference[12] compared wavelet regression and neural network for monthly rainfall predictions for same stations of Assam. Wavelet regression model out preformed ANN model and achieved efficiency index more the 67 percent. The purpose of this study is to investigate the performance of Bayesian network model for monthly rainfall prediction. This study will be useful for water resources planner, managers and agricultural scientists for better water management options for local events in many parts of the state.

II. Study Area and Data

A. Study Area

Assam is one of the seven North-Eastern states of India. The state is situated between latitudes from 24° 8′ N to 28° 2′ N and longitudes from 89° 42′ E to 96° E (Fig. 1). It is situated at the foothills of Eastern Himalayas and receives average annual rainfall of around 300mm from southwest monsoon. Topography and geographic location of Assam cause significant temporal and spatial variation in rainfall over the area. Heavy rainfall is observed at many stations during June to August period due to Southwest Monsoon. During monsoon months monthly rainfall at some stations goes up to 800 mm but during lean period it remains less than 50 mm. Annual mean rainfall at Kamrup station between 1981 and 2000 is shown in Figure 2 which shows monthly variability in rainfall. Temperature also shows similar trends. Maximum temperature in summers lies around 35–38 °C and minimum temperature in winters lies around 6–8 °C. Due to this high temporal variability prediction of rainfall becomes very complicated.

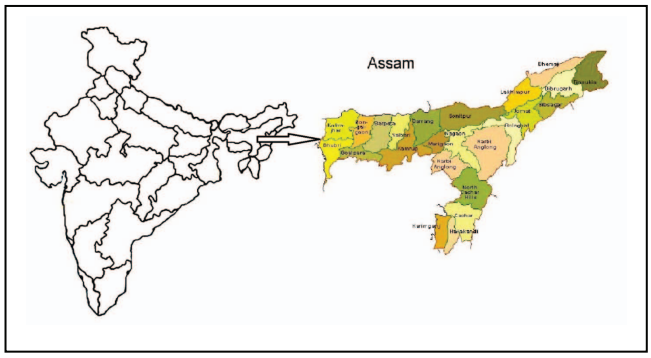

Figure 1: Study area

There is significant spatial variation of rainfall also. The mean monthly rainfall at different station ranges from 168 mm to 297 mm. The inter-nodal dependency property of BN is used for this spatial variation of the rainfall.

The Brahmaputra river flows through Assam from east to west over a length around 800 km. The topography and the warm and humid climate of the state make it home to 51 forest and sub-forest and the confluence of diverse patterns of vegetation. The state experiences heavy rainfall and floods every year which cause river bank erosions, landslides in many parts of the state. Agriculture plays a very important role in the economy of the State. Agriculture accounts for more than a third of the State Domestic Product. Agriculture which is the principal occupation of the rural people who constitute nearly 90% of the total population is also affected by the rainfalls and floods. [1]

B. Data

Monthly rainfall data of 21 stations spread over the whole state is used for this study. Table 1 shows latitude, longitude, maximum, minimum and mean rainfall for all stations. Data for precipitation was obtained from Indian Meteorological Department (IMD) website India water portal (http:// http://www.indiawaterportal.org/met_data) from 1981 to 2000.

TABLE I. Location and Rainfall Details of Stations

St No.	Station rainfall details					
	Station Name	Lat	Long	Min (mm)	Max (mm)	Mean (mm)
1	Jorhat	26° 45' N	94° 13 'E	0.19	620.7	175.4
2	Barpeta	26° 19'N	91° 00' E	0	920.2	240.2
3	Cachar	25° 05' N	92° 55' E	0	876.1	259.6
4	Darrang	26° 45' N	92° 30' E	0	955.7	241.2
5	Dhemaji	27° 29' N	94° 35' E	0.19	685.5	177.7
6	Dhubri	26° 02' N	89° 58'E	0	1020.5	195.6
7	Dibrugarh	27° 29' N	94° 54'E	0.2	703.8	179.5
8	Goalpara	26° 10' N	90° 37'E	0	1040.6	230.2
9	Golaghat	26° 31' N	93° 58' E	0.09	621.1	182.8
10	Hailakandi	24° 41' N	92° 34' E	0	851.3	244.3
11	Kamrup	26° 11' N	91° 44' E	0	1127.1	297.1
12	Karbi Anglong	26° 00' N	93° 30' E	0.03	960.2	264.2
13	Karimganj	24° 52' N	92° 21' E	0	790.8	222.6
14	Kokrajhar	26° 24' N	90° 16' E	0	997.8	217.0
15	Lakhimpur	27°140′ N	94°6′ E	0.2	604.1	177.2
16	Nagaon	26° 21' N	92° 41' E	0	1081.2	290.1
17	Nalbari	26° 25' N	91° 26' E	0	924.6	241.6
18	North Cachar Hills	25.18° N	93.03° E	0	962.2	284.3

19	Sibsagar	26° 59' N	94° 38' E	0.19	598.2	174.8
20	Sonitpur	26° 38' N	92° 48' E	0.02	718.9	195.6
21	Tinsukia	27° 30' N	95° 22' E	0.44	722.2	168.4

Various atmospheric parameters taken:

1. Temperature: We have taken monthly average temperature data for this study. Data was obtained from IMD website India water portal. The maximum and minimum values for monthly mean temperature are 30.275 oC and 12.038 oC.

2. Cloud cover: It is defined as the fraction of the sky obscured by clouds at a particular location and at a particular time. Its value lies between 0 and 1. Zero value represent clear sky and one value represent a scenario when sky is completely covered by clouds. Cloud cover data was obtained from India water portal. Cloud cover data shows very good correlation with rainfall data so we included it for this study. The maximum and minimum values for monthly mean cloud cover are 0.92737 and 0.14967 respectively.

3. Relative Humidity: It indicates the moisture level in air. It is defined as the ratio of water vapor density to the saturation water vapor density. Study on the relationship between relative humidity and rainfall for Uyo Metropolis, Akwa Ibom State, South- South Nigeria found a correlation between RH and rainfall. [13] The correlation between these two parameters is also found to be satisfactorily good for most of the station. We have used Climate Forecast System Reanalysis (CFSR) global meteorological dataset for this data. [14, 15] The maximum and minimum values of monthly mean relative humidity are 0.99493 and 0.231523 respectively.

4. Wind speed: Mean monthly wind speed data is obtained from CFSR global meteorological dataset. Wind speed shows positive correlation with some stations and negative correlation with others. The maximum and minimum values of mean monthly wind speed are 2.374403 m/s and 0.649341 m/s respectively.

5. Southern Oscillation Index (SOI): It is a standardized index to measure the large scale fluctuations in air pressure between the western and eastern tropical Pacific during El Niño (negative value) and La Niña (positive values) events. It is based on the observed sea level pressure differences between Tahiti and Darwin, Australia. Reference [16] presented an investigation on the relation between Indian Monsoon Rainfall (IMR) and Southern Oscillation Index (SOI). Data for SOI was obtained from National Oceanic and Atmospheric Administration (NOAA) website https://www.ncdc.noaa.gov/teleconnections/enso/indicators/soi/#soi-calculation.

We have used monthly data for 20 years from Jan, 1981 to Dec, 2000. The data is divided into standard ration 70:30 for training and testing respectively. Fourteen years data from Jan, 1981 to Dec, 1994 is used for training the model. Rest six years data from Jan, 1995 to Dec, 2000 is used for testing the performance of the model.

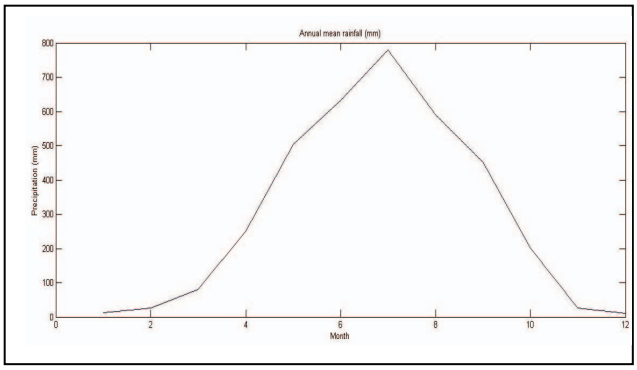

Figure 2: Mean monthly rainfall at Kmarup stations.

III. Bayesian Network

Bayesian Network is a probabilistic graphical model that represents a set of random variables as nodes and their conditional dependencies as directed arrow. A Bayesian network (B) is a combination of Bayesian network structure Bs (qualitative) and conditional probabilities Bp (quantitative). It is also known as Belief network or directed acyclic graphical model. [3] A random variable can be continuous or discrete. The first component, Bayesian network structure (Bs) or DAG consists of nodes connected by directed arrows/links. An arrow from a node X to node Y shows that node Y is dependent on node X and hence X is said to be the parent of Y or Y is said to be the child of X, as shown in Figure 3. Acyclic graph makes sure that no node can be its own parent or its own child. In this paper, a node represents rainfall and atmospheric parameters.

Conditional Probabilities (Bp) for each node is defined which quantitatively relates it to its immediate parents. For nodes having no parents prior-probability function is specified. For each node a conditional probability distribution (CPD) is defined that quantifies the effect of parent node on child node. . For discrete nodes, CPD is defined in the form of tables called Conditional Probability Table (CPT) which has probability for sets of each possible value of child node linked to values of its parents. The Joint Probability for any node can be expressed as the product of several conditional distributions as follows:

$$P_b(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P_b(x_i | x_{i+1}, \ldots, x_n)$$

In Bayesian network a node xi is independent of all other nodes except its parents (πi). Based on the conditional dependencies/independencies in a belief network Joint Probability Distribution is reduced to this form:

$$P_b(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P_b(x_i | \pi_i)$$

Interference is the computation of the probability distribution at a node for a set of query variables for a given set of

evidence variables. Figure 4 shows the overview of the model.

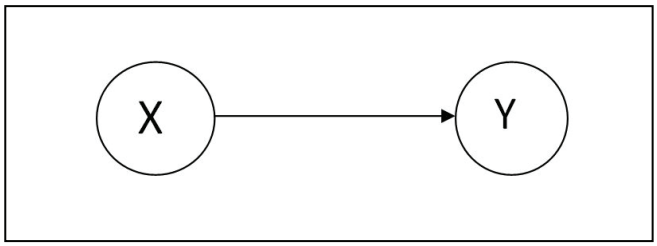

Figure 3: Two nodes X and Y with an arrow which shows Y is dependent on X.

A. Model

The random variable for this study is monthly rainfall at a station. Twenty one nodes are used to represent rainfall at 21 stations. As discussed earlier there is spatial variation of rainfall over the state, but in some cases rainfall at station is dependent/independent of some other station. Bayesian network is used to represent these dependencies/independencies. Rainfall is directly or indirectly dependent on many atmospheric phenomenon. We have used some of the atmospheric parameters which are known to cause/affect rainfall. We have used five atmospheric parameters viz. Temperature, cloud cover, wind speed, relative humidity and southern oscillation index (SOI). The data of these parameters for same period is used to find out dependencies between rainfall at a station and atmospheric parameters. The main idea behind the use of these parameters is to use them as evidence for this model. It is assumed here that these five atmospheric states can predict the rainfall. Using values of these five parameters for a month, rainfall is predicted.

All the nodes are taken as discrete for this model. As rainfall takes place over a large range so the whole data was discretized into three different states named 1, 2 and 3 for monthly data. 1, 2 and 3 represent the cases of Low, Average and High rainfall for that month respectively. Rainfall less than 50 mm is termed as Low, between 50mm and 800mm is termed as Average and above 800mm is termed as High. In similar way discretization was also done for atmospheric parameters in three classes based on the distribution.

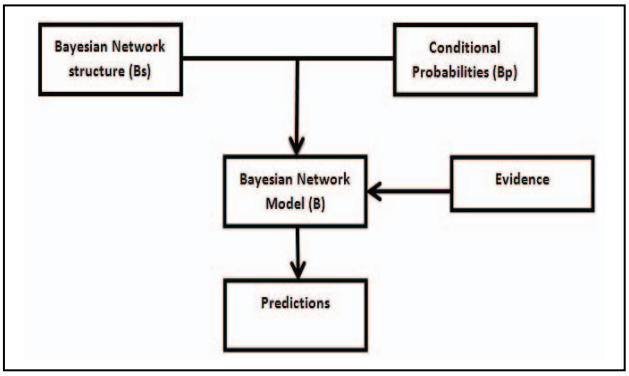

Figure 4: Overview of the model

B. Working

This model works in two stages: Training and Testing. Training is the process of adding a database to model to learn Bayesian structure (Bs) and conditional probabilities (Bp). Training includes structure learning using K2 algorithm and parameter learning using MLE. Data for rainfall and atmospheric parameters is given during this stage to train the model. Testing is the evaluation of the performance of the model. Data is supplied only for evidence nodes (atmospheric parameters) in this stage and model predicts values at other nodes based on Bayesian interference. The predictions obtained are compared with the observed values for evaluating the performance. These steps are explained separately.

1) Structure learning: Structure learning is the process of finding the Bayesian network structure (Bs) or DAG. For our work, we have used K2 algorithm for finding out the dependencies between various stations, which is greedy search algorithm for finding a Bayesian network structure Bs. It begins assuming all nodes have no parents and nodes are ordered. The order has to be decided manually. Network is highly influenced by the order, so it has to given carefully. For each node (Xi) in given order, it adds parent (πi) that has maximum increment in the factor which defines quality of network. The process stops when there is no imporvement by further addition of the nodes. [3] Figure 5 shows the Bayesian network structure obtained after using K2 algorithm. Stations are represented by number as listed in Table 1. Arrows show dependencies of one node on other as discused earlier.

2) Parameter learning: Parameter Learning is the process of quantifying the conditional probabilities (Bp) between the nodes. Maximum likelihood parameter estimation (MLE) is used for this model.

3) Interference: Top-down reasoning is performed for prediction of Rainfall in this model. Evidence is given to parent nodes (i.e. Atmospheric states) and rainfall at different stations is found. Inputs and outputs are in the form of discretized values. We have used Matlab BNT Toolbox for implementing all these processes in MATLAB. [17]

IV. Results and Discussion

The Bayesian network structure (Bs) is obtained by K2 algorithm for which only rainfall data at different stations used. Several order of nodes were given to obtain the best network. The best results were obtained using an order according to the serial no. of stations. Figure 5 shows the DAG obtained by K2 algorithm. K2 algorithm is very sensitive to order of nodes given. A wrong order may lead to a false network.

The second component of Bayesian network, Condition probabilities (Bp) were computed using Maximum Likelihood Parameter Estimation (MLE). The rainfall data was used to compute the inter-station conditional dependencies.

Conditional probabilities are in the form of Condition Probability Table (CPT) as we have used discretized nodes. To add atmospheric parameters, additional nodes were added to the network. The CPTs the added links between atmospheric parameters and stations were also obtained using MLE.

Model is tested for different combinations of atmospheric parameters. We operated model for 13 different cases to find the efficacy of atmospheric parameters used. The same Bayesian network structure (Bs), as shown in Figure 5, is used for all atmospheric parameter combinations as only rainfall data is used for making Bs. The final output of the model is in term of probability of rainfall to be in three categories i.e. Low, Medium and High. The category for which the probability is maximum is taken as the prediction. Accuracy of the model is presented in the form of percentage of correct predictions. Total no. of prediction made by model for testing period is 1512 (21 stations x 6 years x 12 months). Table II shows accuracy of model for different cases.
Temperature (Combination no 1) was found to be the most efficient among first four atmospheric parameters. Wind speed (Combination no 3) has shown the least accuracy which was expected as coefficient of correlation for rainfall and wind speed was found to be negative at some stations. This can also be seen by comparing combination no 11 and 12, as there is no improvement by adding wind speed. Efficiency of wind speed alone is not that good but when used along with temperature and cloud cover gave best results in combination no. 8.

TABLE II. ACCURACY OF MODEL FOR DIFFERENT COMBINATIONS OF ATMOSPHERIC PARAMETERS

Combination No	Atmospheric parameter taken	Accuracy (Percent)
1	Temperature	89.35
2	Cloud cover	86.44
3	Wind speed	61.64
4	Relative humidity	82.74
5	Temperature and cloud cover	89.35
6	Temperature and relative humidity	89.35
7	Cloud cover and relative humidity	86.77
8	Temperature, cloud cover and wind speed	91.27
9	Temperature, relative humidity and wind speed	89.48
10	Cloud cover, relative humidity and wind speed	88.16
11	Temperature, cloud cover and relative humidity	89.81
12	Temperature, cloud cover, relative humidity and wind speed	89.81
13	Temperature, cloud cover, relative humidity, wind speed and SOI	91.14

There is improvement in efficiency in combination no 13 over combination no 12 which signifies the importance of including SOI. Table III shows efficiencies of model with all the atmospheric parameters (combination no 13) for different stations. Thirteen stations have efficiency above 90 percent and four stations (station no 2, 6, 8 and 14) have efficiencies above 95 percent which shows model performed very well for these stations. Performance for other station is also satisfactorily good.

Similar patterns was obtained for wavelet regression based rainfall prediction model for these stations. [12] Stations located in the upper Assam much better than other stations. Stations like Kamrup, Karbi Anglong, North Cachar Hills, Cachar did not perform great.

TABLE III. ACCURACY OF MODEL WITH ATMOSPHERIC PARAMETERS COMBINATION NO. 13 FOR ALL STATIONS.

Station No	Station	Accuracy (Percent)
1	Jorhat	88.8889
2	Barpeta	95.8333
3	Cachar	87.5
4	Darrang	91.6667
5	Dhemaji	91.6667
6	Dhubri	95.8333
7	Dibrugarh	90.2778
8	Goalpara	95.8333
9	Golaghat	88.8889
10	Hailakandi	88.8889
11	Kamrup	86.1111
12	Karbi Anglong	87.5
13	Karimganj	90.2778
14	Kokrajhar	95.8333
15	Lakhimpur	94.444
16	Nagaon	90.2778
17	Nalbari	91.6667
18	North Cachar Hills	88.8889
19	Sibsagar	90.2778
20	Sonitpur	94.4444
21	Tinsukia	88.8889

V. Conclusion

This paper proposes Bayesian network model for mean monthly rainfall prediction at 21 stations in Assam, India. Directed acyclic graph (DAG) represents dependencies of rainfall at different stations which were found using K2 algorithm and conditional probability is found using maximum likelihood approximations. Five different atmospheric parameters viz. Temperature, Cloud cover, Relative humidity, Wind speed and SOI are used. Temperature is found most efficient and wind speed least. SOI is also found important in improving the results. Some station got efficiency above 95 percent whereas other station also got satisfactory results. Models can be developed based on this model by having more number of classes (discretization). This study can be useful for better management of water resources.

References

[1] Govt of Assam (2003) Assam development report. Available at http://hdr.undp.org/sites/default/files/india_2003_en.pdf

[2] Ben-Gal I. (2007) Bayesian Networks in Encyclopedia of Statistics in Quality & Reliability. Wiley & Sons.

[3] Cooper Gregory E, Herskovits Edwards (1992) A Bayesian Method for the Induction of Probabilistic Networks from Data. Machine Learning, 9, 309-347

[4] Heckerman David (1997) Bayesian Networks for Data Mining. Data Mining and Knowledge Discovery 1, 79–119 (1997)

[5] Cheng Jie, Bell David A. ,Liu Weiru (1997) An Algorithm for Bayesian Belief Network Construction from Data in the proceeding of AI & STAT'97

[6] Heckerman David (1990) Probabilistic similarity networks. ARTICLE in NETWORKS · AUGUST 1990

[7] Borsuk Mark E., Stow Craig A., Reckhow Kenneth H. (2004) A Bayesian network of eutrophication models for synthesis, prediction, and uncertainty analysis. Ecological Modelling 173 (2004) 219–239

[8] Cofino Antonio S., Cano Rafael, Sordo Carmen, Gutierrez Jose M. (2002) Bayesian Networks for Probabilistic Weather Prediction in ECAI 2002. Proceedings of the 15th European Conference on Artificial Intelligence, IOS Press, 695 - 700 (2002).

[9] Cano Rafael, Sordo Carmen, Gutierrez Jose M. (2004) Applications of Bayesian Networks in Meteorology. in Advances in Bayesian Networks, Gámez et al. eds., 309-327, Springer, 2004.

[10] Nikam Valmik B, Meshram B.B. (2013) Modeling Rainfall Prediction Using Data Mining Method in 2013 Fifth International Conference on Computational Intelligence, Modelling and Simulation.

[11] Jain Sharad K., Kumar Vijay (2012) Trend analysis of rainfall and temperature data for India. CURRENT SCIENCE, VOL. 102, NO. 1

[12] Goyal Manish Kumar (2014) Monthly rainfall prediction using wavelet regression and neural network: an analysis of 1901-2002 data, Assam, India. Theor Appl Climatol (2014) 118:25 –34

[13] Umoh Augustine Asuquo, Akpan Aniefiok O., Jacob Bernice Bassey (2013) Rainfall And Relative Humidity Occurrence Patterns In Uyo Metropolis, Akwa Ibom State, South- South Nigeria. IOSR Journal of Engineering (IOSRJEN) Vol. 3, Issue 8 (August. 2013), ||V4|| PP 27-31

[14] Fuka D.R., C.A. MacAllister, A.T. Degaetano, and Z. M. Easton (2013). Using the Climate Forecast System Reanalysis dataset to improve weather input data for watershed models. Hydrol. Proc. DOI: 10.1002/hyp.10073

[15] Dile, Y. T. R. Srinivasan, 2014. Evaluation of CFSR climate data for hydrological prediction in data-scarce watersheds: an application in the Blue Nile River Basin. Journal of the American Water Resources Association (JAWRA) 1-16. DOI: 10.1111/jawr.12182

[16] Mooley D. A. And Munote A. A. (1997) Relationships Between Indian Summer Monsoon and Pacific SST/SOI Tendency from Winter to Spring and their Stability. Theor. Appl. Climatol. 56, 187-197 (1997).

[17] Murphy Kevin P. (2001) The Bayes Net Toolbox for Matlab. Computing Science and Statistics, vol 33, 2001. Code available at https://code.google.com/p/bnt