Predictive Analytics for Rainfall Prediction

Lakshmi S R
Department of Computer Science and Engineering,
APS College of Engineering, Kanakpura, Bangalore

Asha N
Dept. of CSE and Engineering
APS College of Engineering, Kanakpura, Bangalore

K C Gouda,
* CSIR Fourth Paradigm Institute
NAL Belur Campus, Bangalore, India

***Abstract*: Rainfall is important for food production plan, water resource management. India is an agricultural country and its economy is largely based upon productivity. Thus rainfall prediction becomes a significant factor in agricultural countries like India. On the growing importance of Rainfall studies in the climate change scenario and High Performance Computing, different Users starting from a farmer to a scientist to a policy maker needs the rainfall prediction well in advance for their application like crop planning, water storage etc. Data discovery from temporal, spatial and spatio-temporal data is critical for rainfall analysis. However, recent growth in observations and model outputs, combined with the increased availability of geographical data, presents new opportunities for the users to implement new techniques such as predictive analytics for developing a predictor which can be used for multi-scale forecasting of rainfall that is from 24 hour forecast to long-range forecast say 2-3 month in advance forecast. Hence we developed predictive analytics system for the efficient and real time prediction of rainfall over India.**

Keywords—Agriculture, Ensemble forecasting, Rainfall Forecasting, Prediction.

I. INTRODUCTION

Agriculture is the predominant occupation in India, accounting for about 52% of employment. Since, India is an agricultural country and its economy is largely based upon crop productivity. The occurrence of prolonged dry period or heavy rain at the critical stages of the crop growth and development may lead to significant reduce crop yield. Rainfall is important for food production plan, water resource management and all activity plans in the nature. Thus rainfall prediction becomes a significant factor in agricultural countries like India. Rainfall forecasting has been one of the most scientifically and technologically challenging problems around the world in the last century. The rainfall data is available for the data mining techniques which can be useful for predicting the rainfall which can very useful for taking decisions over crop planting in the areas.

The Irrigation facilities are inadequate, as revealed by the fact that only 52.6% of the land was irrigated in 2009–10 which result in farmers still being dependent on rainfall, specifically the Monsoon season. A good monsoon results in a robust growth for the economy as a whole, while a poor monsoon leads to a sluggish growth. In the present study a framework for rainfall prediction from past data and present weather condition is generated using the predictive analytics in a High Performance Computing environment where several software's like MATLAB, STATISTICA etc are interfaced with the object oriented language to develop a multi-scale forecasting platform for rainfall prediction.

II. RELATED WORK

There have been many attempts to forecast rainfall. Rainfall forecasting can apply to many time horizons such as short term, medium term, and long term periods. Some authors design systems which can forecast yearly data, some try to forecast monthly data whereas some try to forecast daily data.

N. Sen. [1] has presented long-range summer monsoon rainfall forecast model based on power regression technique with the use of Ei Nino, Eurasian snow cover, north west Europe temperature, Europe pressure gradient Wind pattern, Arabian sea SST, east Asia pressure and south Indian ocean temperature in previous year. The experimental results showed that the model error was 4%.

S. Nkrintra, [2] described the development of a statistical forecasting method for SMR over Thailand using multiple linear regression and
local polynomial-based nonparametric approaches. SST, sea level pressure (SLP),wind speed, EiNino Southern Oscillation Index (ENSO), IOD was chosen as predictors. The experiments indicated that the correlation between observed and forecast rainfall.

T. Sohn, [3] has developed a prediction model for the occurrence of heavy rain in South Korea using multiple linear and logistics regression, decision tree and artificial neural network.

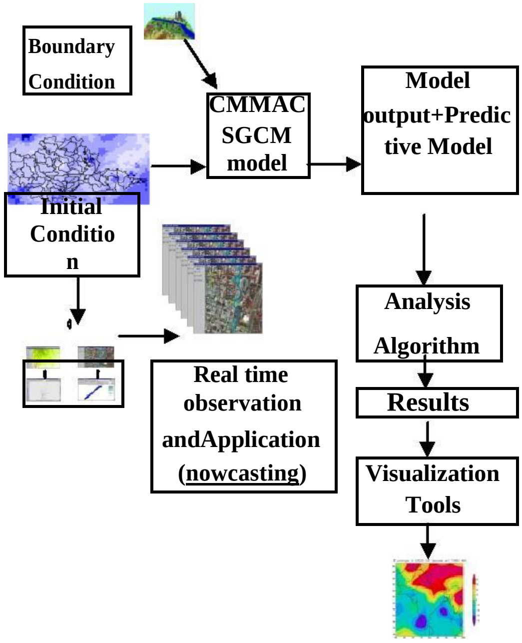
M. T. Mebrhatu [4] modeled for prediction categories of rainfall (below, above, normal) in the highlands of Eritrea. The most influential predictor of rainfall amount was the southern Indian Ocean SST. Experimental results showed that the hit rate for the model was 70%.

H. Hasani [5] proposed human height prediction model based on multiple polynomial regression that was used successfully to forecast the growth potentials of height with precision and was helpful in children growth study.

Vaccari [6] modeled plant motion time series and Nutrient recovery data for advanced life support using multi variable polynomial regression.

But in these studies the approach is very simple as the data set is very small, To handle data mining in a Big data environment as all the climate data are of large size with multi-dimensional (latitude ,longitude, vertical levels, time(day,month,year)) and of large time scale starting from daily to decades i.e 10 years. So in the present project work algorithms will be developed for large climate data analysis to study the climate change over India.

III.SYSTEM DESIGN



The fig 1 shows the high level design of forecasting a rainfall. In this we consider both boundary conditions (eg .It may be land, ocean, forest parameter) as well as initial conditions(ie.,today's condition) data along with the GCM(Global Climate Model) output data. With these output data we apply our predictive analysis algorithm to analyze the data. Finally, the forecasted results will be obtained by the visualization package (ie., Grads).

Fig1. High level design

A. Flow of the Predictive Model

Step 1: Adopting unsupervised data classification technique, such as K-means clustering technique, for clustering of the observed multi-site rainfall data in order to identify the rainfall states present in the rainfall data.

Step 2: Perform Principal Component Analysis (PCA) to reduce the dimensions of the standardized predictor data, i.e. CMMACS climate data set. The dimensionally-reduced climate variables represent a large fraction of the variability contained in the original data.

Step 3: Training the Predictive model(s) to establish relationship between the input data containing current day standardized and dimensionally-reduced climate predictors along with previous day(s) rainfall state and the output data containing the current day rainfall state.

Step 4: Applying bias correction for the GCM output data to obtain bias-corrected GCM data.

Step 5: Obtain principal components of GCM data by performing PCA of the bias-corrected GCM data with the help of principal directions obtained during PCA of CMMACS climate data.

Step 6: Using the trained Predictive model to derive present day rainfall state with the help of Principal components obtained from GCM output and rainfall state of the previous day.

B. Design of Multi-Scale Predictive Analytics

- ***Short Range Prediction(1-6hrs)***

Now casting using real-time satellite data.

eg. IPL match, event organization, closing school etc.

- ***Medium Range Prediction(3-5days)***

Cyclone or heavy rainfall using meso-scale model (WRF) +PA algorithm

Eg. Disaster preparedness, Warning for Fisherman, Closing school.

- ***Long Range Prediction(3-6months)***

Monsoon rain forecasting in April for JJAS using GCM Model + PA algorithm

Eg. Agriculture planning, Water management, food security etc.

- ***Climate Prediction(10-100yrs)***

Understanding Climate Change, Global warming using CSIR climate model +PA algorithm

Eg. Food habit, crop pattern etc

IV. IMPLEMENTATION AND RESULTS

The modules are

1. Monsoon model (long range)
2. weather research forecast model(Short-range)
3. Cyclone model (rainfall day cyclone)

In order to calculate the anomaly of rainfall we use the following formula.

$$Rain(lat,lon) = \frac{1}{50}\sum_{yr=1}^{50} Rain(lat,lon, yr)$$

Anomaly

For year if anomaly (-10 to10): Normal Year (ex 1998)

If anomaly less than (-10): Drought or Deficit Year (ex 2002)

If anomaly more than (+10): Flood or Excess Year (ex 1961).

Regression

Regression is a statistical empirical technique that utilizes the relation between two or more quantitative variables on observational database so that an outcome variable can be predicted from the others.

Regression use two methods Simple linear regression and multiple linear regression models. Regression produces a polynomial describing the relationship between any set of inputs and corresponding output.

Here we have considered past data to develop a regression equation,

- A linear regression equation to predict monsoon rainfall

$Y=aX_1+bX_2$

Where

y=Normalized rain anomaly.

x_1= normalized anomaly of the location. x_2=normalized anomaly of January to April.

The regression equation is utilized in order to predict the accurate results. The overall measure is adopted for the accuracy of the forecasts is the root-mean-square error (rmse).

$$RMSE = \sqrt{\sum_{i=1}^{n} (F_i - O_i)^2 / n}$$

Where f_i and o_i are the forecast and observed Indian monsoon rainfall for the ith year. In the present work an predictive analytics approach is carried out to provide better understanding of the rainfall pattern over all India using 50 years of multi-source data. This work involves development of several novel algorithms like ensemble forecasting, multi-model forecasting of the monsoon rainfall, extreme rainfall events, flood or drought Index etc. over India using predictive analytics. The developed predictive algorithm is compared against the traditional statistical methods of forecasting and proposes new directions. Finally several case studies will be presented using predictive analytics for rainfall predictions, which provide new scientific insights with high societal impacts.

Fig.2 shows the results of 20yr of verification of observed and forecasted rainfall. The arrow along the abscissa denotes the long-term mean rainfall. The two dotted lines represent the actual rainfall.

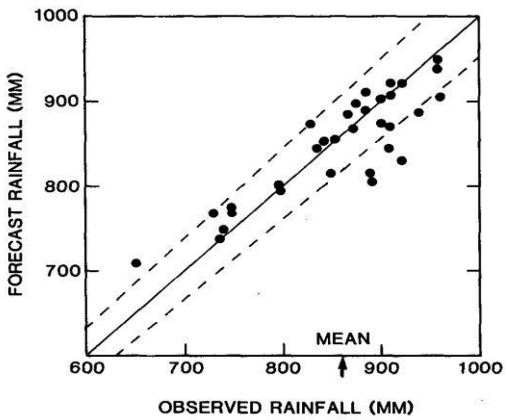


Fig.2 Observed and forecasted rainfall for 20yr of verification

Table 1 shows the comparison of forecasted and observed rainfall over India from the period 1951-1984.where the forecasted results are almost near to the observed data. Table 2 shows the result of forecasted rainfall for monsoon season (June, July, august & September) and this shows that predictive analytics prediction is almost near to observed rainfall data. Fig 3 shows the comparison of predicted model and observed monsoon rainfall.

Year	Rainfall Forecast	Rainfall Observed
1951	748	737
1952	797	792
1953	925	920
1954	891	885
1969	872	829
1970	891	939
1971	911	886
1972	717	653
1973	921	912
1974	767	747
1975	948	960
1976	853	855
1977	805	880
1978	845	908
1979	769	746
1980	825	881
1981	853	842
1982	738	736
1983	904	959
1984	845	835

Table1.Comparison of PA forecasted Rainfall and Observed Rainfall for the period 1951-1984

Year	Raw Prediction	With PA Prediction	Observation
2012	+3	-2	-7
2011	+6	+4	+2
2010	-1	+3	+2
2009	-5	-8	-22

Table2. Rainfall Forecasting for all India

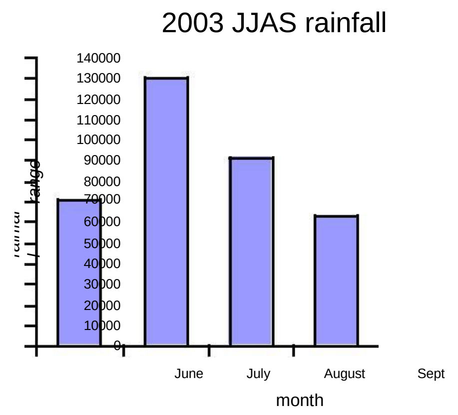


Fig 3. Seasonal Rainfall variation in monsoon months

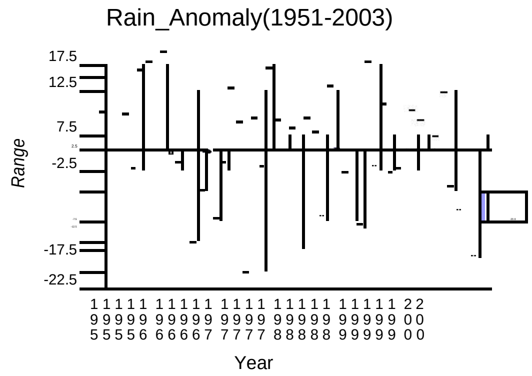


Fig 4. Rainfall anomaly calculated for 53yrs of data

Fig 3. Shows the seasonal rainfall variation during the months of June-Septembers. The results are averaged over 53 years and fig 4 shows the rainfall anomaly calculated for 53 yrs of data.

V. CONCLUSION

The skill of prediction contributes towards developing methodologies for predicting rainfall at local or regional scale over India from large scale GCM output of climatological data. At the end of present work first time a predictive analytics package is on place using which one can generate accurate and efficient spatio-temporal rainfall forecasting at an affordable cost in a cloud computing and Big Data environment.

REFERENCES

1. Klaush Juliseh, "Data mining for Intrusion Detection – A critical review", Applications of Data mining in computer security, Daniel Barbara, Sushil jajodia, Published by Springer.
2. Guhathakurta,P (2005) "Long-range monsoon rainfall prediction of 2005 for the districts and sub-division Kerala with artificial neural network", Current Science, 90, 773-779
3. Rajeevan, M (2001) "Prediction of Indian summer monsoon: Status, problems and prospects", Current Science.
4. D. Mark, "Geographical information science; critical issues in an emerging cross – disciplinary Research domain", URISA Journal vol. 12, February 1999.
5. Tamil Nadu Meteorology Department, Chennai.
6. European Commission, "Forest Fires in Europe 2007, Technical report, Report No. 8, 2008.