Prediction of Spatial Dependency of Multisite Rainfall

Motivation

Rainfall becomes a significant factor in agricultural countries like India. Rainfall prediction has become one of the most scientifically and technologically challenging problems in the world. This is due mainly to two factors: first, it's used for many human activities and secondly, due to the opportunities created by the various technological advances that are directly related to this concrete research field, like the evolution of computation and the improvement in measurement systems . Since ancient times, weather prediction has been one of the most interesting and fascinating domain. Scientists have tried to forecast meteorological characteristics using a number of methods, some of these methods being more accurate than others.

Accurate rainfall prediction will also help in the following-

1. **Agriculture** : Rainfall means crops and crops mean life. Thus rainfall prediction is closely related to agricultural sector, which contributes significantly to economy of our nation.
2. **Disaster Management** : Accurate rainfall prediction will also help in disaster management like floods, droughts etc.
3. **Weather Forecast** : As rainfall prediction relies on several atmospheric components, thus rainfall prediction will also help in accurate weather forecasting, which will help in planning everyday activities and various other long term factors.

Use cases

1. **Agriculture sector** : Accurate prediction of rainfall will help in deciding the optimal sowing and cropping patterns which will decrease the electricity usage for running motor-pumps, seed wastage, labour cost, water-usage .
2. **Consumer :** The reduced cost in the agricultural sector will also reduce the crop cost and this will lead to decrease in prices of various commodities and goods will be cheaper .
3. **Government**: It can help the government to take preventive measures against natural disasters like flood and drought.The crop or the stored food gets damaged due to excessive rain sometimes, so it will also help the government to take preventive measures.
4. **Weather-forecasters**-: This model will also provide some useful insights to meteorologists like spatial dependencies and correlation between rainfall patterns of different sites.

Related Work

A wide variety of rainfall forecast methods are available. There are mainly two approaches to predict rainfall. They are Empirical method and dynamical method. The empirical approach is based on analysis of historical data of the rainfall and its relationship to a variety of atmospheric and oceanic variables over different parts of the world. Some of the methods previously used for Rainfall prediction

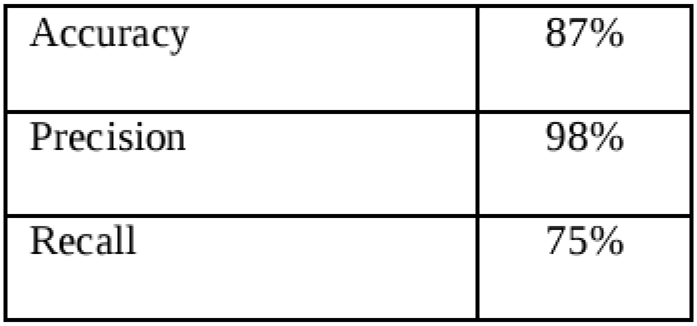
1. **Joseph, Jyothis, and Ratheesh T. K. "Rainfall Prediction Using Data Mining Techniques."** ***International Journal of Computer Applications*** **83.8 (2013)**

Idea : This paper adopts an empirical approach for rainfall prediction and uses data mining techniques like clustering and classification for rainfall prediction.

Brief : The objective here is to analyze the four month rainfall data i.e from June to September for a particular region (here Kerala) over 9 years. These particular months were chosen as they belong to monsoon during which major rainfall occur. Artificial Neural Nets were used to implement the required empirical approach. Because along with analyzing the data it also learns for future prediction. Input parameters used for ANN are Temperature, Pressure, Relative humidity, Wind speed, Precipitable water. In this technique, rainfall values are clustered using subtractive clustering and three classes are identified - low, medium and heavy. Data used was downloaded from official website of (National Oceanic and Atmospheric Administration) NOAA.

Result : Accuracy is the overall correctness of the model and is calculated as the sum of correct classifications divided by the total number of classifications. Precision is a measure of the accuracy provided that a specific class has been predicted. It is defined as Precision = TP/(TP + FP) where TP and FP are the numbers of true positive and false positive predictions for the considered class. Recall is a measure of the ability of a prediction model to select instances of a certain class from a data set. It is also called sensitivity, and corresponds to the true positive rate. It is defined by the formula Recall = Sensitivity = TP / (TP + FN) where TP and FN are the numbers of true positive and false negative predictions for the considered class.

Table 1. Performance Measures



Accuracy	87%
Precision	98%
Recall	75%

2. Luk, Kin C., J.e. Ball, and A. Sharma. "An Application of Artificial Neural Networks for Rainfall Forecasting." *Mathematical and Computer Modelling* 33.6-7 (2001) - Uses artificial neural network and pattern recognition techniques.

Idea: Presented in this paper is a comparison of three types of ANNs (i.e. MLFN (Multilayer Feed forward Network), PRNN (partial recurrent neural network) and TDNN (time delay neural network)) for forecasting rainfall over an urban catchment in western Sydney, Australia.

Brief : The scope of this study was to forecast rainfall one time-step ahead. For the development of the proposed rainfall forecasting models, the rainfall process was assumed to be a Markovian process, which means that the rainfall value at a given location in space and time is a function of a finite set of previous realisations. With this assumption, a model structure can be expressed as where :

$$X(t + 1) = g\ (X(t),X(t - l),X(t - 2, \ldots ,X(t - k + 1)) + e(t),$$

Result : During the development of the alternative ANNs, various network configurations were attempted in order to determine the effect of two key variables, which are: the lag of network, and the number of hidden nodes.For the MLFN, networks with lags 1, 2, 3, and 4 were attempted. In addition, the numbers of hidden nodes tried were 2, 4, 8, 16, 24, 32, 64, and 128. Networks with two layers of hidden nodes were also attempted. The criterion function is Normalised Mean Square error which is compared for various networks.For each type of network, there existed an optimal complexity, which was a function of the number of hidden nodes and the lag of the network. All three networks had comparable performance when they were developed and trained to reach their optimal complexities. Networks with lower lag tended to outperform the ones with higher lag. This indicates that the 15min. rainfall time series have very short term memory characteristics.

3. **Ingsrisawang, Lily, et al. "Machine learning techniques for short-term rain forecasting system in the northeastern part of Thailand."** ***Machine Learning*** **887 (2008): 5358**. - Uses machine learning techniques to do short term rainfall prediction.

Idea: This paper presents the methodology from machine learning approaches for short-term rain forecasting system. Decision Tree, Artificial Neural Network (ANN), and Support Vector Machine (SVM) were applied to develop classification and prediction models for rainfall forecasts.

Brief: The northeastern part of Thailand is an arid region with varied rainfall. To enhance the precipitation in this area, a number of cloud seeding operations have been conducted by the Royal Rain Making Project. Two integrated datasets, so-called GPCM and GPCM+RADAR, provided by Bureau of the Royal Rain Making and Agricultural Aviation and Department of Meteorology, Thailand, were explored in this study. The GPCM dataset consists of 309 daily records including the upper air observations, seeding operations and the average of rain volumes (AVG) from 18 rain gauges at regional weather Stations. A decision tree induction algorithm (C4.5) was used to classify the rain status into either rain or no-rain.

Result: Using the GPCM and GPCM+RADAR datasets, the C4.5 decision-tree induction model can achieve accuracy of 87.06% and 94.41% respectively in forecasting whether rain or no-rain event, but provides somewhat lower accuracy at 62% level in forecasting whether no-rain, few-rain, or moderate-rain event will occur (Table I).

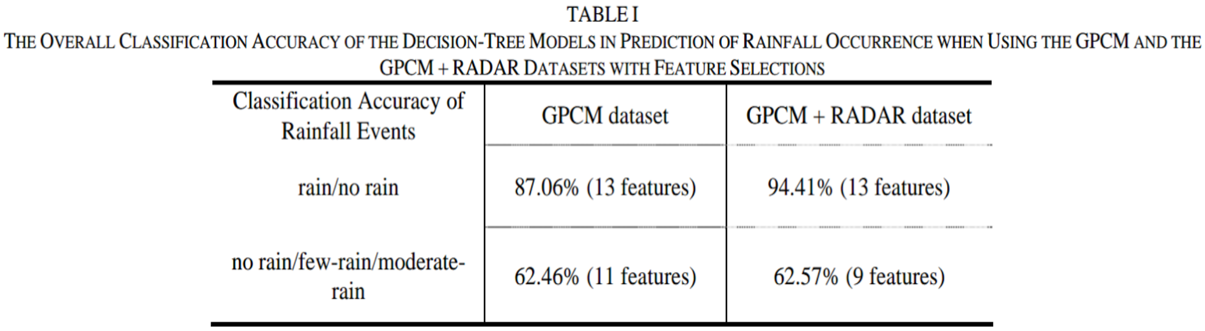


TABLE I
THE OVERALL CLASSIFICATION ACCURACY OF THE DECISION-TREE MODELS IN PREDICTION OF RAINFALL OCCURRENCE WHEN USING THE GPCM AND THE GPCM + RADAR DATASETS WITH FEATURE SELECTIONS

Classification Accuracy of Rainfall Events	GPCM dataset	GPCM + RADAR dataset
rain/no rain	87.06% (13 features)	94.41% (13 features)
no rain/few-rain/moderate-rain	62.46% (11 features)	62.57% (9 features)

4. **Wilby, Rl, Ol Tomlinson, and Cw Dawson. "Multi-site Simulation of Precipitation by Conditional Resampling."** ***Climate Research*** **23 (2003)**

Idea: A single-site, regression-based downscaling method is extended to multi-site synthesis of daily precipitation at stations in Eastern England and the Scottish Borders. Area-averaged precipitation series for each region are downscaled using gridded predictor variables selected from a candidate suite representing atmospheric circulation, thickness and moisture content at length scales of 300 km.

Brief: This paper deals with multi-site generation of daily precipitation at stations in Eastern England (EE) and the Scottish Borders (SB) using an extension to a hybrid regression/weather-generator model (Wilby & Dettinger 2000, Wilby et al. 2002b). The inter-site distances range from 13 to 291 km, and the site elevations from 2 to 253 m above mean sea level. The weather generator, Statistical DownScaling Model (SDSM), was initially developed for downscaling future climate change scenarios at single sites given large-scale climate variables supplied by general circulation models (GCMs). Full technical details and split-sample tests of SDSM are provided by Wilby et al. (1999,2002b), and Wilby & Dettinger (2000).

5. **Buishand, T. Adri, and Theo Brandsma. "Multisite Simulation of Daily Precipitation and Temperature in the Rhine Basin by Nearest-neighbor Resampling." *Water Resources Research* 37.11 (2001)**

Idea : The method of nearest-neighbor resampling is extended to simultaneous simulation of daily precipitation and temperature at multiple locations over a large area (25 stations in the German part of the Rhine basin). Nearest neighbors refer here to historical days for which the observed weather is closes to that of the simulated weather for a given day

Brief : Since the observed weather of historical days is resampled, the dependence between daily precipitation at different sites and that between daily precipitation and temperature is automatically preserved. Many of these dependencies have a complicated structure, which may not be adequately described by parametric models. Prior assumptions about the distribution of weather variables need also not be made.

6. **Singh, Sadhana, S. Kannan, and P. V. Timbadiya. "Statistical downscaling of multisite daily precipitation for Tapi River basin using kernel regression model." *CURRENT SCIENCE* 110 (2016)**

Idea: The study presents fine resolution multisite daily precipitation projection for the Tapi basin using the kernel-regression (KR) based statistical downscaling methodology developed by Kannan and Ghosh with and without conditioned on the estimated rainfall
State.

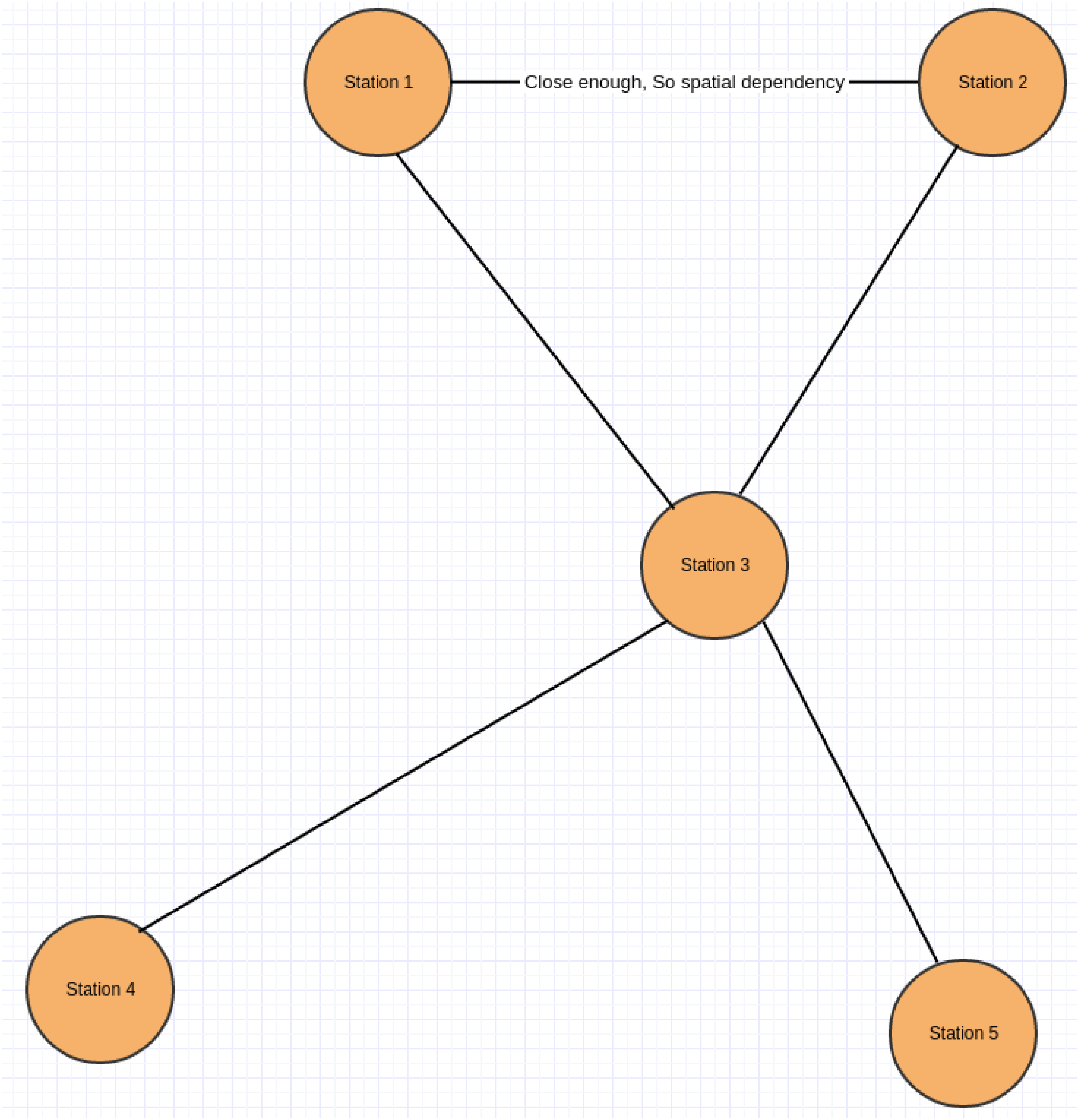
Brief: The study is an attempt to downscale precipitation at a very fine resolution of 0.25 using CMIP-5 GCM data of scenario RCP4.5 (considering this as the most probable scenario) 22,23 and RCP8.5 (considering this as the worst case scenario)23,24 to quantify impact of climate change on water resources of the Tapi Basin up to the end of the 21st century. Furthermore, the downscaled daily precipitation from CMIP-5 GCM (MPI-M) and daily precipitation from Coordinated Regional Climate Downscaling Experiment (CORDEX) South Asia Regional Climate Model (RCM) were compared to evaluate the uncertainty in the KR-based statistical downscaling model for the study area under consideration. The study supports strength of the KR-model in capturing the highly heterogeneous rainfall as well as the interstation cross correlations in the basin.

OUR APPROACH

Given rainfall pattern data for a region over a long time period (around 50 years), aim of this project is to do multisite rainfall prediction. I.e. given a site in the specified region predict the future rainfall pattern.
Although statistical downscaling models already exist, but they fail to capture the correlation and spatial interdependencies between multiple sites, and therefore inadequate to model the variability of the rainfall patterns. Therefore this project aims to create a machine learning model which will model these dependencies thus providing more accurate prediction.

For now we have come up with two major models:

1. **Probabilistic Graphical Model** : Many machine learning techniques such as Markov chains, autoregressive models, and neural networks have been used with limited success. In particular, these models fail to represent spatial and temporal dependencies between neighboring locales. In this study, we examine the use of Bayesian networks to better capture regional dependencies in the limited context of precipitation prediction. We are particularly interested in determining a minimal set of measurement sites sufficient to quantitatively predict local rainfall. Central to these goals, we exploit the interdependence between geographically disparate measurements to evaluate the utility of each existing measurement site and potential new sites.



Each node in the graph represents a weather-site. We try to find out the interdependency of these sites and use them to predict rainfall conditions in any new neighbouring site.

Bayesian Net Construction : As each node in a Bayes network may conditioned upon any other node in the network, exhaustively learning an optimal network structure for all but the smallest networks is computationally intractable. Indeed, this problem is NP-hard. As such, a number of heuristics are commonly used to approximate a globally optimal DAG structure. These include the Metropolis-Hastings Markov Chain Monte Carlo (MCMC) method to sample the DAG space, hill climbing methods to explore node neighbors incrementally, active structure learning.

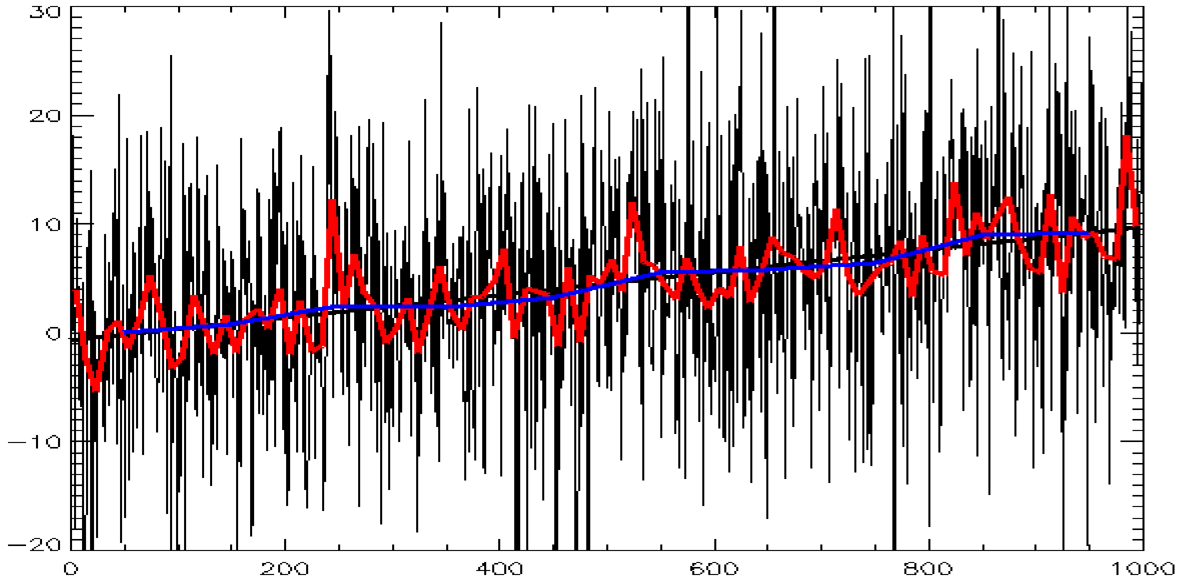
An ideal network structure maximizes the probability of the network given the observed data, i.e. max(P(net|data)). Using Bayes' rule and a constant P(data), max(P(net|data)) = max(P(net,data)/P(data)) = max(P(net, data)). Structure learning algorithms score potential networks based upon this latter property.

Aim is to create a model with as high score as possible.

References :

- Byoungkoo, L., and J. Joseph. "Learning a Probabilistic model of rainfall using graphical models. School of Computer Science." (2006).

2. **Time Series Analysis** : Time series analysis comprises methods of analyzing the time series data in order to extract some meaningful insights from it. Our data is also indexed by time, so time series analysis seems a very natural approach to apply here to use extracted information to predict future.



Time Series analysis assumes that the data consists of a systematic pattern and a random noise which usually makes it difficult to identify the pattern. Time series patterns can be described in terms of two main components: trend and seasonality. Trend represents a general systematic behaviour that changes over time but does not vary within the time range being analysed. Whereas seasonality may have a similar behaviour but it repeats in a systematic intervals of over time.

Most of the time series analysis techniques traditionally used for rainfall prediction fall within the framework of AutoRegressive Moving Average Class (ARMA) of linear stochastic processes. But ARMA model requires data to be stationary, so we use a variation of ARMA called AutoRegressive Integrated Moving Average (ARIMA), which permits the handling of non-stationary data based on differencing of the time series. ARIMA model possesses many appealing features like it allows to forecast future values without requiring data of any other related field like temperature in this case.

ARIMA model is a combination of an autoregressive (AR) process and a moving average process (MA).

AutoRegressive process : Each observation of the time series is made up of random error components (random shock, ε) and a linear combination of prior observations.

Moving average process : Each observation of the time series is made up of a random error component (random shock, ε) and a linear combination of prior random shocks.

ARIMA model is depicted as ARIMA(p,d,q) model, AR(p) refers to order of the autoregressive part, I(d) refers to degree of differencing involved to make data stationary, and MA(q) refers to order of the moving average part.

Four phases are involved in identifying patterns of time series using ARIMA model:

1. Model Identification : The first step is to determine if time series is stationary and if there is any significant seasonality that needs to be modeled.
2. Parameter estimation : To estimate the parameters like p,d,q which can adequately model represent the time series model. This can be done using various algorithms like maximum likelihood algorithm.
3. Diagnostic checking : In diagnostic checking, the residuals from the fitted model are examined against their adequacy. This is usually done using correlation analysis and by goodness of fit test using means of Chi-square statistics.

4. Forecasting : At forecasting stage, estimated parameters are used to calculate new values of the time series, so that we can predict future values.

References:

- Meher, Janhabi, and Ramakar Jha. "Time-series analysis of monthly rainfall data for the Mahanadi River Basin, India." *Sciences in Cold and Arid Regions (SCAR)* 5.1 (2013): 73-84.
- Rivero, Cristian Rodriguez, and Julian Antonio Pucheta. "Forecasting Rainfall Time Series with stochastic output approximated by neural networks Bayesian approach." *Editorial Preface* 5.6 (2014).
- Soltani, S., R. Modarres, and S. S. Eslamian. "The use of time series modeling for the determination of rainfall climates of Iran." *International journal of climatology* 27.6 (2007): 819-829.