Received: April 29, 2019

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

An Adaptive Model for Forecasting Seasonal Rainfall Using Predictive Analytics

Pundru Chandra Shaker Reddy[1,2]* **Alladi Sureshbabu**[3]

[1]Jawaharlal Nehru Technological University, Ananthapuramu, Andhra Pradesh, India
[2]Department of Computer Science & Engineering,
CMR College of Engineering & Technology, Hyderabad, TS, India
[3]Department of Computer Science & Engineering,
JNTUA College of Engineering, Ananthapuramu, AP, India
* Corresponding author's Email: chandu.pundru@gmail.com

Abstract: India is a country which has exemplary climate circumstances comprising of different seasons and topographical conditions like high temperatures, cold atmosphere, and drought, heavy rainfall seasonal wise. These utmost varieties in climate make us exact weather prediction is a challenging task. Majority people of the country depend on agriculture. Farmers require climate information to decide the planting. Weather prediction turns into orientation in the farming sector to decide the start of the planting season and furthermore quality and amount of their harvesting. One of the variables are influencing agriculture is rainfall. The main goal of this project is early and proper rainfall forecasting, that helpful to people who live in regions which are inclined natural calamities such as floods and it helps agriculturists for decision making in their crop and water management using big data analytics which produces high in terms of profit and production for farmers. In this project, we proposed an advanced automated framework called Enhanced Multiple Linear Regression Model (EMLRM) with MapReduce algorithm and the Hadoop file system. In the proposed model (EMLRM) first, we stored the unstructured weather data in hadoop distributed file system (HDFS), process that stored data by using MapReduce Algorithm and build the rainfall prediction model by utilizing Multiple Linear Regression. We used climate data from IMD (Indian Metrological Department, Hyderabad) in 1901 to 2002 period. The experimental outcomes show that the EMLRM provided the lowest value of Root Mean Square Error (RMSE= 0.274) and Mean Absolute Error (MAE= 0.0745) compared with existing methods. The results of the analysis will help the farmers to adopt effective modeling approach for predicting long-term seasonal rainfall.

Keywords: Linear regression, Hadoop, MapReduce, Climate data, Temperature, Rainfall prediction.

1. Introduction

Agriculture plays a vital role in the economy of the nation and each individual need food for their survival. Rainfall can be viewed as the most critical atmosphere component in the hydrological cycle that has significant consequences on the surrounding environment, including human lives. The spatial and temporal circulation of rainfall has a critical impact on the water accessibility of earth surfaces, and hence on the farming activities. Since the agricultural activities and resulting crop production depends on the distribution of rainfall, the forecast of monthly and seasonal rainfall is essentially important for the agricultural planning, flood mitigation approaches. However, precise prediction of seasonal rainfall remains elusive to the researchers. Therefore, seasonal rainfall forecasting becomes plausible amongst the hydrologic researchers around the world [1]. The agriculturists must be helped, with the goal that they will come to know which crop to grow under different climatologically conditions [2]. Fast-growing countries like India 72% of the population depend on agriculture, so we must need accurate weather forecast system for farmers to manage their crop. Cultivation relies upon people as well as on different aspects like atmospheric conditions, water, kind of soil, etc. Weather prediction is always a difficult and challenging task because it is critical to forecasting the upcoming state of the climate [3]. Weather prediction is a subject of meteorology that is furnished by gathering data from the various substations related to the present state of the climate like humidity, temperature, rainfall, wind, etc. These days analysis of a large amount of data is a difficult task and traditional methods are not producing precise results, so we need advanced scientific models/techniques for better weather forecasting [4]. It is repetitive to foresee the climate precisely. Weather datasets are unpredictable, it has several numbers of parameters (temperature, humidity, rainfall, wind speed etc.) involved and complex relations exist among them and oftentimes changes as per global atmospheric changes. Accurate and early predictions can make business opportunities like fishing, airlines, farming, tourism, to name just a few and gives the timing and severity of heavy rainfall like tsunamis, storms, hurricanes, wildfires, floods and other climate events. Unfortunately, enhancing our capacity to predicate climate conditions are a critical task, both computationally and scientifically because of

a) Overseeing and analyzing the huge amount of data sets
b) Expanding model determination
c) Addressing scientific and innovation technology hurdles

Big Data comes with the solution since it turns out to be simple and nearly more affordable to store huge volumes of data. This paper proposes an approach of utilizing Apache Hadoop for processing such huge volumes, variety and speed of climate data. It incorporates the use of Artificial Neural Network which is a helpful approach [5]. ANN is executed on Map-Reduce structure for long-term rainfall forecast. Moreover, a technique will recognize soil and regional investigation which can likewise distinguish trim contingent upon the user’s area. Crop information, soil condition and feasibility of soil are provided so as to benefit for farmers in unequivocal circumstance. Generally, the weather forecast results are not precise. Utilizing big data analytics and predictive analytics to discover when the expectation forecast model might not be exact. Prediction accuracy is improved by enhancing the model. Big data analytics gives the knowledge to give prior decision management, enhance crop production, and minimize unwanted costs related to cultivating, future agriculture, utilization of fertilizers and pesticides. It is significant to remember that none of the predictive models produces 100% precise outcomes for weather forecasting [6]. The main goal of analytics is giving help in decision making, yet an ultimate conclusion is constantly after you.

The primary objective of our paper produces an application that would anticipate climatic and ecological changes. Our task basically focuses around people who live in zones which are inclined conditions calamities, for example, cyclones, floods and furthermore to help farmers in adjusting to atmosphere brilliant climate estimating framework utilizing enormous data approach which increments the productivity and profitability of agriculturists. Aside from predicting the everyday climate, this project would make awareness and alert among people [7]. This paper exhibits an improved automated forecasting model based on Hadoop framework system for proficient and adaptable climate data investigation and predicting system. Conventional data analysis is suitable for only structured data whereas this is works for unstructured data also. The weather forecast is for the most part critical for business class, agriculturists and so forth where they need to plan their work as indicated by the climate [8]. The drawback of the current framework is essentially the procedures utilized as a part of information extraction in turn reflecting the productivity of the model. In this project, we utilize the MapReduce framework from Hadoop to process the offline information gathered from different sources to anticipate the base furthermore, the most extreme rainfall of the specific zone on the earlier year's data.

1.1 Big data

The word Big Data introduced in 2005 which states to datasets that are enormous, additionally high in velocity and variety. So traditional models and methods are not suitable to process the big data. Big data made immense business and interpersonal opportunities in each area, empowering the disclosure of formerly invisible patterns and the advancements of new insights for decision making [9]. The keyword Big Data is currently utilized wherever in our regular day to day existence and it is a recent development and besides going to control the world in future and has grown in light of the way that people and various associations makes growing use of data raised advances [10]. Present scenario big data sizes are running from a TB to ZB in individual data set. As per the survey conducted by IDC, the digital world will grow from 130EB to 40,000 EB (40 trillion GB) in next 15 years (2005 to 2020) i.e. digital data will grow twice every two years until 2020. 85 % of the data in the universe today has been delivered over the most recent two years only.

i. Characteristics of Big data

Big Data has numerous properties depicted by n V's characteristics. Bunch of V's properties of the big data was accumulated from different research distributions to have Nine V's qualities categorized below.

Volume: Big data name itself is related to size and that is tremendous. The volume of the data accepts an aggressively critical part in choosing motivation out of information. Whether a particular data truly be considered as a Big Data or not, it relies upon the size of data.

Variety: It identifies to disparate sources, the structured and unstructured nature. In the old scenario, we considered spreadsheets and databases were only sources of data by the greater part of the applications. In present scenario data in various forms as a text, jpeg, audio, video, observing smart gadgets, PDFs, etc. is also being considered in the examination applications. This assortment of semi-organized and unstructured information speaks to specific issues for storage, extracting and breaking down data.

Velocity: keyword velocity represents accelerate the data generation. How fast the information is delivered and taken care of to meet the demand, decides certified potential in the information. Big Data Velocity stems the agility at which information streams in from origins like business frames, application logs, frameworks and online SMNs, sensors, Mobile gadgets, and so on. The stream of data is gigantic and ceaseless.

Variability: it refers that data streams might be highly inconsistent in regular intervals, day by day and seasonal data can be difficult to manage, particularly with unstructured information included.

Veracity: It describes the inclinations, noise and irregularity of data.

Validity: The information is right, precise for the utilization. Undoubtedly, accurate and correct information is crucial to make valid decisions.

Volatility: When the support period ends, we can without a lot of a stretch pulverize it.

Visualization: it suggests convoluted diagrams that can join a couple of components of data while as yet remaining sensible and intelligible.

Value: It has a low-regard thickness due to removing an enticement from enormous data. Valuable data ought to be extricated from the tremendous amount of data.

ii. Advantages of Big Data Processing

- Business associations utilize further intelligence while making decisions
- Enhanced customer service
- Early discovery of risk to the products
- Better operational proficiency

1.2 Map reduce

MapReduce is a programming model for composing applications that can process Big Data in parallel on different hubs. MapReduce gives expository capacities to analyzing enormous volumes of complex data. Classical Enterprise Systems regularly have an incorporated server to store and process data. Traditional methods are absolutely not convenient to process immense volumes of adaptable data; they can't be obliged to definitive DB servers. In addition, the central framework makes excessively of a bottleneck while preparing different documents at the same time. MapReduce separates a task into small parts and assigns them to different systems called nodes. Afterward, the outcomes are gathered at one place and coordinated into a single frame outcome. MapReduce framework involves two significant assignments, specifically Map and Reduce. The Map assignment uses an arrangement of data and exchanges it into another form of data, where particular components are separated into tuples key – value sets. The Reduce assignment takes the result from the Map as an input, consolidates those data tuples key – value sets into a smaller set of tuples [11]. The MapReduce executing different numerical algorithms to partition a task into small parts and assign them to different frameworks. In other terms, the MapReduce algorithm helps in sending the Map and Reduce works to proper servers in a cluster. The main objective of the Map-Reduce system in Hadoop is to extract knowledge from the sets of data which is stored. MapReduce is a handling huge number of datasets in parallel utilizing more number of computers functioning in a cluster [12]. We can expand the mapper class with our own particular guideline for dealing with multiple inputs in a particular way. Map master node trains worker nodes to process the input data and Hadoop execute shuffle task and master node gathers outcomes from all to answer the overall query described in Fig. 1.

1.3 Hadoop

Hadoop is an open source wide-ranging data handling framework that supports distributed

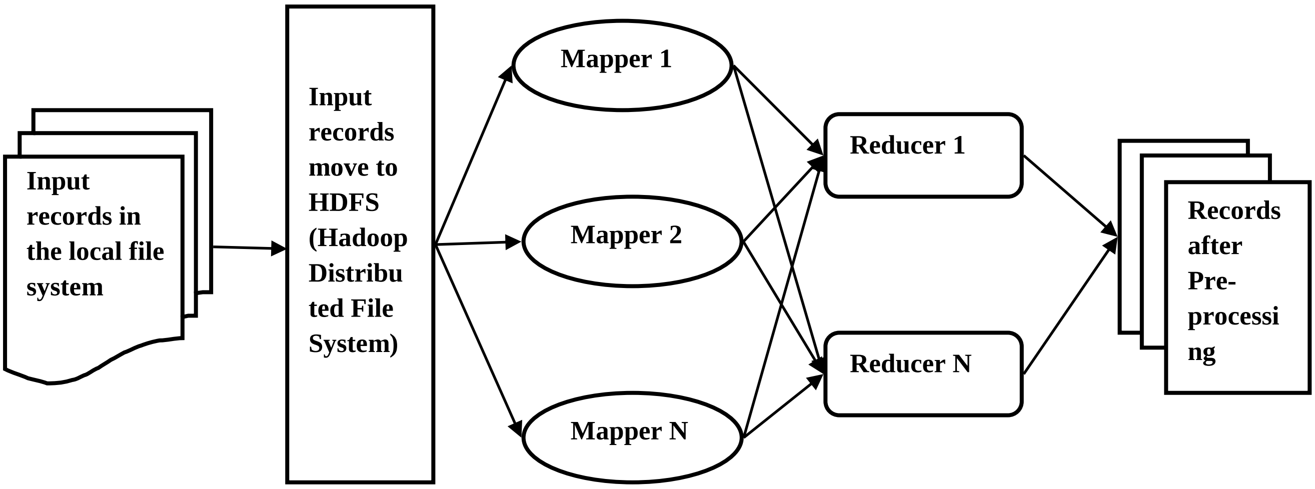
Input records in the local file system
Input records move to HDFS (Hadoop Distributed File System)
Mapper 1
Mapper 2
Mapper N
Reducer 1
Reducer N
Records after Pre-processing

Figure.1 A Map-reduce task that converts raw input to a required format

processing of huge blocks of data utilizing basic programming strategies. The Apache Hadoop project consists of the HDFS and Hadoop Map Reduce notwithstanding different modules. Its modules give simply to utilize languages, graphical interfaces and tools apparatuses for handling the information on a large number of systems. Hadoop cluster is a group of machines networked together in a single location. The user can acknowledge tasks to Hadoop from his work area system in a remote area from the Hadoop cluster. Primary parts of Hadoop will be HDFS, MapReduce. HDFS is a distributed file system for administration of extensive data index sizes of gigabytes and petabytes. Also, MapReduce is a programming algorithm for handling and processing an enormous measure of unstructured data in parallel is based on the division of a large dataset into smaller autonomous blocks. Climate condition investigation and its forecast have a tough task and it needs great efforts. India being a nation with farming as an essential occupation, there is an extraordinary need to forecast information with enhanced exactness.

In the agricultural field, selecting the crop is a most challenging task for farmers and crop productivity depends on rainfall, temperature, wind speed and soil type, etc., if the person is newer in this field then difficult to predict the weather conditions. At first, the farmers have to predict the next season rainfall, then only they can plan the crops based on weather conditions. To overcome these issues machine learning techniques are used to predict the rainfall based on previous weather parameters.

In this paper, we introduce a model of long term seasonal rainfall forecast with weather data from IMD, Hyderabad of Nellore district, Andhra Pradesh utilizing 6 parameters which are temperature (min, max, avg.), cloud cover, vapor pressure and precipitation. Our work will forecast the accurate rainfall for farmers in planting and water management. This research exhibits the efficiency of Multiple Linear Regression Model in forecasting long-term seasonal rainfall estimating. It was performed utilizing the past values of the weather indices as the potential predictors of long-term seasonal rainfall. Seasonal rainfall prediction can be useful to a wide range of users in different sectors like farming, water supply, and storm management [13]. The results of the investigation might be the benchmark for future ages in anticipating seasonal rainfall. Further, the article is divided as follows. Section 2 is discussed existing works; in section 3 we explain data preparation and location, section 4 proposes the methodology of work, in section 5 we discussed experiment execution process, in section 6 illustrates the assessment of model and section 7 states the conclusion of the paper.

2. Related work

In recent times, there is huge research in metrological data analysis by using Hadoop and MapReduce framework. It is a novel framework for illuminating certain sorts of distributable issues and preparing huge datasets. In this way, to manage high dimensions and tremendous datasets, different researchers have proposed various models to solve these issues. Wei Fang, V.S. Sheng, Xue Zhi Wen and Wubin Pan [14] designed and implemented SVM classification technique using Hadoop and MapReduce framework to predict the rainfall from large amount data. They utilized feature selection and reduction algorithm associated with the dataset. The proposed model serves as an application that allows a huge amount of data to be easily analyzed and classified to predict the storm information from the available cluster. They concluded as the proposed system enhances the performance in terms of precision and efficiency. Q. Feng, X. Wen, and J. Li [15] implemented wavelet analysis-support vector machine coupled model (WA-SVM) for predicting rainfall for next 1, 3 and 6 months. This model was obtained by a hybrid of discrete wavelet transform (DWT) and support vector machine (SVM) methods. For this study authors considered monthly rainfall values of China region and evaluated outcomes on the basis of RMASE and MSE. Finally, they concluded as WA-SVM producing superior results compared other models and it was very useful in monthly rainfall forecasting. WA-SVM is predicting precise rainfall for monthly basis only and it is not suitable for seasonal rainfall forecasting.

A. Joseph and M. Lakshmi [16] proposed a model of processing huge volumes of climate data using Hadoop and ANN implemented on MapReduce algorithm for short-term rainfall forecast. This model predicting next day rainfall using previous data (temperature) and they concluded as use different weather parameters with machine learning algorithms to improve the prediction accuracy. V. Sharma, Umit Cali, Veit Hagenmeyer, Ralf Mikut and JAG Ordiano [17] described a method which uses loop non-linear autoregressive artificial neural network (CL-NAR-ANN) model for next day prediction of power generation without using any past data. It is a cost-effective and accurate prediction system without using any data if communication breaks with weather provider also it works. The outcomes show that the CL-NAR-ANN approach produces seasonal forecasts and surpass other NWP free techniques by a margin of 8% in terms of RMSE. This model is limited to next day forecasting only and it does not work for long term forecast.

Ming-Chang Wu and Gwo-Fong Lin [18] proposed toolbox Climate Learn algorithm for rainfall prediction using ML algorithms and climate network analysis and they provided new prediction methods for the occurrence of rainfall in user location. The proposed method of rainfall prediction is based on limited data and it does not handle the huge amount of datasets. Tulasi Sunitha Manepalli and Chamakuzhi Subramanian [19] proposed daily rainfall prediction model called Regression Automata (RA) models of four stations in Queensland State. They concluded that the proposed model is working good and predicting precise results when compared with existing models. This model does not handle unstructured data.

Pushpa Mohan and Kiran Kumari Patil [20] implemented Self Organizing Map (SOM) is proposed along with Latent Dirichlet Allocation (LDA) for weather and crop forecast. It is producing accurate and improved results compared to traditional models up to 7-23%. Various examinations have been inspected to recognize the fitting model for the forecast of seasonal rainfall. I. Hossain, R. Esha and M. A. Imteaz[21] proposed a non-linear regression technique for long term rainfall forecasting and assessed the model efficiency in terms of RMSE, MAE. Authors selected case study as Australian Capital Territory (ACT) and they suggest that the dependents which have the highest correlation with the independents do not necessarily give the least errors in rainfall anticipating. The results of the study will help the agricultural and watershed- related authorities to adopt efficient modelling approach by anticipating long-term seasonal rainfall.

To overcome the above mentioned problems, the EMLR model is implemented to produce the precise and timely seasonal rainfall forecasting. In any case, just a single atmosphere driver isn't equipped for imitating the precise rainfall. Multi-predictors techniques have a higher forecasting ability than single predictor models. By the by, there may exist different attributes of seasonal rainfall patterns with the same rainfall totals. Along these lines, a more intensive look at the appropriate models of seasonal rainfall development turns out to be basically critical. Since there exist noteworthy relationships between's seasonal rainfall and month average values of weather parameters, past values are considered for this work. Enhanced Multiple linear regression model is a helpful approach to a method the relationship between two or more than two explanatory attributes and a response attribute by fitting a linear equation to observed the data. There are two benefits of using this approach to analyze the data.

i. It is the ability to evaluate the relative consequences of one or more predictor attributes to the response value.
ii. it has the capability to determine outliers or anomalies

3. Data Preparation and Location

We assembled weather data from IMD (Indian Meteorological Department, Hyderabad), India. The dataset contains the total of forty-seven attributes includes temperature (min, max, avg), humidity, wind speed, precipitation, wind direction, cloud cover, visibility, atmosphere pressure and other outcomes of the main variables in daily, monthly and yearly basis. For our research, we utilized dataset as monthly mean values of all climate parameters period of 1901-2002(110 years). We selected region for our work as Nellore district in Andhra Pradesh, India.

3.1 Data pre-processing

The data is produced by sensors is unstructured, that turns into a complex task to investigate it. The temperature (max, min, avg), cloud cover, rainfall and other weather parameters rehash up each year dependent on various environments. The averages of climate attributes are handled and computed by utilizing Hadoop, it would be desirable to the upcoming weather forecast. These high volumes of data are loaded onto an HDFS which consists of the number of clusters [22]. This section describes the data gathering and pre-processing of climate dataset. The raw monthly average of weather dataset with the location is collected. These datasets will be in continuous values as the numeric format. The dataset has 10% missing values and 3% outliers and these are rectified because due to this noise we cannot get accurate results. After prepossessing the data, we need to select appropriate attributes which are important for our model. Using feature selection method out of 47 we have chosen 6 parameters which are temperature (min, max, avg.), cloud cover, vapor pressure and precipitation.

4. Methodology

The issue discussed in this article is that of prediction of accurate seasonal rainfall for the following season for next coming years based on the past weather dataset, by utilizing corresponding earlier data. The model was applied to Nellore district which is located on the southeast coast of India, with water of Bay of Bengal. The proposed model is shown in Fig. 2.

Our proposed model having the following steps

1. Data collection from IMD
2. Pre-processing of climate data
3. Apply the model
4. Predict the rainfall
5. Visualization of results
6. Transform the results to the agriculture field

In the MapReduce method, mappers do parallel handling of data utilizing a function defined in it and reducers will summarize the outcomes. At the point when the jar file is termed as the fundamental class runs which receive the inputs and mapper class is conjured here. The mapped input is passed onto the reducer stage which includes every one of the values and computes its mean and composes it to the result file. The task is separated into subtasks executing in parallel with no information bury dependencies. The task is separated into subtasks executing in parallel with no information bury dependencies.

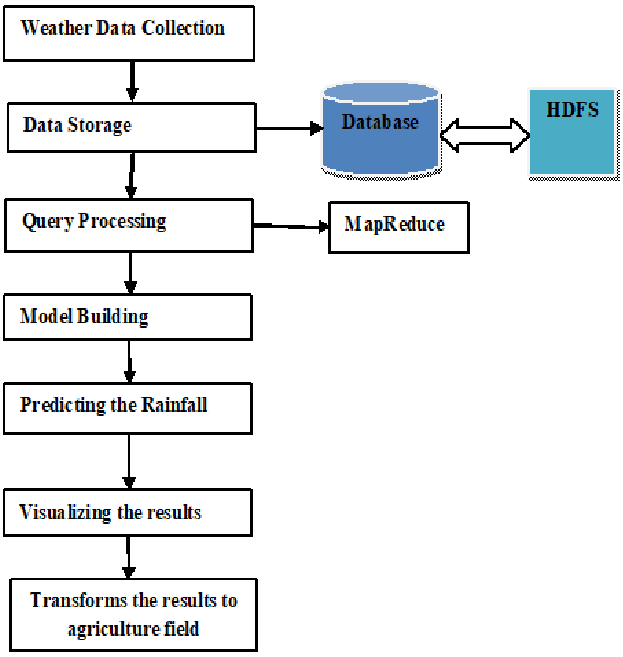

Figure. 2 Flow diagram of the proposed method

This MapReduce based model pre-processed weather data is given as input. Quantifiability of the result is the primary assets of utilizing the MapReduce method. It has been utilized for determining the rainfall on the following month, utilizing rainfall and temperature, etc. estimations of a traverse of past n months. Map reduce execution is shown in Fig. 3 above “the outcomes show adding more numbers of frameworks to the network accelerate the data processing” [23]. In the MapReduce enormous data is processed in parallel. Data is at first portioned over the hubs of a cluster and stored in a distributed file system (DFS).

The calculation of the two functions is expressed formally shown below.

$$map\ (w1,\ R1) \rightarrow list\ (w2,\ R2) \qquad (1)$$

$$reduce\ (w2,\ list(R2)) \rightarrow list\ (w3,\ R3) \qquad (2)$$

The pseudocode of MapReduce is shown in Algorithm1. Below program is used to calculate the monthly maximum temperature.

A Map function is utilized to extract all the months-years and temperatures (Key/value pairs) showed in the text, and these sets are directed to an intermediate temporary space determined by the MapReduce. Through intermediate processing by the Map function, the key/value pairs are clustered as per to the key, so that each month-year is followed by a list of temperatures. Then, a Reduce function is only to find the max value through a whole list. The result is the monthly maximum temperature. Fig.3 demonstrates the intermediate results of each progression of the execution procedure of MapReduce, including Map and Reduce stages, which both utilize all nodes in the cluster. Between the Map and Reduce stages, there is an intermediate stage, which concatenates the intermediate outcomes with the same key into a list. The list will be utilized by the Reduce function to result in the maximum temperature of a certain year month wise. Similarly, the avg min temperature, avg temperature, avg vapor pressure, avg rainfall and cloud cover computed the utilizing the Hadoop-MapReduce program.

MapReduce algorithm procedure

1. Weather data as input is loaded directly into mappers using map (input key, input value) function.
2. It divided data set into smaller subtasks and start computation parallel on each task and finally, it returns a list of <Key, List<Value>> sorted pairs to reduce phase.
3. The reducer takes merged and sorted pairs from the mapper and perform reduce operation to get final (key, value) pairs.

Algorithm 1:

```
map(String input key, String input value):
#input key: document name
#input value: document contents
each year $y$, month m and temperature $t$ in input value:
EmitIntermediate($y$,m, $l$,$t$);
reduce(String output key, Integer intermediate _values):
#output key: year, month
# intermediate _values: a list of temperature values
int maxValue = Integer.MAX VALUE;
for each $t$ in intermediate values:
maxValue = Math.max($t$);
Emit(year, month, maxValue);
```

5. Model evaluation

5.1 Enhanced multiple linear regression model using R tool

Climate dataset which is pre-processed by the Hadoop-MapReduce framework that means outcomes of MapReduce is loaded into R software for prediction and put away in separate vectors. The dataset contains 6 attributes as we discussed in earlier sections with monthly average values and target variable is rainfall. R Tool supports in build function for prediction and visualization. The relation between these vectors is established utilizing the lm() function. Then, predict () function is used to forecast the average monthly rainfall based on the relationship between input parameters. Once data is pre-processed, then data is divided into training and testing with a percentage of 80, 20 respectively and we scaling the attributes into interim keeping in mind the end goal to stay away from impacts of range for the model. In this work, EMLRM was executed to achieve a constant relationship between rainfall and other weather attributes.

Multiple linear regression is represented in the formula as

$$y= \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \epsilon \quad (3)$$

Where x_1, $x_2 \ldots x_n$ is independent predictor variable, β_0 is intercept and ϵ is an error term of regression and y is target variable.

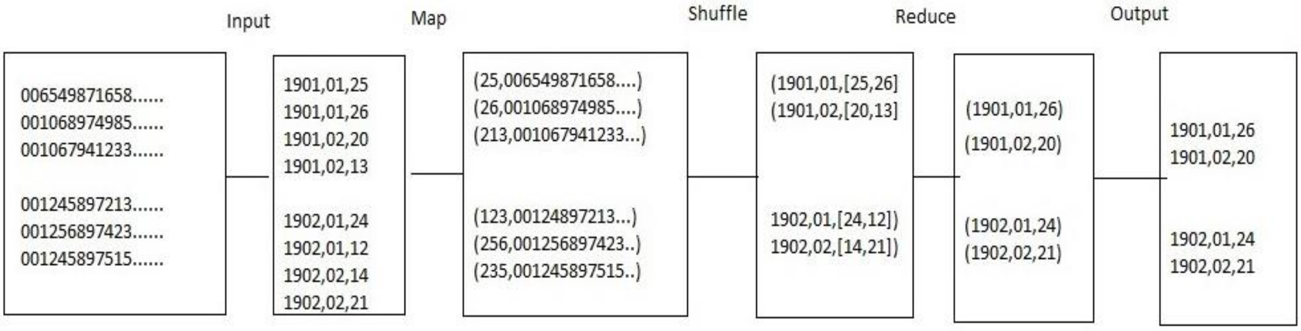

Figure. 3 Systematic execution steps of the MapReduce program for climate data

$$Rainfall = \beta_0 + (Mean\ Temp) \times \beta_1 + (Cloud\ Cover) \times \beta_2 + (Vapour\ Pressure) \times \beta_3 \qquad (4)$$

Where $\beta_1, \beta_2, \beta_3$ are regression coefficients

The outcomes of prediction are visualized graphically to make decisions based on predicted values. Climate forecast is utilized for different purposes like cultivating, fishing, traveling, etc. This paper gives the information to build up the web application which acquires the data including location and individual points of interest of the user and by utilizing the MapReduce and regression methods for training the data. The required data about the climate and its forecast is sent to the user.

5.2 Model building and evaluation

In this paper, we utilized enhanced Multiple linear regression approach for rainfall forecast based on three attributes (mean Temperature, cloud cover and vapor pressure).
Where the mean temperature calculated as the average value of min and max temperature.

Steps:

1. Loading or reading the dataset into R
2. Split the dataset in to train as 80% and test as 20 %
3. Build the prediction model by utilizing lm() as follows

```
PredModel← lm (rainfall ~ MeanT+CC+VP, data = trainData)
```

4. Assess the model by using given the equation
   pred <- predict(PredModel, newdata = test)

5. Compare the actual values with predicted values and compute the value of R-squared for our model on the test dataset.

6. Results and discussions

To assess the model, we need to compute performance criteria relate to RMSE (root mean square error) and Mean Absolute Error (MAE) shown as follows

$$RMSE = \sqrt{\frac{\sum_{m=1}^{N}(f(m)-g(m))2}{N}} \qquad (5)$$

$$MAE = \frac{1}{N}\sum_{m=1}^{N}|f(m) - g(m)| \qquad (6)$$

Where f(m) is actual mean rainfall and g(m) is predicted mean rainfall.

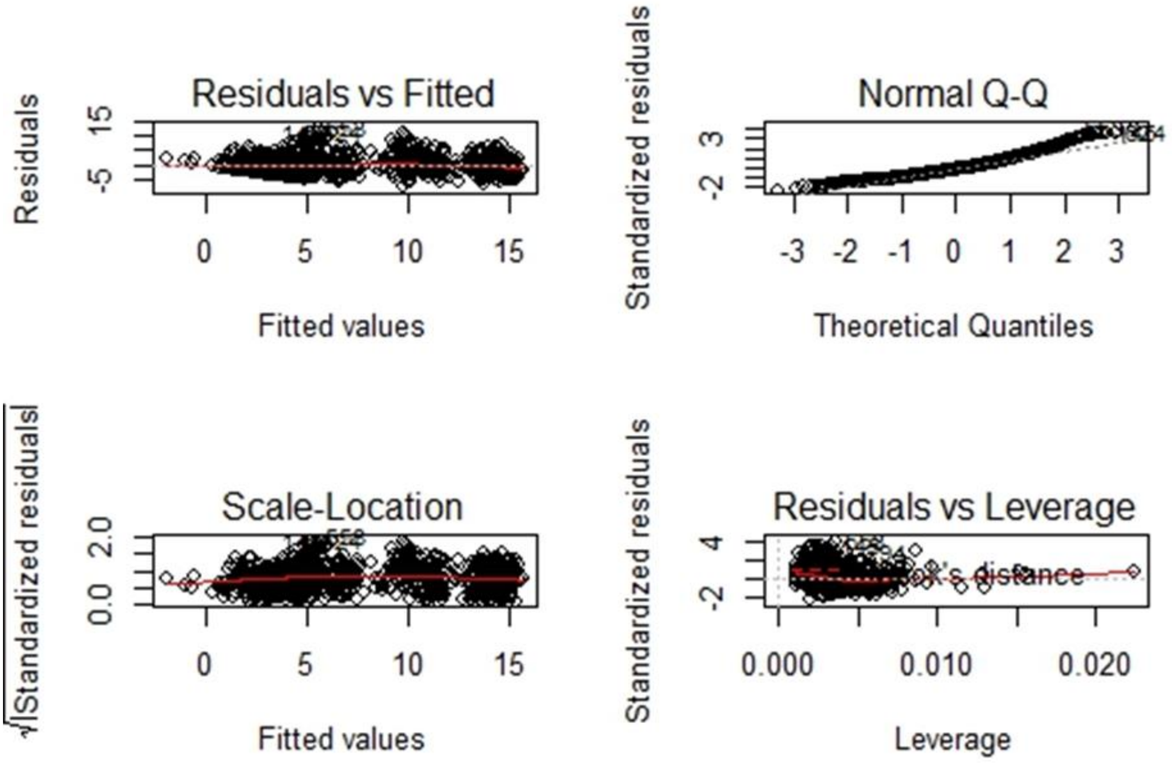

Figure. 4 A residual plot of the EMLR model

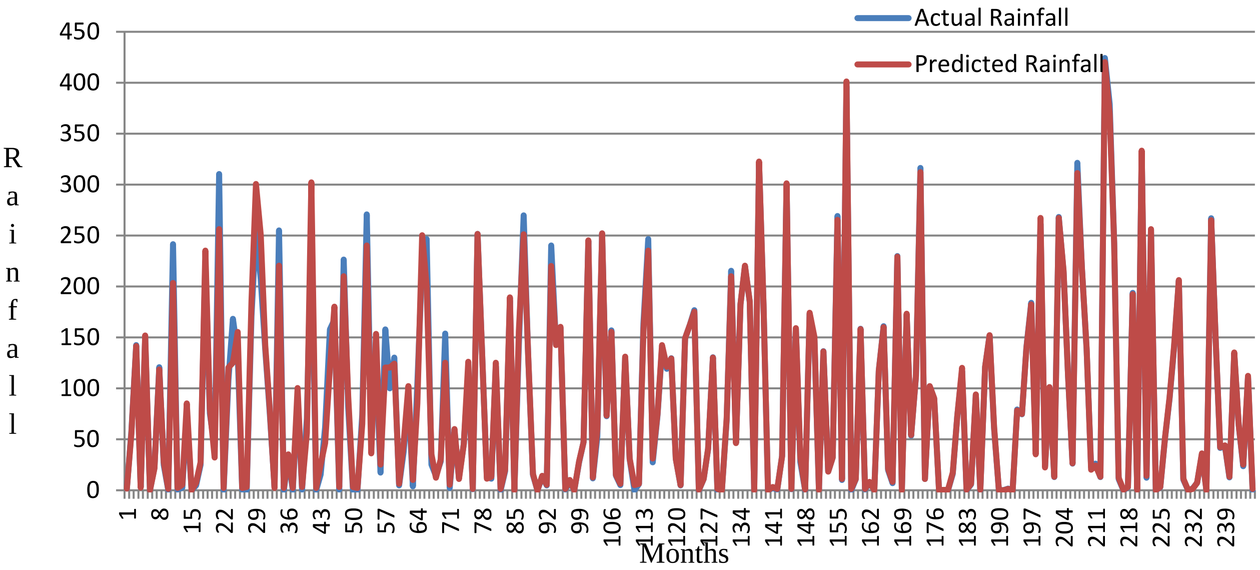

Figure. 5 Actual and predicted rainfall values trend series

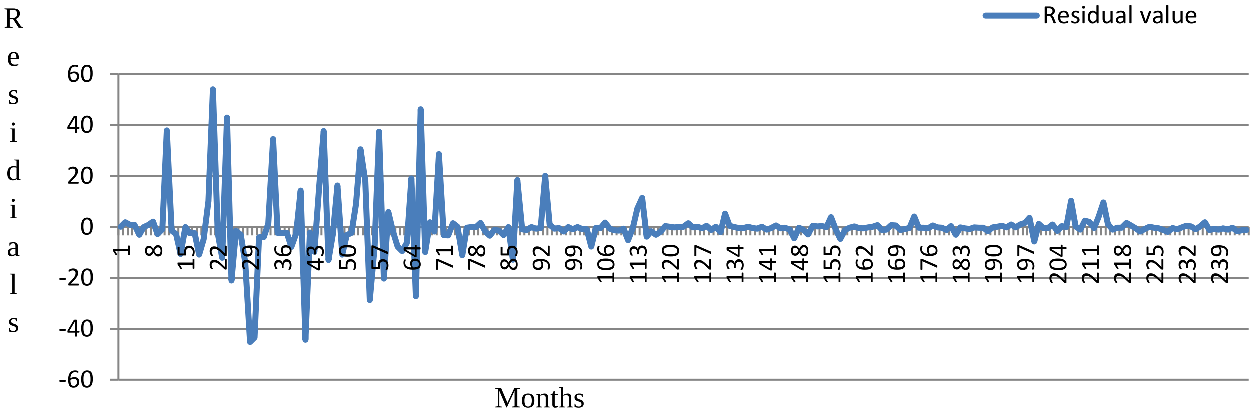

Figure. 6 Residual trend analysis

$$Multiple\ R\text{-}squared = 1 - \frac{SSE}{SST} \quad (7)$$

$$SSE = \sum(P - Q)^2 \quad (8)$$

$$SST = \sum(P - mean(P))^2 \quad (9)$$

Where P represents Actual values of rainfall and Q represents predicted values of rainfall. Where SSE is the sum of the square of residuals, SST is the total sum of squares. It is computed by adding the squares of difference among the actual value and the mean value. R-squared is 1 that implies that it is an outstanding forecast model and it implies 0 that indicates no enhancement over the basic model. Tables 1 and 2 showed the model results and Fig. 4, 5,6 are describes the residual plot, trends of actual & predicted values and residual respectively of the proposed model.

Table. 1 Comparison of actual and predict rainfall values

SNo.	Actual Values (a)	Predicted Values (b)	Residuals (a-b)
1	1.624	**1.84895**	-0.22492
2	57.63	**55.756**	1.874
3	142.431	**141.5697**	0.8613
…..	………	**…………..**	………………..
245	214.807	**210.7268**	5.0802

Table. 2 Model Evaluation

RMSE	**0.274**
MAE	**0.0745**
R-squared	**0.967**

6.1 Comparative analysis

The predicted values of rainfall obtained in this

Table. 3 Comparison of works with existing methods

SNo	Methodology used	RMSE	MAE
1	ANN [17]	10.49	5.41
2	Non-Linear Regression Modeling Technique [21]	15.62	14.2
3	WA-SVM [15]	12.689	7.828
4	**EMLRM**	**0.274**	**0.0745**

study can be used in agricultural and water resource planning, hydrological model study and climate change study. In Table 3 we compared the proposed model (EMLRM) with existing methods in the terms of RMSE and MAE. Mean Absolute Error (MAE) and Root mean squared error (RMSE) are two of the most common metrics used to measure accuracy for regression models. Lower values of RMSE and MAE are indicating better fit the model for prediction. Non-Linear Regression modeling technique [1] is used to predict the long term seasonal rainfall in Australia based on existing climate data and the model produces RMSE and MAE 15.62, 14.2 respectively. WA-SVM [15] model is a combined of wavelet analysis and support vector machine used to predict short term rainfall of next 1, 3, 6 months based on existing weather datasets. WA-SVM model is evaluated and it gives results as RMSE (12.689) and MAE (7.828). ANN [17] implemented on MapReduce algorithm for short term rainfall forecast, and it predicted next day rainfall based on temperature data. The outcome of the model is assessed based on RMSE (10.49) and MAE (5.41). Our proposed method gives lesser RMSE and MAE values compared to existing models.

7. Conclusions and future work

Climate data is very unique and clamorous in nature and has a huge volume of data. An adequate and economical solution for preparing a gigantic volume of climate data is highly needed. Precise rainfall prediction is essential for agriculture dependent countries like India to yield management. This article describes an enhanced multiple linear regression model using Hadoop-MapReduce. This project is an endeavor to perceive how big data solutions can be used in the field of climate forecast and it is made to propose best appropriate yield and its data dependent on agriculturist location and climate condition to crop and fertilizer management. Executing this experiment on Hadoop-MapReduce makes it quicker and adaptable. The utilization of these kinds of advancements for extensive scale data investigations can possibly extraordinarily upgrade the climate forecast too. Regardless of whether the data estimate increases to terabytes or petabytes, the same structure holds great for rainfall anticipating. The experimental outcome of the proposed strategy (EMLRM) indicated better performance over existing strategies. All things considered, the proposed strategy accomplished lower RMSE & MAE values as 0.274 & 0.0745 respectively. More thorough statistical comparison of the result of the proposed method shows that it is the optimal approach for predicting rainfall. The proposed design has been contrasted with other cutting-edge models. The outcomes recommend that our proposed design beat other models regarding the MSE, RMSE and in terms of various forecasting techniques, our experimental outcomes demonstrate that created model execute the outstanding for monthly rainfall based on various criteria. As future work, we intend to enhance our model accuracy in weather forecast using ANN for coming months prediction with the location for exact to decision making for agriculture field.

References

[1] [1] E. Ricciardelli, A. Cersosimo, D. Cimini, and F. D. Paola, "Analysis Of Heavy Rainfall Events Occurred In Italy By Using Numerical Weather Prediction, Microwave And Infrared Technique", In: *Proc. of International Conf. On Geoscience and Remote Sensing Symposium*, pp.1-8, 2018.

[2] B. Anurag, M. Prakash, V. Kanna, and P. Choudhary, "Weather Forecasting using Map-Reduce", *International Journal of Innovative Research in Computer and Communication Engineering*, Vol. 5, No. 9, pp.1-8, 2017.

[3] P. ChandrashakerReddy and A. Sureshbabu, "Survey on Weather Prediction using Big Data Analystics", In: *Proc. of International Conf. On Electrical, Computer and Communication Technologies*, pp.1-6, 2017.

[4] M. Senthilkumar, N. Manikandan, U. Senthilkuma, and R. Samy, "Weather Data Analysis Using Hadoop", *International Journal of Pharmacy and Technology*, Vol.8, No.4, pp.21827-21834, 2016.

[5] V. Dagade, L. Mahesh, A. Supriya. and K. Priya, "Big Data Weather Analytics Using Hadoop", *International Journal of Emerging Technology in Computer Science & Electronics*, Vol.14, No.2, pp.194-199,2015.

[6] M. Joshi, S. Shaikh, and P. Waghmode, "Farmer Buddy-Weather Prediction and Crop Suggestion using Artificial Neural Network on

Map-Reduce Framework", *International Journal of Computer Applications*, Vol. 159, No. 7, pp. 1-3,2017.

[7] C.P. Shabariram, K.E. Kannammal, and T. Manojpraphakar, "Rainfall Analysis and Rainstorm Prediction using MapReduce Framework", In: *Proc. of International Conference on Computer Communication and Informatics*, pp.1-6, 2016.

[8] Q. Xiaoyun, K. Xiaoning, Z. Chao, J. Shuai and M. Xiuda, "Short-Term Prediction of Wind Power Based on Deep Long Short-Term Memory", In: *Proc. of International Conference on Asia-Pacific Power and Energy*, pp.1148-1152, 2016.

[9] R. Basvanth and B.A. Patil, "Weather Prediction Based on Big Data Using Hadoop Map Reduce Technique", *International Journal of Advanced Research in Computer & Communication Engineering*, Vol.5, No.6, pp.1-6, 2016.

[10] K.A. Ismail and M. Abdulmajid, "Big Data Prediction Framework for Weather Temperature Based on MapReduce Algorithm", In: *Proc. of International Conference on Open Systems*, pp. 1-6,2016.

[11] Doreswamy and G. Ibrahim, "Big Data Techniques: Hadoop And Map Reduce for Weather Forecasting", *International Journal of Latest Trends in Engineering and Technology*, Special Issue, pp.194-199, 2016.

[12] S. Selvaragini, and E. Venkatesan, "Big Data Techniques For Weather Forecasting", *International Journal of Pure and Applied Mathematics*, Vol.116, No.18, pp.195-201, 2017.

[13] M. Navid and N.H. Niloy, "Multiple Linear Regressions for Predicting Rainfall for Bangladesh", *Science PG Communications*, Vol.6, No.1, pp.1-4, 2018.

[14] W. Fang, V.S. Sheng, W. XueZhi and W. Pan, "Meteorological Data Analysis Using MapReduce", *Hindawi Publishing Corporation Scientific World Journal*, Vol.4, No.3, pp.1-10, 2014.

[15] Q. Y. Feng and R. Vasile, "Climate Learn: A Machine-learning Approach for Climate Prediction using Network Measures", *Journal of Geosci. Model Dev. Discuss*, Vol.15, No.4, pp.1-18, 2016.

[16] A. Joseph and M. Lakshmi, "Storm Analysis with Raw Rainfall Dataset by using Artificial Neural Network and Min-Max Algorithms", *Indian Journal of Science and Technology*, Vol. 9, No.10, pp.1-5, 2016.

[17] V. Sharma, C. Umit, V. Hagenmeyer, R. Mikut, and J. Ordiano, "Numerical Weather Prediction Data Free Solar Power Forecasting with Neural Networks", In: *Proc. of International Conference on Future Energy Systems*, pp.604-609, 2018.

[18] W. Ming-Chang and L. Gwo-Fong, "The Very Short-term Rainfall Forecasting for a Mountainous Watershed by means of an Ensemble Numerical Weather Prediction System in Taiwan", *Journal of Hydrology*, Vol.546, pp.60-70, 2017.

[19] M.T. Sunitha and C. Subramanian, "Time Series Analysis of Large Scale Rainfall Data Using Regression Automata Models", *International Journal of Intelligent Engineering and Systems*, Vol.11, No.6, pp.118-127, 2018.

[20] P. Mohan and K.K. Patil, "Deep Learning Based Weighted SOM to Forecast Weather and Crop Prediction for Agriculture Application", *International Journal of Intelligent Engineering and Systems*, Vol.11, No.4, pp.167-176, 2018.

[21] I. Hosain, R. Esha and M.A. Imteaz, "An Attempt to Use Non-Linear Regression Modelling Technique in Long-Term Seasonal Rainfall Forecasting for Australian Capital Territory", *GeoSciences*, Vol.8, No.8, pp.282-293, 2018.

[22] K. Namitha, A. Jayapriya and G. Santhosh Kumar, "Rainfall Prediction using Artificial Neural Network on Map-Reduce Framework", In: *Proc. of International Symposium on Women in Computing and Informatics*, pp.1-6, 2015.

[23] J.A. Suyatno, F. Nhita, and A. A. Rohmawati , "Rainfall Forecasting in Bandung Regency using C4.5 Algorithm", In: *Proc. of International Conf. on Information and Communication Technology*, pp.1-6, 2018.