Novel Approach for Spatiotemporal Weather Data Analysis

Radhika T V[1], Dr.K C Gouda[2], Dr. S Sathish Kumar[3]
Dept. of Computer Science & Engineering, RNS Institute of Technology, Bengaluru, India[1]
Council of Scientific & Industrial Research (CSIR), Fourth Paradigm Institute (4PI), Bengaluru, India[2]
Dept. of Information Science & Engineering, RNS Institute of Technology, Bengaluru, India[3]

***Abstract*—Massive volumes of multidimensional array-based spatiotemporal data are generated by climate observations and model simulations. The growth in climate data leads to new opportunities for climate studies at multiple spatial and temporal scales. Managing, analyzing and processing of big climate data is considered to be challenging because of huge data volume. In this work multidimensional climate data such as precipitation and temperature are processed and analyzed in the Spark MapReduce Platform, since Spark platform is computationally more efficient than Hadoop-MapReduce Platform of same configuration. In temporal scale monthly and seasonal analysis of climate data has been carried out to understand the regional climate. The prediction of Rainfall on monthly and seasonal time scales is very much important for planning and devising agricultural strategies and disaster management, etc. As the prediction of climate state is very challenging, in this study an attempt is being made for the prediction of the rainfall using the time series analysis in the same framework. As a test case the time series approach such as Auto Regression Integrated Moving Average (ARIMA) has been used for the prediction of rainfall. The proposed approach is evaluated and found to be accurate in the analysis and prediction of climate data and this will surely guide for the researcher for better understanding of the climate and its application to multiple sectors.**

***Keywords*—Spatiotemporal; big climate data; spark; ARIMA**

I. INTRODUCTION

Big climate data are preferably provided to scientists for on-demand processing and for analyzing critical problems which may help them to relieve from time consuming computational tasks. Since processing of big climate data requires efficient data management approaches, scalable computing resources and complex parallel computing algorithms, so dealing with this problem is considered to be more challenging task. To address these kinds of challenges, high performance computing technologies have been applied to climate data analysis, modeling and prediction [3]. Many exhaustive big data analytics applications have evolved based on big climate data, also the emergence of various technologies such as Internet of Things (IoT), cloud computing and many advanced Big Data analytics tools have begun to investigate on climate, as well as established various intelligent analytic platforms and new technological advancements have further emphasized its importance and potential impacts on climate science and Big data science development [14][15].

Traditional Big Data techniques are usually incapable of handling large amounts of spatiotemporal data. For example, research has added spatial indexing, spatiotemporal indexing [1] and trajectory analytics features to Hadoop. One of the basic idea of using spatiotemporal data, with respect to large spatial database systems, is the emergence of moving objects [4]. A moving object is a spatial object that varies in geographical position or dimensions over a time period [6]. For Example, Rainfall in one region differs from others; also a river that switches its path over a geologic time scale may be represented as a moving line. Moving region can also be indicated by a hurricane which switches its dimension and geographic position as it evolves [16]. Thus there is a need for High Performance Computing (HPC) environment to process big spatiotemporal data.

The number of cores in HPC environment is persistently getting increased depending on the application's requirement. Since these applications generate large volumes of spatiotemporal data, that will ultimately be stored and accessed in parallel [15]. The Scientific applications like weather prediction models use standard high-level libraries and data formats such as Network Common Data Format-4 (NetCDF- 4) and Hierarchical Data Format 5 (HDF 5), which will helps to store and operate on the dataset that is situated inside a parallel file system interface. There are various file formats and software libraries in order to reduce the restrictions imposed by plain binary files. Because of which NetCDF file format has been introduced for systematic reading and writing of various kinds of scientific data, mainly for array data. NetCDF file is composed of various kinds of data, which includes BYTE, CHAR, SHORT, LONG, FLOAT, and DOUBLE. The main intention of NetCDF file is to store rectangular arrays of data such as Interactive Data Language (IDL) arrays [12]. NetCDF files are self-descriptive; that is, every file consists of the basic information required to read.

Big spatiotemporal data have gained huge attention in recent years. Analyzing such massive amount of multidimensional data is one of the most common requirements today and processing of this data is considered to be most challenging task. The ability to assess global concerns such as climate change and natural disasters, as well as their influence on different sectors such as agriculture and disease, requires efficient data processing. This is challenging not only because of the large data volume, but also because of the intrinsic high-dimensional nature of the climate data. The

emergence of Apache sparks provides quicker solution for big spatiotemporal data analysis and processing speed has reduced drastically compared to the traditional way of processing multidimensional data with multi core processors.

In the proposed work we have used Spark MapReduce Framework for processing of big spatiotemporal data at multiple spatial and temporal scales. In the proposed work we have also considered time series rainfall data, we have read rainfall data (precipitation) of past 11 years (2010-2020) and identified the Box-Jenkins time series seasonal ARIMA approach for prediction of rainfall for Bengaluru region on monthly scales. Seasonal ARIMA model(2, 0, 2) (0, 0, 0) for rainfall was identified the best model to forecast rainfall for next 5 years with confidence level of 76 percent by analysing last 11 year's data (2010-2020).

Apache Spark is an integrated platform for cluster computing to facilitate efficient big data management and analytics [13]. It is a non-proprietary, distributed computing scheme which enhances the MapReduce framework. Spark system is made of various main modules including Spark core and various high level libraries such as Spark's MLlib for machine learning, GraphX for graph analysis, Spark Streaming for stream processing and Spark SQL for structured data processing [17]. It functions as a consolidated tool for Machine learning, SQL, Streaming and Graph processing and it supports batch, interactive and stream processing.

Spark is considered to be one of the excellent platform for Data Scientists as it has number of data-centric tools which may assist the data scientists to move forward ahead of the problems that is pertinent in a single machine and also it assist data engineers since it has an integrated method that takes out the need to utilize various special-purpose tools for streaming, machine learning, and graph analytics [13]. More importantly Spark is very essential for researchers, as the platform fosters new opportunities and ideas to design and develop distributed algorithms and also to test their performance in various clusters [9].

The rest of this paper is organized as follows: Section 2 provides overview of various research on Hadoop based approaches to process array-based multidimensional spatiotemporal data; Section 3 presents our proposed Spark-based approach to process multidimensional spatiotemporal data and provides highlights on prediction of rainfall using ARIMA model; Section 4 describes evaluation results of our proposed work by performing sequence of experiments; finally, Section 6 gives summary of the proposed research and envisage on future enhancement.

II. Background Study

Big Data analytics has evolved with advanced opportunities for research, development, business and innovation. It has been identified by four Vs: volume, velocity, veracity and variety and may deliver value via processing of Big Data [2]. The conversion of these four Vs into the 5th (value) is one of the magnificent challenges for processing capacity. The emergence of Cloud Computing as a new standard is to provide computing as a utility service is to deal with various processing needs such as on demand services, pooled resources, elasticity, broad band access and measured services. The capability of delivering computing capacity promotes a possible solution for the conversion of Big Data's 4 Vs into the 5th (value). The continuously increasing volume of big data has accelerated technological developments and practical applications.

Earth is composed of complex dynamic system; as big data analytics works with vast amounts of climate data, it poses greater challenges in climate research than in any other field [3]. Climate change is the present concern throughout globe and also a data-intensive subject, making it one of the main research area for big data experts in recent decades [4][5]. The anomalous growth of climate data makes climate data to be a candidate in the Big Data research. The climate scientists have been exploring on historic data to understand the physics and dynamics, merge millions & billions of daily global observational records and undertake simulations of various climate-change scenarios, all of which leads to huge volumes of data [8].

Extremities in climate such as floods, droughts, and cold and heat-waves may lead to considerable impact on society, ecology and also on the economy globally [7]. Thus spatiotemporal data acquisition, analysis, management and processing are considered to be more important, which will be helpful for various sectoral applications. Spatiotemporal data refers to the data which is connected to both space and time and is considered to be at least 2-dimensional and often 3-dimensional, such that the volume of data gets increased at tremendous speed [8]. Since the general database cannot manage such large volumes of data, there is a need of large database software to play a significant role in the management of spatiotemporal data. Big data is collected from a range of sources, archived, and processed in a variety of computing modes, including cloud computing, mobile computing, edge computing, and wearable computing.

Spatiotemporal data mining is the process of identifying interesting patterns and critical information from spatiotemporal data. Discovering weather patterns, anticipating earthquakes and storms, exposing the progressive history of towns and regions, and identifying global warming trends are the examples of such processes. The unusual rise in spatiotemporal data, combined with the introduction of new technologies, has increased the demand for automatic spatiotemporal knowledge realization. Spatiotemporal data mining techniques are very much essential for many organizations which take decisions based on huge spatiotemporal datasets. As these data are multidimensional in nature, the complexities of such data and their interrelationships create computational and statistical challenges [11].

Researchers of climate science have been exposed to ample of recognized resources of Big climate data for analysis and prediction ,for instance, the NASA Global Climate Change (climate.nasa.gov), the Climate Observing System (GCOS), NASA Center for Climate Simulation (nccs.nasa.gov), Global Earth System Grid Federation (esgf.llnl.gov), the National Center for Atmospheric Research (ncar.ucar.edu), United Nations Global Pulse

(unglobalpulse.org), the Climate Data Guide (climatedataguide.ucar.edu), and many other international and national climate analysis and monitoring centers over the world.

Multi-dimensional, array-based data model are mainly used to represent Climate data .The GRIB, HDF and NetCDF are the three most commonly used data formats to store climate data. HDF5/NetCDF4 was mainly developed to enable support for nested structures, ragged arrays, unsigned data types, chunking data structure, and caching techniques which ultimately helps to systematically organize climate science data and to have control over the changing computer models [10]. Meanwhile in order to flexibly use data as multi-dimensional arrays, many software and libraries such as Panoply, h5py, and NetCDF-Java were introduced.

These real time standard software and data formats have added major benefits to store, acquire, examine and exchange climate data. Also there are number of tools available for performing climate data analytics and visualization, one of such tool is Apache Open Climate Workbench, a Python-based tool to carry out interpretations on climate science employing remote sensing rainfall data taken away from various sources and also using climate model outputs.

Since the above mentioned tools and libraries deal with only discrete machines and have restrictions on cloud computing systems, compatibility with HPC and scalability. The absence of proper libraries leads to difficulty in dealing with variety, veracity, format and resolution of Big Climate Data that give rise to a challenge in the emergence of advanced computing technologies.

1) Big climate data management and analytics: In [19] authors have presented a case study supervised by Deutscher Wetterdienst (DWD) which includes storage of array based multidimensional raster data with hands-on exposure on extraction and processing of gridded meteorological data sets. As the big data acquire various challenges such as repositioning, managing and processing with high computational requirements [18]. One of the key resolutions to this is achieved through the database system having the capability of parallel processing and distributed storage. In [19] authors have conducted a study on processing of the multi-temporal satellite image data using SciDB, which is an array-based database mainly used to accumulate, manage and perform computations on such data. The main goal of the proposed work is to provide elastic solution using SciDB to accumulate and execute time series analysis on multi-temporal satellite imagery.

In [21] authors have illustrated the working of SpatialHadoop, It is regarded as one of the first capable open-source MapReduce frameworks to support spatiotemporal data. The working of ST-Hadoop have been illustrated in [20],which has given a support for spatio-temporal data and considered to be one of the first proficient open-source MapReduce framework. In [22] authors have introduced SciHadoopa, a Hadoop plugin that aids scientists in identifying logical queries in data models based on arrays. SciHadoop was used to run queries as map/reduce programs over the logical data model. Authors have shown implementation of a SciHadoop paradigm for NetCDF data and evaluate the performance of five separate optimizations that address the following goals representing an integrated aggregate function query.

2) Time series analysis for rainfall prediction: Time series analysis is a statistical technique that deals with time series data, or trend analysis. Time series data means that data is in a series of particular time periods or intervals. The data is considered to be in three types, such time series data which includes set of observations on the values that a variable takes at different times, Cross-sectional data which is the data made of one or more variables, collected at the same point in time, Pooled data which is a combination of time series data and cross-sectional data. Various research groups attempted to predict rainfall on a seasonal time scales using different techniques. Below we have discussed existing work done related to rainfall prediction using ARIMA.

Climate and rainfall are highly non-linear and complicated phenomena, which require classical, modern and detailed models to obtain accurate prediction. Authors in [23] have considered various statistical models for prediction of rainfall time series data for designing a model, models such as the statistical method based on autoregressive integrated moving average (ARIMA), the emerging fuzzy time series(FST) model and the non-parametric method(Theil's regression) were used. To evaluate the prediction efficiency, they have used 31 years of annual rainfall data from year 1982 to 2012 of Ibadan South West, Nigeria. ARIMA (1, 2, 1) was used to derive the weights and the regression coefficients, while the theil's regression was used to fit a linear model. The performance of the model was evaluated using Mean Squared Forecast Error (MAE), Root Mean Square Forecast Error (RMSE) and Coefficient of determination.

To forecast future climatic data, the ARIMA model was utilized. The authors in [24] have proposed ARIMA based daily weather forecasting tool which they have considered as case study for predicting weather of Varanasi. The authors have implemented the ARIMA algorithm in R to create an ARIMA-based weather forecasting tool. The Indian Meteorological Department provided 65 years of daily meteorological data (1951-2015) for this study. The accuracy of the model was calculated according to the root mean square error (RMSE) estimated for each forecasting. They approximated future values for the following fifteen years using ARIMA (2, 0, 2) for rainfall data and ARIMA (2, 1, 3) for temperature data. The root means square error values for rainfall and temperature data were 0.0948 and 0.085, respectively, indicating that the technique functioned correctly. The outcome of this can be further used for the management of solar cell station, agriculture, natural resources and tourism. The error is regarded to be minimal by observing at the values of RMSE, indicating that the ARIMA model has forecasted the data properly.

III. Research Methodology

In the proposed work we have considered Spark MapReduce framework which is considered to be one of the excellent platform for Data Scientists as it has number of data-centric tools which may assist the data scientists to move forward ahead of the problems that is pertinent in a single machine.

1) Data analysis and processing using spark map reduce: As weather data is considered to be multidimensional array based, so in the proposed work, we have considered precipitation and temperature data of Bengaluru region for rainfall prediction and also seasonal weather analysis has been carried out on other states of India. Various experimentation are carried out by reading, analyzing and processing the data. NetCDF data are procured from National Center for Environmental Prediction (NCEP) & India Meteorological Department (IMD) has been used. Following are the work carried out.

- Initially Raw station-level NetCDF based temperature and precipitation data of Bangalore district which is located between 12^o latitude and 77^o longitude has been read in the Google Colab Environment. The data considered for analysis is from Jan 2010-Dec 2020 (11 years data). Data from each year are displayed as a single plot and also 11 years data is also plotted as single graph to analyze past 11 years data and use it for processing to assist in future prediction.
- Mean value has been computed for every year (Jan-Dec) using past 11 years data and plotted as a single point in a graph for analysis.
- Mean value has been computed for all 11 years using Spark MapReduce Platform and plotted as a single graph. This step is considered to be more important as data is effectively processed using spark MapReduce platform for analysis and future weather prediction. The detailed diagram illustrating how the data is processed using spark platform can be seen in (Fig. 1). Following are the steps
  - First step is to import and execute main library files for setting up Spark MapReduce functions in google colab environment.
  - Raw station level precipitation data (pr_wtr) of banglore (from Jan 2010-Dec 2020) is read individually and data frame for each year are created.
  - New data frame is created by adding years as columns (total No. is 12) ie from 2010-2020 and values of corresponding year are placed in the appropriate place and convert the datframe to .csv file.
  - Split Data: As spark MapReduce works by splitting the data and assigning key-value pairs (key is day and value is pr_wtr). In this step, we split the data row wise, and perform read operation using spark.read.option() function. Each row refers to daily data of every year i.e. row 1 is Day 1 data from 2010 to 2020, similarly next row is day 2 data from 2010 to 2020. Same applies till last row which is day 365 from the year 2010-2020. Temporary last column is created in data frame to hold the final row-wise mean value.
  - Map phase: This step computes sum of all the values in each row and calculate 'n' value, where 'n' is No. of columns(2010-2020).we use the formula, n=lit(len(df.column)-1.0) and use the value of it in the next step.
  - Reduce Phase: Row mean is calculated using reduce function of spark ie using reduce ((add,(col(x) for x in df.column[1:]))/n).alias("11 years mean") and the same is displayed.
  - Aggregate Phase: This step Aggregates all mean values, place that in last column created in step 4.
  - Finally display the aggregated precipitation value as a single plot.

2) Seasonal analysis: In the proposed work, we have considered seasonal analysis of temperature and precipitation data of Bangalore to analyze the state of weather during various seasons such as pre-monsoon (March 1-May 31), monsoon (June 1 to September 30), post-monsoon (October 1 to December 31). Various graphs were shown to illustrate season-wise analysis of the weather of particular region such as Bangalore. Comparison of weather status of different cities is undertaken. The results of the same are discussed in Section IV.

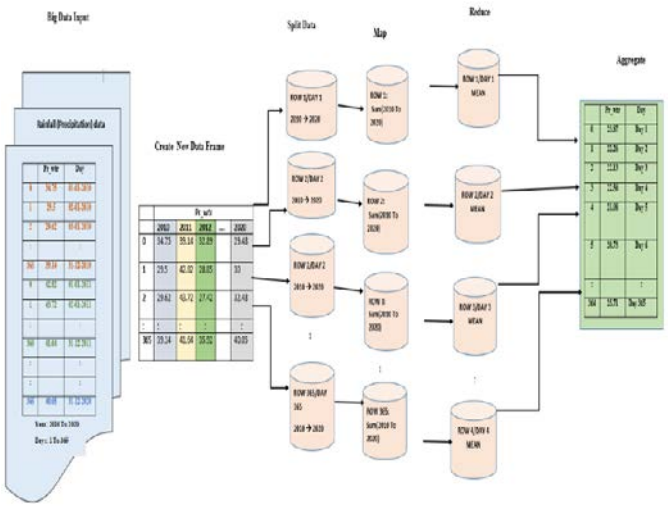


Fig. 1. Spark MapReduce Model to Compute Aggregated Climate Parameters, namely, Precipitation and Temperature.

3) Time series forecasting using ARIMA model: In this study, we have considered past 11 years data and trained them using ARIMA (autoregressive integrated moving average) model. The trained model is used for future forecasting.ARIMA is a class of models that predicts a given time series based on its own past values. An ARIMA model is one where the time series was differenced at least once to make it stationary.

The working principle behind autoregressive (AR) model is that there is a relationship between the present value and the past values. It means that the present value is equal to past values adding with some random value. Moving average (MA) model says that present value is related to the residuals of the past. AR is not capable of forecasting nonlinear data; it can be utilized for data which are linearly related. Using AR and MA together will give best results. But it can be used for stationary weather data and forecasting short term weather. So the proposed work considers ARIMA model which works good for long-term rainfall prediction. We worked on ARIMA (2,0,2) for rainfall data**.**Following steps are used for time series forecasting of rainfall using ARIMA

1) Plot the data.
2) Make the data stationary.
3) Identify the model technique best suited for rainfall forecasting. In the proposed work we have used ARIMA model.
4) Build the model.
5) Compute the mean and Root Mean Squared error (RMSE) value. Use the same for finding accuracy of model.
6) Do the future forecasting based on accuracy of ARIMA model.

Generalized equation used in ARIMA model is as as shown below (1).

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p}\, \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \cdots + \phi_q \varepsilon_{t-q} \quad (1)$$

Where α is intercept term, β_1 is is the coefficient of lag1 that the model estimates, Y_{t-1} is the coefficient of lag1 that the model estimates.

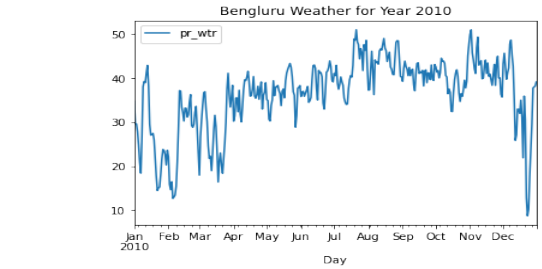
IV. Results and Discussion

The proposed work is executed in Google Colab environment. Python code is used for implementation and necessary libraries were imported. Following are the results. In Fig. 2(a-c) precipitation rainfall data is read individually and plotted as separate graph. Whereas Fig. 2(d-e) shows 5 years and 11 years plot as single graph.

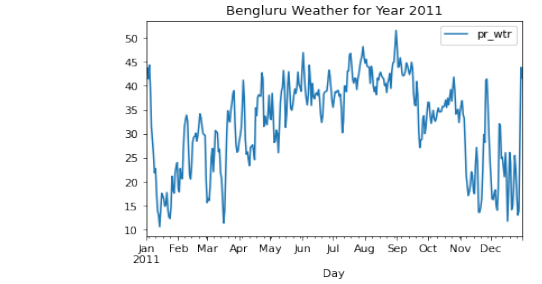
Next we have computed Mean value for every year (Jan-Dec) using 11 years data (2010-2020) and plotted as a single point in a graph for analysis. The same is plotted in line and bar graph as shown in the Fig. 3(a-b).

Mean value has been computed for 11 years (Jan 2010-Dec 2020) using Spark MapReduce Platform and plotted as a single graph. Fig. 4 shows how aggregated mean value is placed in new data frame and shows the final plot after applying to Spark MapReduce.

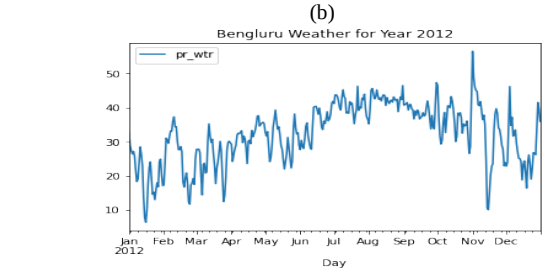
The Table I shows overview of daily dataset of perceptible water (in mm) for rainfall prediction from the years 2010 to 2020. These daily data of past 11 years has been processed using Spark MapReduce Platform which gives aggregated result as shown in Table II. The same result is used in the analysis and prediction of future rainfall.



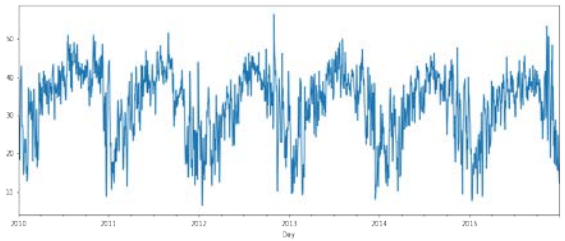
(a)



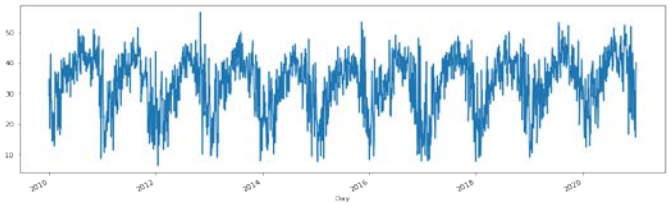
(b)



(c)

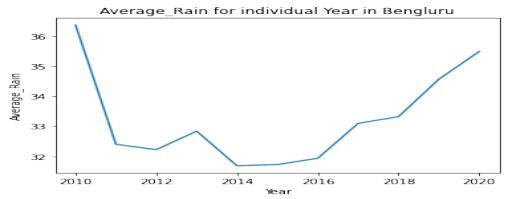


(d)

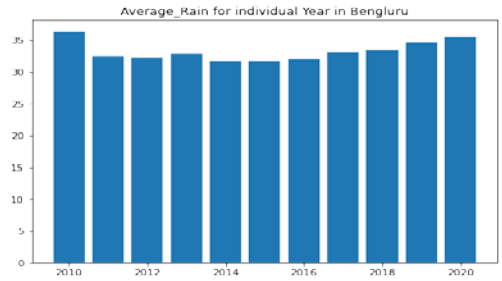


(e)

Fig. 2. (a-e): Raw station-level 11 Years Daily Data (of Bangalore Region) has been Read and Plotted Individually and also as a Single Graph for Comparison.



(a)



(b)

Fig. 3. (a-b): Mean Value for Every Year (Jan-Dec) using 11 Years Data (2010-2020) and Plotted as a Single Point in a Graph for Analysis.

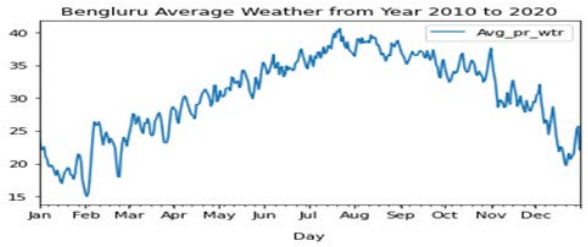


Fig. 4. Final Plot from Aggregated Mean Precipitation Values of 11 Years (2010-2020) using Spark MapReduce Model.

TABLE I. DAILY RAINFALL (PRECIPITABLE WATER) DATASET FROM THE YEAR 2010-2020

Days	Pr_wtr				
	2010	2011	2012	……..	2020
0	34.75	39.14	32.89	……..	29.48
1	29.5	42.82	28.85	……..	30
2	29.62	43.72	27.42	……..	32.48
……	……	……	……	……..	……
364	39.14	41.64	35.92	……..	40.05

TABLE II. AGGREGATED DAILY RAINFALL DATA (PRECIPITABLE WATER) RESULT AFTER PROCESSING 11 YEARS DATASET IN SPARK MAPREDUCE PLATFORM

	Day	Pr_wtr
0	Day 1	23.87
1	Day 2	22.26
2	Day 3	22.13
3	Day 4	22.56
4	Day 5	21.08
5	Day 6	20.73
:	:	:
364	Day 365	25.71

The seasonal analysis of Bangalore weather using temperature data of 2012 has been shown below in Fig. 5. The bar plots shows pre-monsoon, monsoon and post monsoon temperature.

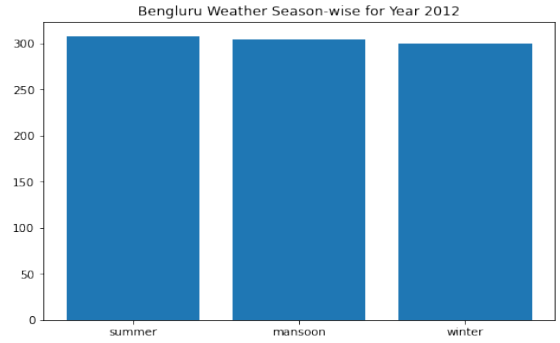


Fig. 5. Seasonal Temperature Analysis of Bangalore for the Year 2012.

The sample first five rows of precipitation dataset (pr_wtr) for the year 2010 is shown in Fig. 6. Time series forecasting is done using ARIMA model. We have worked on ARIMA (2, 0, 2) for rainfall data. Past 11 years rainfall data is trained using the model and the same is used for future prediction. Fig. 7(a-b) shows ARIMA forecasting results. Metrics used for evaluation are Mean error (ME) and Root Mean Squared Error (RMSE). Average error is calculated as shown below in equation (2).

$$\text{Average error} = \text{ME/RMSE} * 100 \quad (2)$$

The accuracy for rainfall data considered in our work using ARIMA (2, 0, 2) model is found to be 76.8.

```
import pandas as pd
from statsmodels.tsa.seasonal import seasonal_decompose
precip = pd.read_csv('/content/5yearsdata.csv', index_col ='Day',parse_dates = True)
precip.head()
```

Day	Unnamed: 0	pr_wtr
2010-01-01	0	34.750000
2010-01-02	1	29.500000
2010-01-03	2	29.619995
2010-01-04	3	27.850006
2010-01-05	4	24.699997

Fig. 6. Sample View of First Five Rows of Precipitation Dataset for the Year 2010.

```
Performing stepwise search to minimize aic
 ARIMA(2,0,2)(0,0,0)[0] intercept   : AIC=9582.453, Time=2.79 sec
 ARIMA(0,0,0)(0,0,0)[0] intercept   : AIC=13179.579, Time=0.10 sec
 ARIMA(1,0,0)(0,0,0)[0] intercept   : AIC=9813.216, Time=0.16 sec
 ARIMA(0,0,1)(0,0,0)[0] intercept   : AIC=11344.685, Time=0.34 sec
 ARIMA(0,0,0)(0,0,0)[0]             : AIC=18092.745, Time=0.03 sec
 ARIMA(1,0,2)(0,0,0)[0] intercept   : AIC=9643.610, Time=0.68 sec
 ARIMA(2,0,1)(0,0,0)[0] intercept   : AIC=9648.076, Time=2.18 sec
 ARIMA(3,0,2)(0,0,0)[0] intercept   : AIC=9584.343, Time=3.06 sec
 ARIMA(2,0,3)(0,0,0)[0] intercept   : AIC=9583.823, Time=3.02 sec
 ARIMA(1,0,1)(0,0,0)[0] intercept   : AIC=9652.861, Time=0.50 sec
 ARIMA(1,0,3)(0,0,0)[0] intercept   : AIC=9626.157, Time=0.85 sec
 ARIMA(3,0,1)(0,0,0)[0] intercept   : AIC=9610.014, Time=2.91 sec
 ARIMA(3,0,3)(0,0,0)[0] intercept   : AIC=9586.439, Time=2.77 sec
 ARIMA(2,0,2)(0,0,0)[0]             : AIC=9593.282, Time=1.18 sec

Best model:  ARIMA(2,0,2)(0,0,0)[0] intercept
Total fit time: 20.583 seconds
9582.45289119955
```

(a)

```
[47] test['pr_wtr'].mean()

     31.687875667230117

[50] from sklearn.metrics import mean_squared_error
     from math import sqrt
     test['pr_wtr'].mean()
     rmse=sqrt(mean_squared_error(pred,test['pr_wtr']))
     print(rmse)

     7.35869865474459
```

(b)

Fig. 7. (a-b): Preview and Results of ARIMA (2 ,0, 2) Model.

V. CONCLUSION

Vast amounts of climate data are being generated rapidly by satellite observations and numerical climate models. Agriculture, tourism, water, electricity, wildfire management, and other sectors are all require climate data. The utility of climatic data depends on timely analysis. Existing technologies, such as Apache Hadoop, which are based on the idea of breaking problems down into smaller chunks and solving them on a cluster of commodity servers, have emerged as a possible solution for analysing huge climate datasets. Apache Spark has recently emerged as a viable alternative to Hadoop's disk-based architecture. The proposed work considers analysis and processing of big spatiotemporal data using Spark MapReduce platform. Multidimensional NetCDF based precipitation and temperature data from NCEP and CSIR-4PI are considered for analysis. Analysis shows that Spark platform is computationally more efficient (double the No. of times) than Hadoop - MapReduce Platform of same configuration. Monthly and seasonal analysis of climate data has been carried out. Time Series prediction approach such as ARIMA (2,0,2) model was used for forecasting future rainfall of Bangalore region, results shows that ARIMA performs well for long term weather prediction. Performance analysis of the model has been carried out using NetCDF data of NCEP and CSIR-4PI Bangalore.

REFERENCES

[1] Z. Li, F. Hu, J. L. Schnase, D. Q. Duffy, T. Lee, M. K. Bowen, and C. Yang. A spatiotemporal indexing approach for efficient processing of big array-based climate data with MapReduce, International Journal of Geographical Information Science, pages 17–35, 2017.

[2] Chaowei Yang, Manzhu Yu, Fei Hu,Yongyao Jiang,Yun Li, Utilizing cloud computing to address big geospatial data challenges, Journal of Computers, Environment and Urban Systems, 2016, http://dx.doi.org/10.1016/j.compenvurbsys.2016.10.010.

[3] James H. Faghmous and Vipin Kumar, A big data guide to understanding climate change: The case for theory-guided data science, Journal of Big Data, Vol. 2, No. 3, Sep 2014, Pages 155–163, https://doi.org/10.1089/big.2014.0026.

[4] Markus Götz, Christian Bodenstein, Matthias Richerzhagen, Gabriele Cavallaro. On Scalable Data Mining Techniques for Earth Science, Procedia Computer Science, December 2015, Volume 51, Pages 2188–2197.

[5] Jinsong Wu, Song Guo, Jie Li,Deze zeng, Big data meet green challenges: Greening big data. IEEE Systems Journal, Volume: 10, Issue: 3, 19 May 2016, Pages 873 – 887.

[6] Ralf Hartmut Guting, M. H. Bohlen, Martin Erwig, Christian S. Jensen, Nikos A. Lorentzos, Markus Schneider, and Michalis Vazirgiannis, A foundation for representing and querying moving objects, ACM Transactions on Database Systems (TODS), Vol. 25, No. 1, March 2000, Pages 1–42.

[7] Sebestyén Viktor, Czvetkó Tímea, Abonyi János, The Applicability of Big Data in Climate Change Research: The Importance of System of Systems Thinking, Frontiers in Environmental Science, Volume 9, March 2021,DOI:10.3389/fenvs.2021.619092.

[8] Yang C., Clarke K., Shekhar S., Tao C.V, Big Spatiotemporal Data Analytics: a research and innovation frontier, International Journal of Geographical Information Science, April 2020, https://doi.org/10.1080/13658816.2019.1698743.

[9] Fei Hu, Chaowei Yang, Daniel Q. Duffy, Michael Bowen, Weiwei Song, Tsengdar Lee, Mengchao Xu and John L. Schnase, ClimateSpark: An in-memory distributed computing framework for big climate data analytics, Journal of Computers and Geosciences, March 2018, Pages 154-166, https://doi.org/10.1016/j.cageo.2018.03.011.

[10] Christopher Bartz, Konstantinos Chasapis, Michael Kuhn, Petra Nerge & Thomas Ludwig, A Best Practice Analysis of HDF 5 and NetCDF- 4 Using Lustre, International Conference on High Performance Computing, ISC High Performance 2015: High Performance Computing, volume 9137, Pages 274-281.

[11] Gowtham Atluri,Anuj Karpatne, Vipin kumar, Spatio-Temporal Data Mining: A Survey of Problems and Methods, ACM Computing Surveys, Volume 51, Issue 4, Article No.: 83, July 2019 Pages 1–41,https://doi.org/10.1145/3161602.

[12] R. Rew, G. Davis, NetCDF: an interface for scientific data access, Volume: 10, Issue: 4, July 1990, pages: 76 – 82, DOI: 10.1109/38.56302.

[13] Salman Salloum, Ruslan Dautov, Xiaojun Chen, Patrick Xiaogang Peng, Joshua Zhexue Huang, Big data analytics on Apache Spark, International Journal of Data Science and Analytics, Springer International Publishing Switzerland 2016, Pages 145-164.

[14] Abdul Salam, Internet of Things for Sustainable Human Health, Book chapter in Internet of Things for Sustainable Community Development. Internet of Things, Springer, January 2020, Pages 217-242, https://doi.org/10.1007/978-3-030-35291-2_7.

[15] Pankaj Mudholkar and Megha Mudholkar, Internet of Things (IoT) and Big Data: A Review, International Journal of Management, Technology and Engineering, Volume 8, Issue XII, December 2018, ISSN NO: 2249-7455, Pages 5001-5007.

[16] Mark McKenney, Niharika Nyalakonda, Jarrod McEvers, Mitchell Shipton, Pyspatiotemporalgeom: A Python Library for Spatiotemporal Types and Operations, Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, October 2016, Article No.: 93, Pages 1–4.

[17] R. Rew, G. Davis, NetCDF: an interface for scientific data access, IEEE Journal of Computer Graphics and Applications, Volume: 10, Issue: 4, July 1990, Pages 76 – 82.

[18] Dimitar Misev, Peter Baumann, Jürgen Seib, Towards Large-Scale Meteorological Data Services: A Case Study, Journal of Datenbank Spektrum- Springer, Volume 21,Issue 1,Pages183–192, 22nd September 2012.

[19] A. Joshi, E. Pebesma, R. Henriques, M. Appel, SCIDB Based Framework For Storage And Analysis Of Remote Sensing Big Data, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLII-5/W3, Capacity Building and Education Outreach in Advance Geospatial Technologies and Land Management,pp.10–11 December 2019, Dhulikhel, Nepal.

[20] Louai Alarabi, Mohamed F. Mokbel, A Demonstration of STHadoop: A MapReduce Framework for Big Spatiotemporal Data, proceedings of the VLDB Endowment, Vol. 10, No. 12, August 2017.

[21] Ahmed Eldawy, Mohamed F. Mokbel, Demonstration of SpatialHadoop: An Efficient MapReduce Framework for Spatial Data, proceedings of the VLDB Endowment, Volume 6, Issue 12, August 2013, Pages 1230-1233, https://doi.org/10.14778/2536274.2536283.

[22] Joe B. Buck, Noah Watkins, Jeff LeFevre, Kleoni Ioannidou, SciHadoop: Array-based query processing in Hadoop, Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis, November 2011, Article No.: 66, Pages 1–11 https://doi.org/10.1145/2063384.2063473.

[23] Timothy Olatayo, A. I. Taiwo, Statistical Modelling and Prediction of Rainfall Time Series Data, Global Journal of Computer Science and Technology: Interdisciplinary, Volume 14, Issue 1 Version1.0, 2014, Online ISSN: 0975-4172 & Print ISSN: 0975-4350.

[24] Nikita Shivhare, Atul Kumar Rahul, Shyam Bihari Dwivedi and Prabhat Kumar Singh Dikshit, ARIMA based daily weather forecasting tool: A case study for Varanasi, Journal Mausam 70(1), January 2019, Pages 133-140.