Computers and Electrical Engineering 000 (2017) 1–15

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Spatial cumulative sum algorithm with big data analytics for climate change detection☆

Gunasekaran Manogaran*, Daphne Lopez

School of Information Technology and Engineering, VIT University, Vellore 632014, Tamil Nadu, India

ARTICLE INFO

Article history:
Received 9 January 2017
Revised 6 April 2017
Accepted 6 April 2017
Available online xxx

Keywords:
Hadoop Distributed File System
Big data
Climate change
Data analytics
Weather sensor data

ABSTRACT

Big data plays a vital role in the prediction of diseases that occur due to climate change. For such predictions, scalable data storage platforms and efficient change detection algorithms are required to monitor the climate change. However, traditional data storage techniques and algorithms are not applicable to process the huge amount of climate data. This paper presents a scalable data processing framework with a novel change detection algorithm. The large volume of climate data is stored on Hadoop Distributed File System (HDFS) and MapReduce algorithm is applied to calculate the seasonal average of climate parameters. Spatial autocorrelation based climate change detection algorithm is proposed in this paper to monitor the changes in the seasonal climate. The proposed climate change detection algorithm is compared with various existing approaches such as pruned exact linear time method, binary segmentation method, and segment neighborhood method.

1. Introduction

"Big Data" is defined by volume, velocity, and variety of data. Big data is very complex to process by traditional data processing techniques and tools. Nowadays, data generation sources like telescopes, satellite, sensor networks, social networks, wearable devices, mobile devices, streaming machines and high throughput instruments are continuously generating a large volume of data. Recently, big data analytics has been applied in various domains, such as healthcare, business process, scientific research, natural resource management, share marketing, social networking, community administration and climate modeling. Climate data is observed from various advanced sensor technologies and is used to represent the seasonal changes. Weather data collected from different climate laboratory and advanced computing technologies are used to give valuable information to the world. Meteorological data is most often used to predict the weather and other climate-related phenomena. In addition, climate data is also used for various purposes that lead to a significant development in weather forecasting, rocket launching, and public health.

However, climate data collected from various sources are used to identify the seasonal changes. In general, the climate laboratories generate data in unstructured format. Statistical techniques or machine learning algorithms are used to get meaningful information from the raw data. For example, statistical techniques are used to identify the number of precipitation days for a specific region. In this regard, the World Climate Data Monitoring Programme (WCDMP) is developed by WMO's World Climate Programme (WCP) that focuses on management and collection of large climate data observed from the global climate system [1]. Researchers and officials from the climate department use term Climate "normals" to compare

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Associate Editor Dr. R. Varatharajan.

* Corresponding author.
E-mail addresses: gunavit@gmail.com (G. Manogaran), daphnelopez@vit.ac.in (D. Lopez).

http://dx.doi.org/10.1016/j.compeleceng.2017.04.006
0045-7906/

with other or past climate conditions. In general, normal climate is calculated using an average of the climate parameter (e.g. maximum temperature) over a period. To check whether the day-to-day climate is normal or not, everyday climate is compared with the past climate period from 1st January 1961 to 31st December 1990.

World Weather Records (WWR) is originally developed by the world climate organization in 1923. The primary goal of World Weather Records (WWR) is to maintain the huge size of records such as monthly temperature, wind speed, rainfall, precipitation and pressure data that are collected from thousands of weather stations around the globe. In recent years, the number of stations is increased noticeably. Especially, many weather stations have been collecting metrological data in a continuous manner. World Metrological Organization (WMO) has been collecting the day-to-day weather data in the form of digital since 1920. Metrological data collected from WMO are digitally published by nine issues. It include 1921–30, 1931–40, 1941–50, 1951–60, 1961–70, 1971–80, 1981–90, and 1991–2000 [2]. World Meteorological Organization (WMO) Commission for Climatology (CCl) maintains NoSQL based database to store the massive amount of data related to world weather extremes and various abnormal conditions. The Commission for Climatology (CCl) published the world weather extreme data that are available online to users. The essential role of this database is to maintain the huge extreme of various climate parameters respect to space and time. The database consists of following climate parameters it includes maximum/minimum observed temperature, wind speed and most precipitation on earth respect to space and time. In addition, the CGI database also maintains the most destructive earthquake, hurricanes, floods, storms, and tornadoes.

World Meteorological Organization uses space temperature on the land to measure the world surface temperature. Ships and buoys are used to measure the sea surface temperature whereas rain gauge and precipitation are measured with the help of satellites. The US Climate Prediction Centre, US National Climatic Data Centre and Global Precipitation Climatology Project (GPCP) are maintaining the database to store the precipitation data for the globe. Recently, researchers from bioinformatics and climate modeling are identified the correlation between the disease and climate with help of big data analytics. Merging of climate and health data has become a major role in big data analytics. This paper uses MapReduce framework to process the huge climate data and to find the seasonal changes.

The structure of the manuscript is explained as follows: Section 1 introduces the need for big data analytics in a climate data processing system. Section 2 reviews the recent work in big climate data analytics. Section 3 discusses the proposed framework for seasonal climate changes. Section 4 discusses the proposed algorithm for seasonal climate change detection. Comparison of various change detection algorithms is discussed in Section 5. Result and discussion, performance evaluation are discussed in Section 6. Finally, Section 7 concludes the work.

2. Related work

Global climate change is considered as one of the top challenges of 21st century. Many researchers are doing their research in big data analytics based climate modeling and bioinformatics [3,4]. Data science is a major field to understand the global climate modeling. Thus, big data plays a vital role in almost every domain [5]. For example, the United States Environmental Protection Agency has identified the software to store and share the climate data [6]. This software is used to monitor the present climate change and future possible changes in the seasonal climate data. In addition, Enviratlas also used to find the relationship and impact of climate change in society, healthcare and ecosystems [7]. Similarly, NASA also developed the Climate Analytics-as-a-Service (CAaaS) in cloud computing to enable the best performance in following domains such as data proximal analytics, scalable data administration (big data), software machine virtualization, high-performance computing, adaptive analytics and synchronized API development [8]. In addition, researchers from Brazil has developed Platform as a Service (PaaS) based cloud computing platform called EUBrazilCC project to store and analyze big climate data [9].

Big data also plays a vital role in various healthcare applications. Data generation sources in healthcare domain are increased, and it requires advanced big data tools and techniques to process such huge volume of data. It is observed that various improvements have been made in the day-to-day clinical system. This advancement is used to develop knowledge from huge clinical records and improve business insights. Recently, many research work has been done to reduce the overall cost and improve the disease diagnosis in healthcare [10–13]. Moreover, the impact of big data and cloud computing has been increased noticeably [14–18]. In addition, there is a need to provide security and privacy in healthcare big data analytics. Katagi and Moriai have reviewed state-of-the-art cryptography algorithms in Internet of Things [19]. Bi et al. have discussed the benefit of the DPA resistant circuits to provide low power consumption and high security [20]. Similarly, Lian has discussed the algorithms and applications in multimedia encryption [21]. Bi et al. have reviewed the applications of transistor technologies in hardware security [22,23].

Bates et al. have developed six use cases to reduce the overall cost of healthcare [24]. These use cases are applied to the following domains to reduce the cost for patients, health record management, triage, readmissions, disease diagnosis, drug recommendation and healing optimization. Hermon and Williams have identified four uses cases of big data in healthcare it includes patient administration and healthcare delivery, medical decision support system, medical support services and customer behavior [25]. Moreover, various big data analytical solutions are developed to reduce the cost of treatment path, drug recommendation and healthcare delivery. Recently, researchers from bioinformatics and climate modeling have identified the correlation between the disease and climate with the help of big data analytics. Merging of climate and health data becomes a significant role in big data analytics. Many researchers from University of California, Johns Hopkins University, and IBM have identified the various platforms and software to model the climate change and dengue. Research and

Please cite this article as: G. Manogaran, D. Lopez, Spatial cumulative sum algorithm with big data analytics for climate change detection, Computers and Electrical Engineering (2017), http://dx.doi.org/10.1016/j.compeleceng.2017.04.006

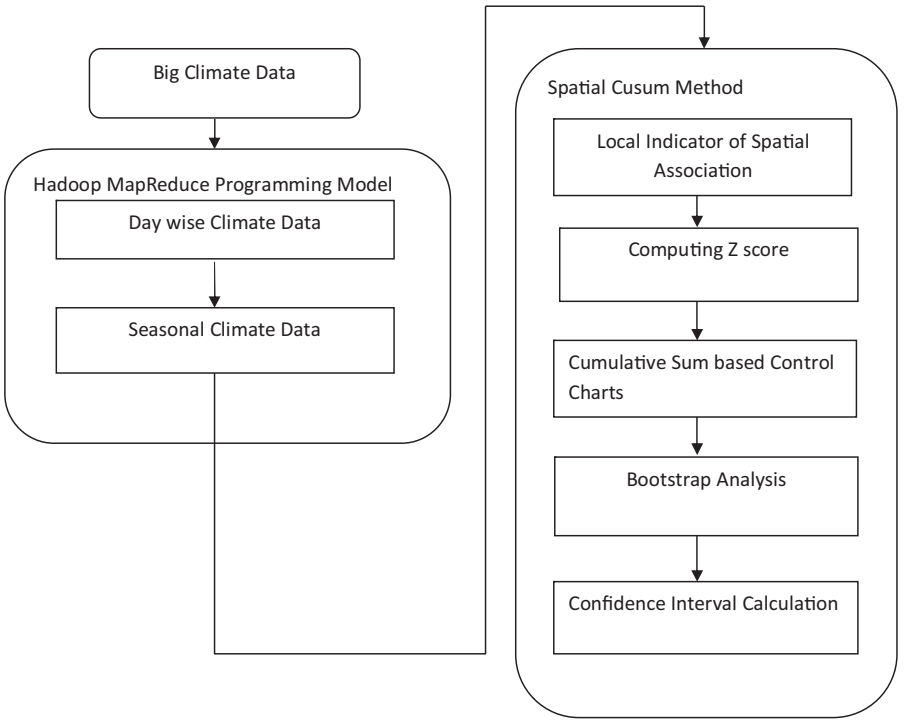


Fig. 1. Proposed framework for climate change detection.

```
Date,Longitude,Latitude,Elevation,MaxTemperature,MinTemperature,Precipitation,Wind,Relative
Humidity,Solar
1/1/1979,79.0625,13.8942003250122,409,24.001,17.365,0.3278742264,1.566199145656,0.8886190936,12.30403348,
1/2/1979,79.0625,13.8942003250122,409,25.817,14.111,0.0102996792,1.835927473671,0.873943504333,19.493721,
1/3/1979,79.0625,13.8942003250122,409,25.231,12.991,0,1.90932099044504,0.850141727895694,20.09739132,
1/4/1979,79.0625,13.8942003250122,409,25.162,12.907,0,1.86838639889342,0.835131190267102,20.104311348,
1/5/1979,79.0625,13.8942003250122,409,25.908,11.847,0.0781059168,1.6394340241707,0.802525084434,20.14002,
1/6/1979,79.0625,13.8942003250122,409,25.51,13.303,0.1441958544,1.87781918316264,0.834196067495133,18.93,
1/7/1979,79.0625,13.8942003250122,409,26.117,12.981,0,1.6757060896263,0.795431155405995,20.292809688,
1/8/1979,79.0625,13.8942003250122,409,26.701,14.226,0,1.49576174732698,0.766072592823415,20.412171432,
1/9/1979,79.0625,13.8942003250122,409,27.036,15.321,0,1.86507940597598,0.740040020348654,20.148522372,
1/10/1979,79.0625,13.8942003250122,409,25.813,13.84,0,2.09980346349538,0.808438424639 49,20.31598458,
1/11/1979,79.0625,13.8942003250122,409,25.215,12.002,0,1.99474999541674,0.794919321398066,20.483435052,
1/12/1979,79.0625,13.8942003250122,409,25.506,10.974,0,1.594475807156,0.790534103060224,20.69099928,
1/13/1979,79.0625,13.8942003250122,409,26.329,14.381,0,1.71615448541087,0.76192037123461,20.681832168,
1/14/1979,79.0625,13.8942003250122,409,26.614,12.775,0,1.99117412616105,0.767678930018668,20.69740926,
1/15/1979,79.0625,13.8942003250122,409,26.903,13.428,0,2.10653682505959,0.723011561155546,20.887990308,
1/16/1979,79.0625,13.8942003250122,409,27.077,13.008,0,1.6733786779527,0.767045136865706,20.72435904,
1/17/1979,79.0625,13.8942003250122,409,27.086,13.333,0,1.82442404034068,0.808374437532221,20.569895568,
```

Fig. 2. Weather station data.

development team from IBM have used mathematical models for vector-borne disease and Spatio Temporal Epidemiological Modeler (STEM) tool with big data to identify the diseases like dengue and malaria based on the climate change. Pfeiffer et al. have identified the Spatio-temporal model based on the big data analytics to predict animal and human health risks.

3. Proposed framework

The proposed framework for climate change detection and raw weather station data are shown in Figs. 1 and 2 respectively. The big climate data is reduced with the help of Hadoop MapReduce framework spatial cumulative sum algorithm is proposed to monitor the seasonal changes in the climate data. MapReduce algorithm is used to create a table in Apache HBase with the help of Apache Hive. The huge day wise climate data from 1979 to 2016 is reduced to seasonal data with the help of Apache MapRedcue framework. Fig. 3 shows the original weather data generated from various weather stations in the study area Tamil Nadu, India. Results generated from the MapReduce framework is shown in Figs. 4 and 5.

Please cite this article as: G. Manogaran, D. Lopez, Spatial cumulative sum algorithm with big data analytics for climate change detection, Computers and Electrical Engineering (2017), http://dx.doi.org/10.1016/j.compeleceng.2017.04.006

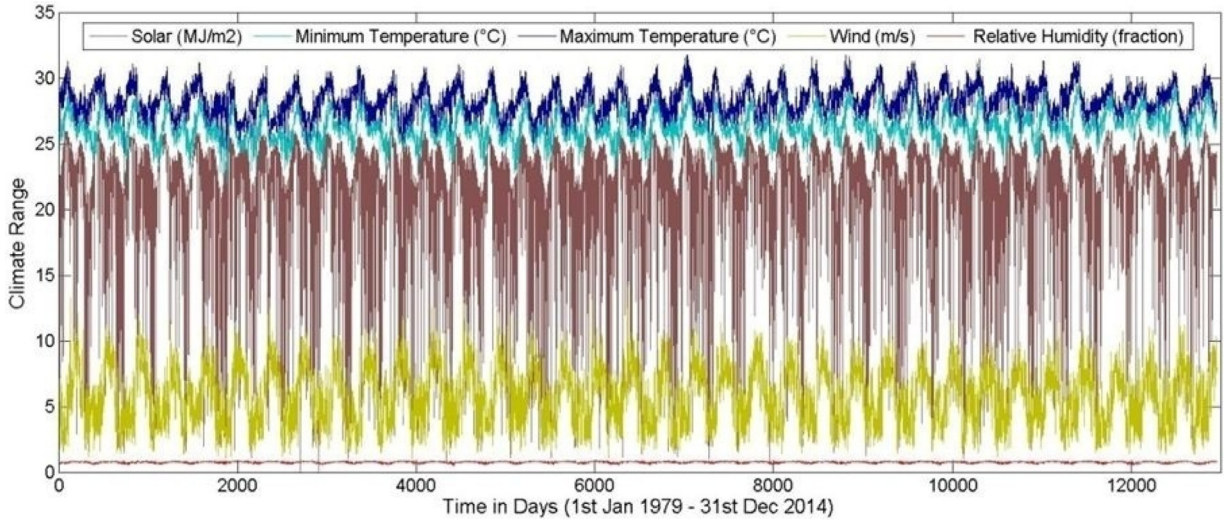


Fig. 3. Daily climate range.

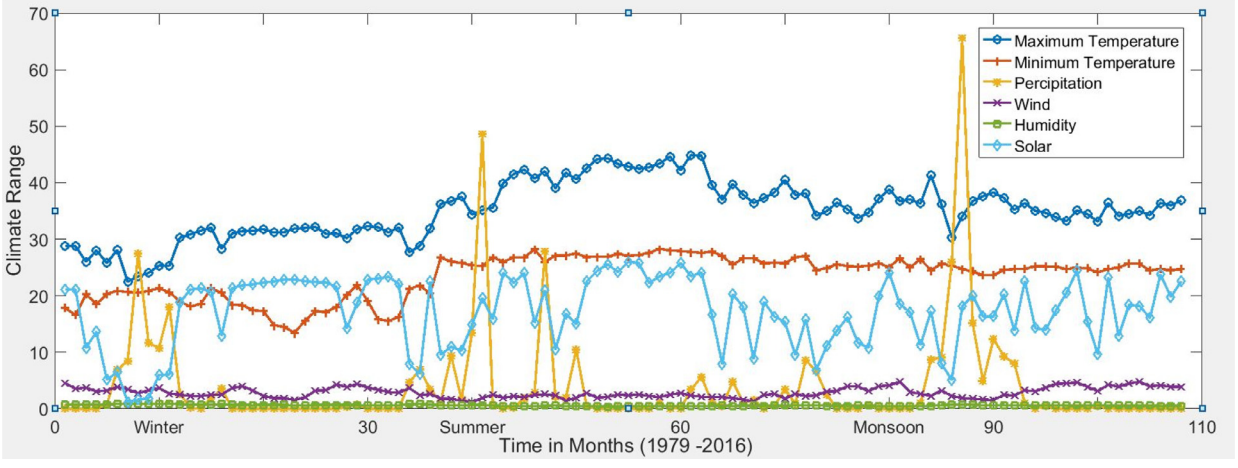


Fig. 4. Seasonal mean climate data (1979–2016).

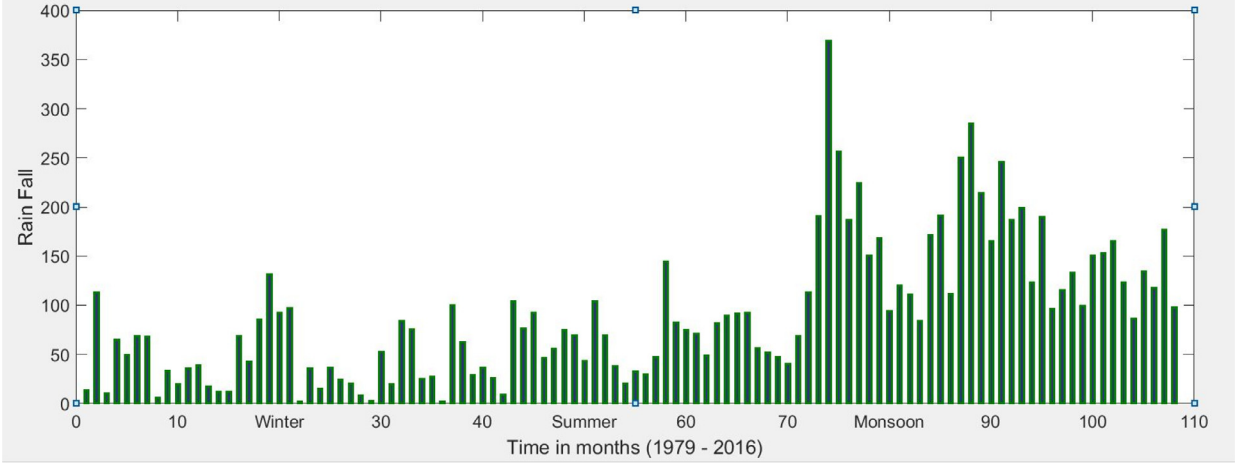


Fig. 5. Rain fall (1979–2016).

Please cite this article as: G. Manogaran, D. Lopez, Spatial cumulative sum algorithm with big data analytics for climate change detection, Computers and Electrical Engineering (2017), http://dx.doi.org/10.1016/j.compeleceng.2017.04.006

Algorithm 1: *Hive Algorithm to Create Table in HDFS.*

1. **Input:** Field name and data type of weather data
2. **Output:** Weather table
3. **if** (Weather_Table_Name ≠ NULL) **or** (Weather_Table _Field ≠ NULL)
4. **then**
5. **for each** column in Weather_Table _Field ∈ Day_wise_ Weather_Table
6. **do**
7. **add** field name and field data type < date Date, latitude Double, longitude Double, elevation Double, maximum_temperature Double, minimum_temperature Double, precipitation Double, wind Double, relative humidity Double, solar Double>
8. **for each** row in Weather_Table_Fields ∈ Day_wise_ Weather_Table
9. **do**
10. Table_Fields **terminated by** ','
11. **Return** Hive Create Table query

Algorithm 2: *Hive Algorithm to Store Big Weather Data from Local File System to HDFS.*

1. **Data:** Weather data collected from multiple weather stations
2. **Input:** Weather data 'climatedata.txt' in Local File System
3. **Output:** Day_wise_Climate_Table in HDFS
4. **if** (Table_Name ≠ NULL)
5. **then**
6. **load data local** inpath '${env:home}/climatedata.txt'
7. **into table** Day_wise_Climate_Table;
8. **Return** Hive Load Table query

Algorithm 3: *MapReduce Algorithm to calculate the seasonal average of weather parameters from day wise weather data stored in HDFS.*

1. **Data:** Weather data collected from multiple weather stations
2. **Input:** Weather data 'climatedata.txt' from HDFS
3. **Output:** Seasonal_Average_Weather_Parameters to HDFS
4. **Function: Mapper**
5. **method** INITIALIZE
6. Sum = new ASSOCIATIVE ARRAY
7. Count = new ASSOCIATIVE ARRAY
8. **method** MAP⟨String ncp, double cpv⟩
9. #nwp=Weather parameter name
10. #wpv=Weather parameter value(day wise)
11. Sum{nwp} = Sum{nwp} + wpv
12. Count{nwp} = Sum{nwp} + 1
13. **method** CLOSE
14. for all term ncp ε Sum do
15. Emit Intermediate(string nwp, pair(Sum{nwp}, Count{nwp}))
16. **Function: Reducer**
17. **method** REDUCE(String ncp, pairs[(sum_1, $count_1$)...])
18. double final_sum = 0.0;
19. **method** REPEAT
20. for all pair (Sum, Count) ε pairs[(sum_1, $count_1$)to (sum_{122}, $count_{122}$)] do
21. final_sum+ = sum;
22. double seasonal_mean_wpv = final_sum/122;
23. return(String nwp, double seasonal_mean_wpv);

4. Proposed algorithm for seasonal climate change detection

This section explains the existing approaches for climate change detection, such as cumulative sum method and bootstrap analysis method.

4.1. Cumulative sum method

Cumulative sum methods are used to detect the slow and drastic changes in the mean value of a quantity of interest. This method is used in many application it includes monitor the changes in the production environment, disease expectation, fish populations, deforestation and crime analysis. This paper uses cumulative sum method to monitor the changes in the climate. Taylor has developed the change point analysis method with the help of cumulative sum charts (CUSUM) and bootstrapping analysis methods. CUSUM control chart can be calculated as follows:

Calculate the average for 'n' data points X_1, X_2, ..., X_n, by the following equation:

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} \tag{1}$$

Please cite this article as: G. Manogaran, D. Lopez, Spatial cumulative sum algorithm with big data analytics for climate change detection, Computers and Electrical Engineering (2017), http://dx.doi.org/10.1016/j.compeleceng.2017.04.006

The cumulative sum value S_I is calculated based on the following equation:

$$S_I = S_{i-1} + \left(X_i - \bar{X}\right) \; for \; i = 1, 2, \ldots, \; n \tag{2}$$

Where,

$S_0 = 0$

Calculate the maximum and minimum S_{max} and S_{min}, by the following equation

$$S_{max} = \max_{i=0,1,\ldots,n} S_i \tag{3}$$

$$S_{min} = \min_{i=0,1,\ldots,n} S_i \tag{4}$$

Calculate the S_{diff} Values to find the changes in cumulative sum value S_I, by the following equation:

$$S_{diff} = S_{max} - S_{min} \tag{5}$$

4.2. Bootstrap analysis method

Cumulative sum method is used to find noticeable shift or change in the average. However, the bootstrap analysis is an additional metric used to determine the significant changes calculated by randomly reordering the original 'n' values. This bootstrap analysis is used to verify the changes initially calculated from the cumulative sum value S_I. The confidence level is calculated for the bootstrap analysis results. A single bootstrap consists of the following steps:

- **Step 1:** Reorder the original 'n' values X^0_1, X^0_2, ..., X^0_n using sampling without replacement method.
- **Step 2:** Calculate the cumulative sum values S^0_0, S^0_1, ..., S^0_n.
- **Step 3:** Calculate the maximum, minimum and difference S^0_{max}, S^0_{min}, and S^0_{diff} values.
- **Step 4:** Identify the bootstrap difference S^0_{diff} is less than the original difference S_{diff}.

The vital role of the bootstrap analysis is to identify the significant level of the cumulative sum results if no change has occurred. Number of bootstrap analyses is performed to calculate the confidence level of the changes

Confidence level CI can be calculated, by the following equation

$$Confidence \; Level \; (CI) = 100 * \frac{X}{N}\% \tag{6}$$

Where,

N = Number of bootstrap samples performed
X = Number of bootstraps for which $S^0_{diff} < S_{diff}$

In general, 90%, or 95% confidence level are required to identify the significant change in the original data. In addition, it is not possible to perform bootstrap analysis for n!. However, 1000 bootstrap analysis is sufficient to find the significant changes in the original data.

4.3. Proposed change detection method

This paper uses spatial autocorrelation based cumulative sum algorithm to monitor the climate change.

4.3.1. Spatial autocorrelation

Spatial Autocorrelation is used to find the correlation of the variables over space. Spatial statistics group identified many approaches to measure the spatial autocorrelation between the variables, such as Moran's I, Geary's C, Getis' G and the Join Count Analysis. This study uses local Moran's I spatial auto correlation to find the correlation between the climate parameter with itself over space. Local Moran's I spatial auto correlation can be calculated, by the following equation:

$$Z \, \text{score} \, z_i = \frac{x_i - \bar{x}}{SD} \tag{7}$$

$$\text{Local Moran's} \, I_i = z_i \sum_j w_{ij} z_j \tag{8}$$

Where,

n = Number of spatial locations indexed by i and j
x_i = Variables

$\bar{x}$ = Mean of x

$$SD = \sqrt{\frac{\sum \left(x - \bar{\bar{x}}\right)^2}{n}} \tag{9}$$

w_{ij}= Spatial standardized weight matrix

Algorithm 4: *Proposed Algorithm for climate change detection.*

1. ***Input:*** *Seasonal weather data*
2. ***Output:*** *Seasonal changes*
3. ***Let*** C_1, C, …, C_i *represent the 'i' data points*
4. ***Calculate*** *the Standard Deviation for 'i' data point* $SD = \sqrt{\frac{\sum (C - \bar{\bar{C}})^2}{n}}$
5. ***Calculate*** *Z score Value* $z_i = \frac{C_i - C}{SD}$
6. ***Calculate*** *Spatial Local Moran's I Value* $I_i = z_i \sum_j w_{ij} z_j$
7. ***Calculate*** *the average* $\bar{I} = \frac{I_1 + I_2 + \ldots + I_n}{n}$
8. ***Initialize*** *cumulative sum* $S_0 = 0$
9. ***Calculate*** *the other cumulative sums* $S_I = S_{i-1} + (I_i - \bar{I})$ ***for*** $i = 1, 2, \ldots, n$
10. ***Calculate*** $S_{max} = \max_{i=0,1,\ldots,n} S_i$
11. ***Calculate*** $S_{min} = \min_{i=0,1,\ldots,n} S_i$
12. ***Calculate*** $S_{diff} = S_{max} - S_{min}$
13. ***Sampling without replacement:*** ***Generate*** *a bootstrap sample of 'n' units, denoted* $X^0{}_1$, $X^0{}_2$, …, $X^\circ{}_n$, *by randomly reordering the original 'n' values*
14. ***Calculate*** *the bootstrap CUSUM* $S^0{}_i$ ***for*** $i = 1, 2, \ldots, n$ ($S^0{}_0$, $S^0{}_1$, …, $S^0{}_n$)
15. ***Calculate*** *the maximum, minimum and difference of the bootstrap CUSUM, denoted* $S^0{}_{max}$, $S^0{}_{min}$, *and* $S^0{}_{diff}$.
16. ***Determine*** *whether the bootstrap difference* $S^0{}_{diff}$ *is less than the original difference* S_{diff}.
17. ***Let*** *N be the number of bootstrap samples performed and let X be the number of bootstraps for which* $S^0{}_{diff} < S_{diff}$
18. ***Calculate*** *Confidence Level* $CI = 100 * \frac{X}{N}\%$

5. Comparison of various change detection algorithms

The proposed spatial CUSUM based change detection algorithm is compared with various existing change detection approaches, such as Pruned Exact Linear Time (PELT), binary segmentation and segment neighborhood method.

5.1. PELT algorithm

The Pruned Exact Linear Time (PELT) method is used to detect multiple changes in the large datasets. The proposed algorithm is tested with larger regions of the genomes. The proposed approach primarily used to minimize the cost function over possible numbers and locations of change points. PELT method uses a new method to find the minimum of such cost functions. Hence, a result generated from the PELT method has optimal number and location of change points.

5.2. Binary segmentation algorithm

The binary segmentation algorithm is used to analyze the variance in the multiple homogeneous groups. The proposed algorithm uses multiple comparison methods for analyzing various groups. A likelihood ratio test is also proposed in this algorithm to identify the differences among the resulting groups.

5.3. Segment neighborhoods

The segment neighborhood method is used to estimate the parameters of the model that describe the boundaries of each segment neighborhood. The proposed algorithm also efficiently used in the least squares and maximum likelihood estimation. This algorithm is effectively used to model the changes in haemagglutinin protein of influenza virus.

6. Result and discussion

This paper uses Cumulative sum control charts to find the differences of each sample value from the target value. CUSUM control charts are also called as time-weighted control chart used to monitor the small shifts in the mean of a process. The traditional CUSUM control chart is used to monitor the changes in rainfall, precipitation, maximum temperature, minimum temperature, humidity, wind speed and solar and the results are shown in Figs. 6 to 12 respectively. The cumulative sum is not the cumulative sum of the values. Instead, it is the cumulative sum of differences between the values and the average (target value). Because the mean is subtracted from each value, the cumulative sum also ends at zero. A CUSUM control chart is used to identify the cumulative sums (CUSUMs) of the deviations of each sample value from the target value. Moreover, small drifting in the mean vale will lead to steadily increasing or decreasing cumulative deviation values. However,

Please cite this article as: G. Manogaran, D. Lopez, Spatial cumulative sum algorithm with big data analytics for climate change detection, Computers and Electrical Engineering (2017), http://dx.doi.org/10.1016/j.compeleceng.2017.04.006

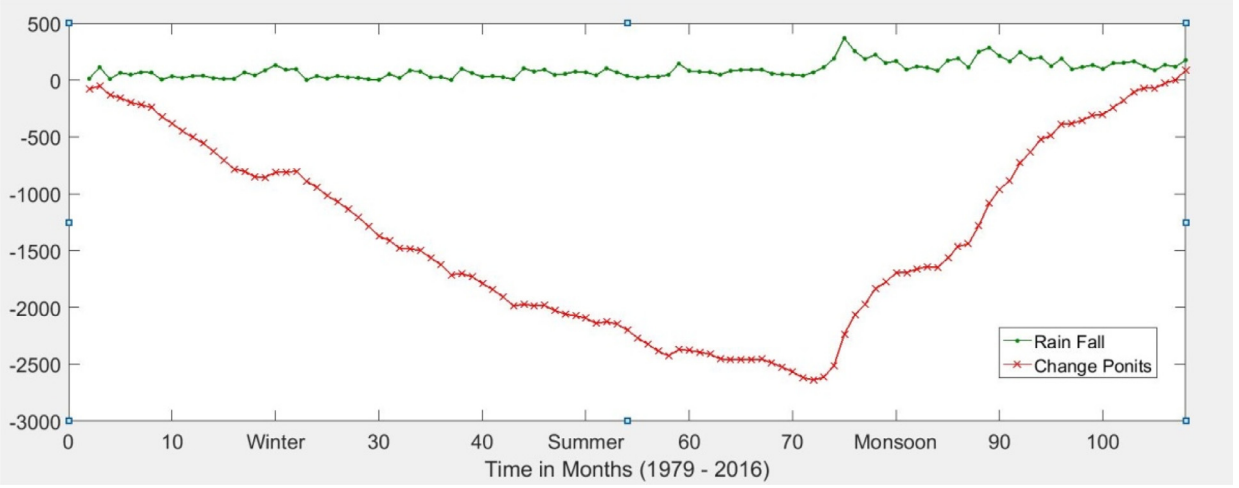


Fig. 6. Cusum control chart for rain fall ('X' axis represents the time in months (1979–2016) and 'Y' axis represents the rain fall change points).

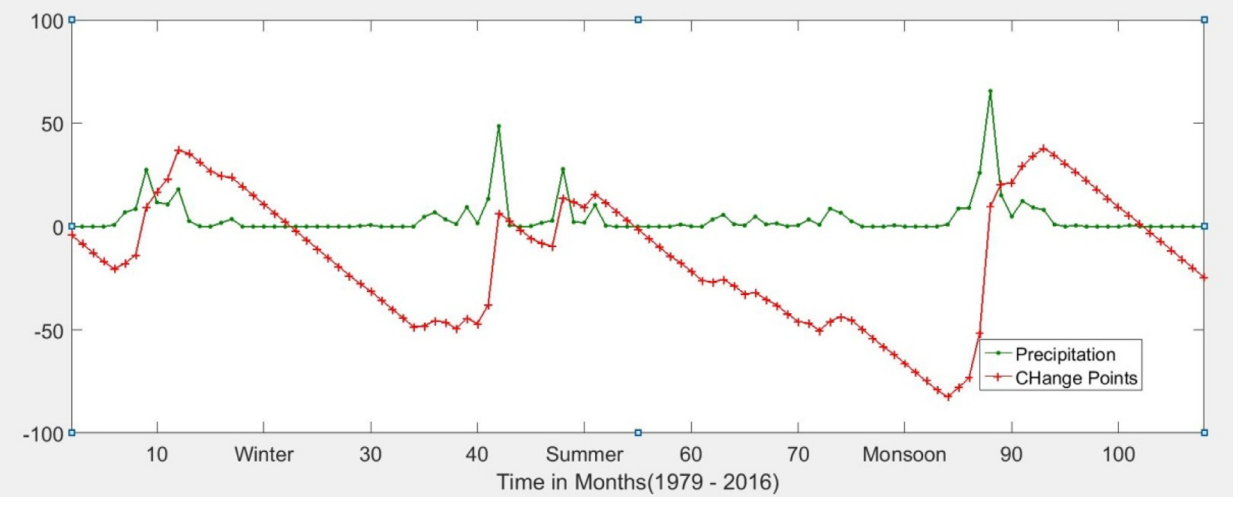


Fig. 7. Cusum control chart for precipitation ('X' axis represents the time in months (1979–2016) and 'Y' axis represents the precipitation change points).

interpreting the changes identified by the control charts can still be difficult. In order to analyze these changes, a change-point analysis can be performed. The CUSUMs are also named as change points. Fig. 6 represents the CUSUM control chart for identifying the CUSUMs of the deviations of rainfall from the target value.

As shown in Fig. 6, the rainfall CUSUMs in summer is more deviations from the target value. Fig. 7 represents the CUSUMs for precipitation. The change points for precipitation are varied based on the seasons. Figs. 8–10 represent the change points for minimum temperature, maximum temperature, and humidity respectively. More deviations in minimum temperature, maximum temperature, and humidity are identified during the summer season. Figs. 11 and 12 represents the change points for wind speed and solar respectively. As shown in Figs. 11 and 12, more deviations are occurred in wind speed and solar during the monsoon and winter season respectively. Various significant climate changes are identified with the help of Pruned Exact Linear Time (PELT) method, binary segmentation method, segment neighborhood method, and the results are depicted in Tables 1 –6. Fig. 13 shows the original changes in various climate parameters during 1979–2016. The original change is identified with the help of a mixture of Pruned Exact Linear Time (PELT) method, binary segmentation method, and segment neighborhood method. Table 1 represents the significant changes in the rainfall for the year 1979–2016. As shown in Table 1, the major change in rain fall is identified during summer 2014 with +256 mm difference. Another significant change in rainfall is identified during monsoon 1984 with −249 mm difference. Table 2 represents the significant changes in the maximum temperature for the year 1979–2016. Table 2 depicts the maximum temperature change during summer 2004 with + 7.662 °C.

In addition, the second significant change during summer 2014 with + 6.382 °Celsius. Table 3 represents the significant changes in the minimum temperature for the year 1979–2016. The most significant change in the minimum temperature is

Please cite this article as: G. Manogaran, D. Lopez, Spatial cumulative sum algorithm with big data analytics for climate change detection, Computers and Electrical Engineering (2017), http://dx.doi.org/10.1016/j.compeleceng.2017.04.006

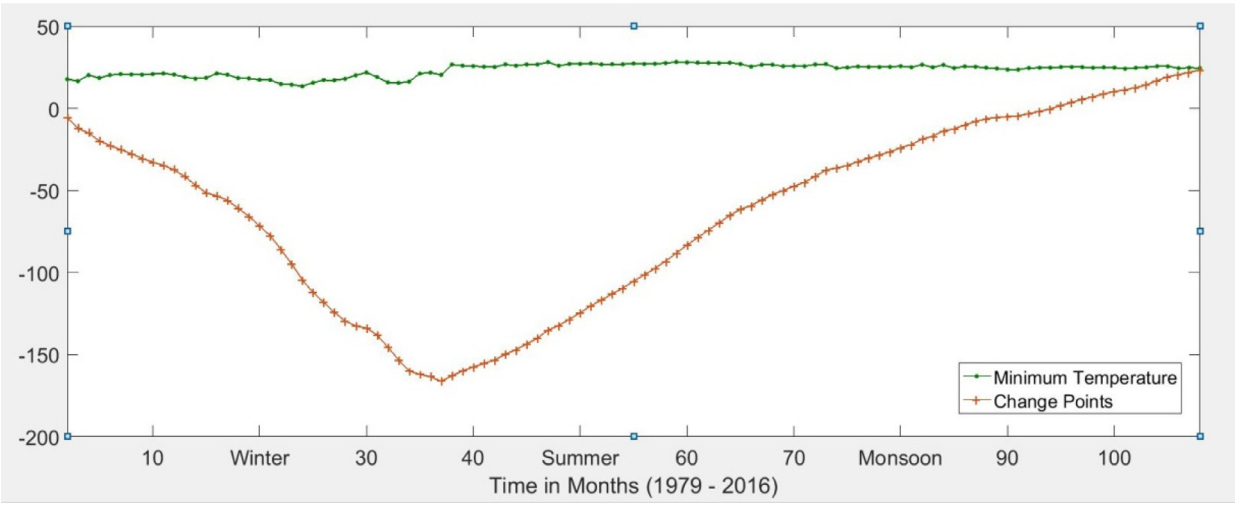


Fig. 8. Cusum control chart for minimum temperature ('X' axis represents the time in months (1979–2016) and 'Y' axis represents the minimum temperature change points).

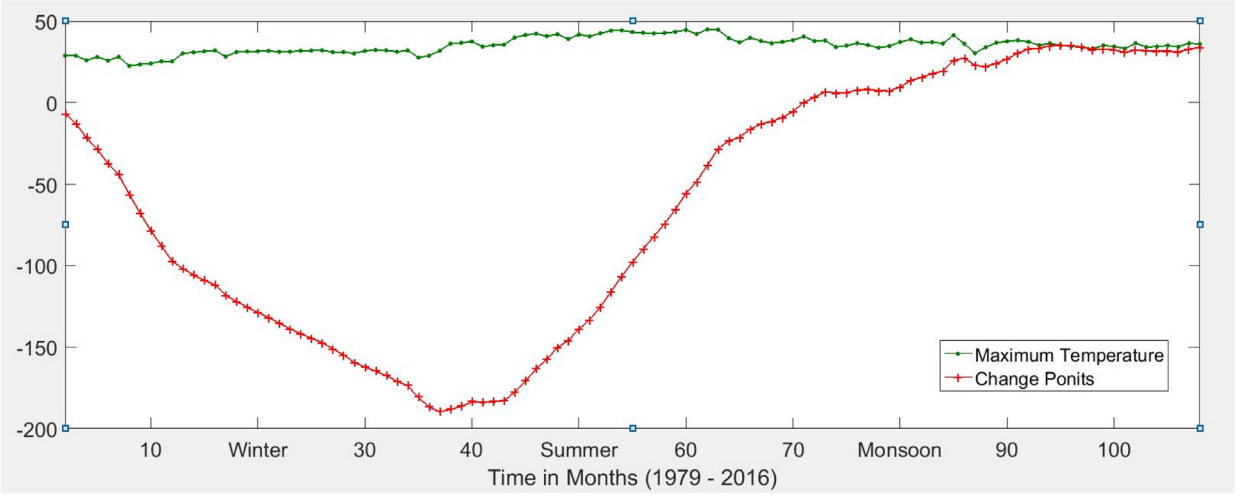


Fig. 9. Cusum control chart for maximum temperature ('X' axis represents the time in months (1979–2016) and 'Y' axis represents the maximum temperature change points).

Table 1
Rain fall changes.

Year	Change point	Rain Fall		Rain Fall change	Level of change
		From	To		
2012 Summer	72	89	112	+23	5
2014 Summer	74	112	368	+256	1
1984 Monsoon	81	368	119	−249	2
1991 Monsoon	88	119	249	−130	4
1996 Monsoon	93	249	118	+131	3

−5.572 °C, +4.4933 during summer 1981 and winter 1996 respectively (Table 3). Table 4 represents the significant changes in the precipitation for the year 1979–2016. The results depict the most significant change in precipitation during monsoon 1989, 1991 with −44 mm and +39 mm respectively. Table 5 represents the significant changes in the solar for the year 1979–2016. The most significant change in solar is identified with the value −18.204 during winter 1991. Table 6 represents the significant changes in the wind speed for the year 1979–2016. The maximum change in the wind speed is identified during summer 2005 with −1.259 m/s.

Please cite this article as: G. Manogaran, D. Lopez, Spatial cumulative sum algorithm with big data analytics for climate change detection, Computers and Electrical Engineering (2017), http://dx.doi.org/10.1016/j.compeleceng.2017.04.006

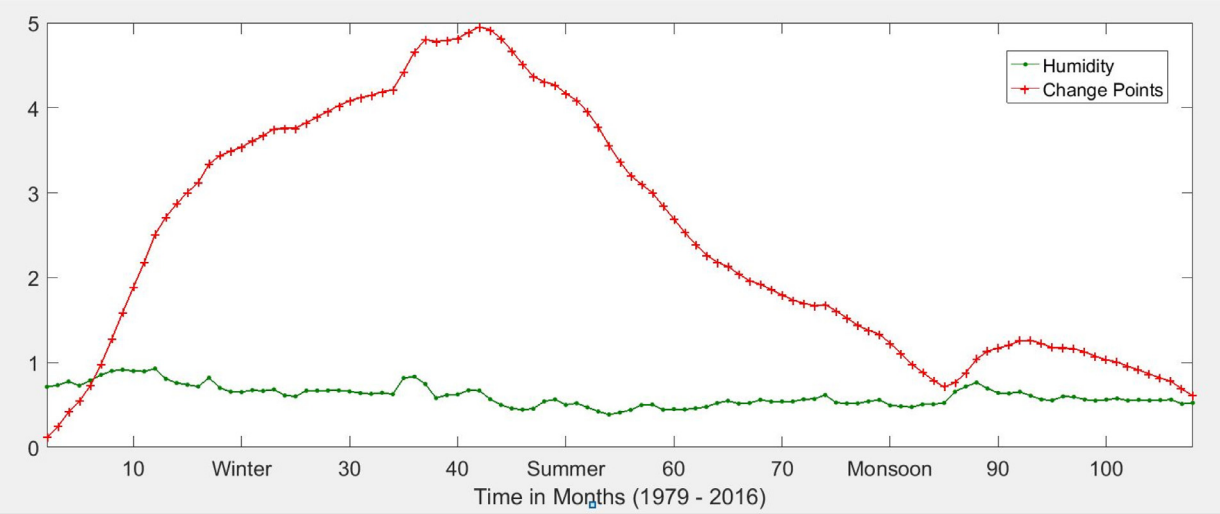


Fig. 10. Cusum control chart for humidity ('X' axis represents the time in months (1979–2016) and 'Y' axis represents the humidity change points).

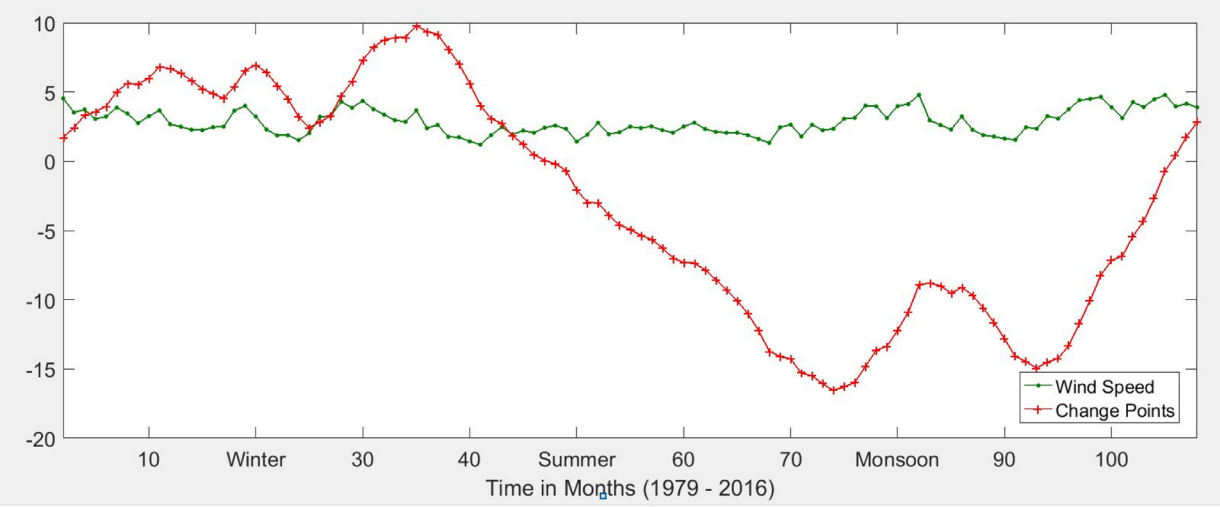


Fig. 11. Cusum control chart for wind speed ('X' axis represents the time in months (1979–2016) and 'Y' axis represents the wind speed change points).

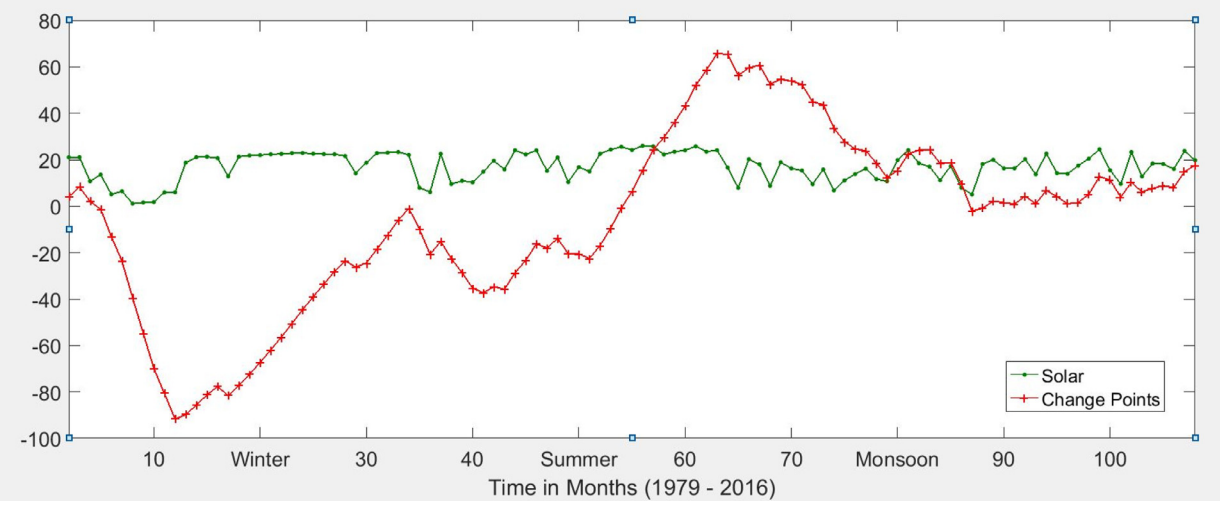
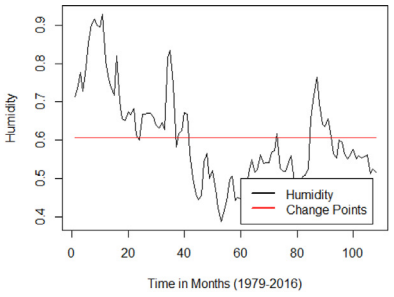
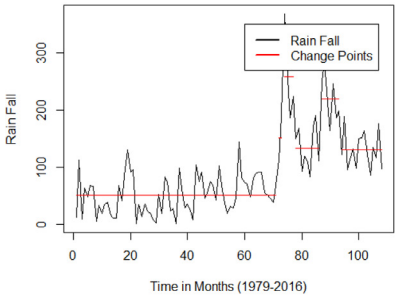


Fig. 12. Cusum control chart for solar ('X' axis represents the time in months (1979–2016) and 'Y' axis represents the solar change points).

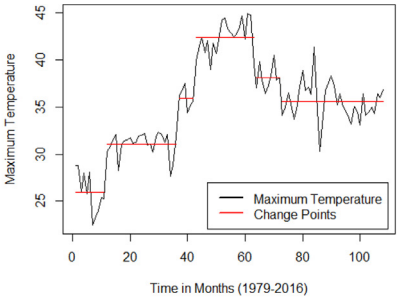
Please cite this article as: G. Manogaran, D. Lopez, Spatial cumulative sum algorithm with big data analytics for climate change detection, Computers and Electrical Engineering (2017), http://dx.doi.org/10.1016/j.compeleceng.2017.04.006



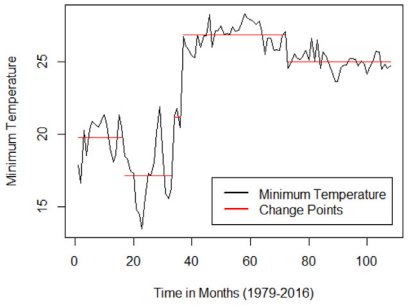
(a) Humidity Changes in fraction



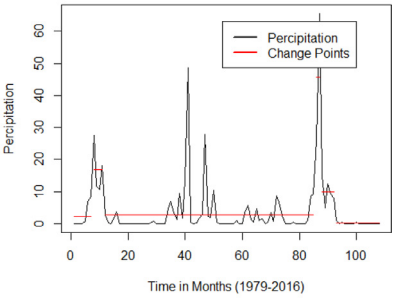
(b) Rain Fall Changes in mm



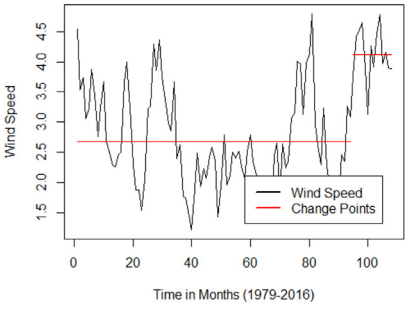
(c) Minimum Temperature Changes in ‘c



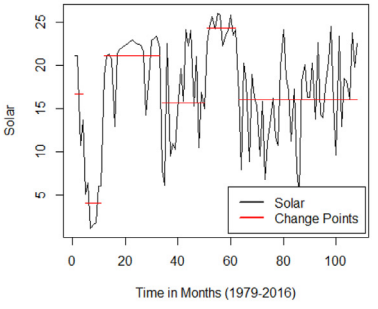
(d) Minimum Temperature Changes in ‘c



(e) Precipitation Changes in mm



(f) Wind Speed Changes in m/s



(g) Solar Changes in MJ/m2

Fig. 13. Climate change points detection. (a) Humidity changes in fraction. (b) Rain fall changes in mm. (c) Minimum temperature changes in ‘c. (d) Minimum temperature changes in ‘c. (e) Precipitation changes in mm. (f) Wind speed changes in m/s. (g) Solar changes in MJ/m2.

Please cite this article as: G. Manogaran, D. Lopez, Spatial cumulative sum algorithm with big data analytics for climate change detection, Computers and Electrical Engineering (2017), http://dx.doi.org/10.1016/j.compeleceng.2017.04.006

Table 2
Maximum temperature changes.

Year	Change point	Maximum temperature		Maximum temperature change	Level of change
		From	To		
1990 Winter	12	35.1206	30.278	+4.8426	5
2015 Winter	37	30.278	36.229	−6.051	3
1981 Summer	43	36.229	41.524	−5.295	4
1999 Summer	59	41.524	44.609	−3.085	7
2004 Summer	64	44.609	36.947	+7.662	1
2010 Summer	70	36.947	40.512	−3.565	6
2014 Summer	74	40.512	34.13	+6.382	2

Table 3
Minimum temperature changes.

Year	Change point	Minimum temperature		Minimum temperature change	Level of change
		From	To		
1996 Winter	17	23.4753	18.582	+4.4933	2
2013 Winter	34	18.582	21.247	−2.655	3
1981 Summer	42	21.247	26.819	−5.572	1
2013 Summer	73	26.819	25.74	+1.079	4

Table 4
Precipitation changes.

Year	Change Point	Precipitation (mm)		Precipitation change	Level of change
		From	To		
1987 Winter	8	35	18	+17	3
1992 Winter	13	18	3	+15	4
1989 Monsoon	83	3	47	−44	1
1991 Monsoon	85	47	8	+39	2
2002 Monsoon	92	8	1	+7	5

Table 5
Solar changes.

Year	Change Point	Solar		Solar change	Level of change
		From	To		
1984 Winter	5	17.178	4.231	+12.965	2
1991 Winter	12	4.231	22.435	−18.204	1
2014 Winter	35	22.435	17.543	+4.892	5
1991 Summer	52	17.543	24.221	−6.678	3
2003 Summer	63	24.221	18.325	+5.896	4

Table 6
Wind speed changes.

Year	Change Point	Wind speed		Wind speed change	Level of change
		From	To		
2005 Summer	95	2.862	4.121	−1.259	1

The proposed spatial CUSUM based change detection algorithm is compared with Pruned Exact Linear Time (PELT) method, CUSUM with Bootstrap, binary segmentation method and segment neighborhood method and the results are depicted in Tables 7 and 8. Fig. 14 shows the performance of the proposed spatial CUSUM based change detection algorithm. As shown in Table 5, the performance evaluation of change detection methods is comparatively analyzed with the help of precision value. The precision is defined by,

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

Experimental results prove that Spatial CUSUM based climate change detection algorithm performed well when compared with Pruned Exact Linear Time (PELT) method, binary segmentation method, and segment neighborhood method. The PELT algorithm identifies 20 original changes in the seasonal climate with the precision of 74.07. Similarly, BinSeg, SegNeigh, and CUSUM with Bootstrap methods determine the original changes in the seasonal climate with the precision of 77.77,

Please cite this article as: G. Manogaran, D. Lopez, Spatial cumulative sum algorithm with big data analytics for climate change detection, Computers and Electrical Engineering (2017), http://dx.doi.org/10.1016/j.compeleceng.2017.04.006

Table 7
Comparison of climate change prediction by various methods.

Climate Parameters	Year	Change detection methods				Spatial cusum
		PELT	BinSeg	SegNeigh	Cusum with bootstrap	
Rain Fall Change	2012 Summer	Y	Y	Y	Y	Y
	2014 Summer	Y	Y	Y	Y	Y
	1984 Monsoon	N	Y	N	Y	Y
	1991 Monsoon	N	N	Y	N	N
	1996 Monsoon	Y	Y	Y	Y	Y
Maximum Temperature Change	1990 Winter	N	Y	N	N	Y
	2015 Winter	Y	Y	Y	Y	Y
	1981 Summer	N	Y	Y	Y	Y
	1999 Summer	Y	Y	N	Y	Y
	2004 Summer	N	N	Y	Y	Y
	2010 Summer	Y	N	Y	Y	Y
	2014 Summer	Y	Y	Y	Y	Y
Minimum Temperature Change	1996 Winter	Y	Y	Y	N	Y
	2013 Winter	Y	Y	N	Y	Y
	1981 Summer	Y	N	Y	N	N
	2013 Summer	Y	Y	Y	Y	Y
Precipitation Change	1987 Winter	Y	Y	Y	Y	Y
	1992 Winter	N	Y	Y	N	N
	1989Monsoon	Y	N	Y	Y	Y
	1991 Monsoon	Y	Y	Y	Y	Y
	2002 Monsoon	Y	Y	Y	N	Y
Solar Change	1984 Winter	Y	Y	Y	Y	Y
	1991 Winter	Y	N	Y	Y	Y
	2014 Winter	Y	Y	Y	N	N
	1991 Summer	Y	Y	N	Y	Y
	2003 Summer	Y	Y	Y	Y	Y
Wind Speed Change	2005 Summer	N	Y	T	Y	Y

Note: 'Y' represents 'Yes: change detected' and 'N' represents 'No change is not detected'.

Table 8
Performance evaluation of change detection methods.

Validation Metrics	Change detection methods				
	PELT	BinSeg	SegNeigh	Cusum with bootstrap	Spatial cusum
Correctly Predicted (True Positive)	20	21	22	20	23
Wrongly Predicted (False Positive)	7	6	5	7	4
Precision	74.07	77.77	81.48	74.07	85.18

81.48, and 74.07 respectively. Spatial CUSUM based climate change detection algorithm achieved the precision of 85.18. The SegNeigh method is also performed well when compared to Pruned Exact Linear Time (PELT) method, binary segmentation method. SegNeigh based change detection method achieved the precision of 81.48. In addition, Spatial CUSUM based climate change detection algorithm predicts 23 seasonal changes in the climate data.

7. Conclusion

Big data refers to voluminous amounts of structured or unstructured data that becomes complex to process by using traditional data processing techniques and platforms. In other words, big data is difficult to store, process and visualize using state-of-art technologies. Big Data has gained much attention from many private organizations, public sector, and research institutes. The push towards collecting and analyzing large amounts of data in diverse application domains has motivated us to use a variety of applications such as Health and Human welfare, Nature and natural processes, Government and the public sector, commerce, business and economic systems, social networking and the internet, and computational and experimental methods. This paper uses day wise weather data collected from Global Weather Data for SWAT Inc. The large climate data is stored into the HDFS in distributed manner, and MapReduce algorithm is applied to calculate the seasonal average of various climate parameters such as maximum temperature, minimum temperature, precipitation, wind, relative humidity and solar. In this paper, a novel climate change detection algorithm is proposed to monitor the changes in the seasonal climate. The proposed climate change detection algorithm is compared with various existing approaches such as Pruned Exact Linear Time (PELT) method, binary segmentation method, and segment neighborhood method. The experimental results prove the efficiency of the proposed climate change detection algorithm. As a future work, we intend to use the climate change values to predict the seasonal diseases.

Please cite this article as: G. Manogaran, D. Lopez, Spatial cumulative sum algorithm with big data analytics for climate change detection, Computers and Electrical Engineering (2017), http://dx.doi.org/10.1016/j.compeleceng.2017.04.006

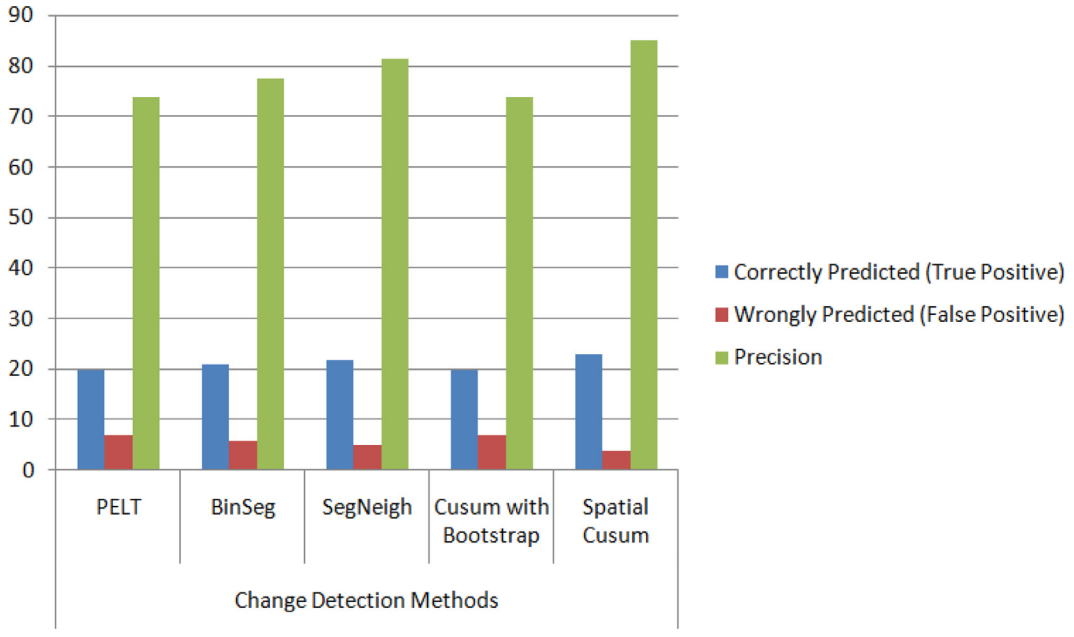


Fig. 14. Precision comparison of various change detection methods.

References

[1] Webteam W. Welcome. WCDMP | WMO [Internet]; 2017. Wmo.int.[cited 7 March 2017]. Available from: http://www.wmo.int/pages/prog/wcp/wcdmp/index_en.php.
[2] World Meteorological Organization [Internet]. World meteorological organization [cited 7 March 2017]. Available from: https://public.wmo.int/en.
[3] Lee JG, Kang M. Geospatial big data: challenges and opportunities. Big Data Res 2015 Jun 30;2(2):74–81.
[4] Nativi S, Mazzetti P, Santoro M, Papeschi F, Craglia M, Ochiai O. Big data challenges in building the global earth observation system of systems. Environ Model Softw 2015 Jun 30;68:1–26.
[5] Faghmous JH, Kumar V. A big data guide to understanding climate change: the case for theory-guided data science. Big data 2014 Sep 1;2(3):155–63.
[6] EnviroAtlas | US Environmental Protection Agency [Internet]. Enviroatlas.epa.gov. [cited 7 March 2017]. Available from: http://enviroatlas.epa.gov/enviroatlas.
[7] Pickard BR, Baynes J, Mehaffey M, Neale AC. Translating big data into big climate ideas. Solutions 2015;6(1):64–73.
[8] Schnase JL, Duffy DQ, Tamkin GS, Nadeau D, Thompson JH, Grieg CM. MERRA analytic services: meeting the big data challenges of climate science through cloud-enabled climate analytics-as-a-service. Comput Environ Urban Syst 2017 Jan 31;61:198–211.
[9] Fiore S, Mancini M, Elia D, Nassisi P, Brasileiro FV, Blanquer I. Big data analytics for climate change and biodiversity in the EUBrazilCC federated cloud infrastructure. In: Proceedings of the 12th ACM international conference on computing frontiers. ACM; 2015 May 6. p. 52.
[10] Lopez D, Gunasekaran M, Murugan BS, Kaur H, Abbas KM. Spatial big data analytics of influenza epidemic in Vellore, India. In: 2014 IEEE international conference on big data (big data). IEEE; 2014 Oct 27. p. 19–24.
[11] Lopez D, Gunasekaran M. Assessment of vaccination strategies using fuzzy multi-criteria decision making. In: Proceedings of the fifth international conference on fuzzy and neuro computing (FANCCO-2015). Springer International Publishing; 2015. p. 195–208.
[12] Lopez D, Sekaran G. Climate change and disease dynamics-a big data perspective. Int J Infect Dis 2016 Apr 1;45:23–4.
[13] Lopez D, Manogaran G. Big data architecture for climate change and disease dynamics. The human element of big data: issues, analytics, and performance. USA: CRC Press; 2016.
[14] Manogaran G, Thota C, Kumar MV. MetaCloudDataStorage architecture for big data security in cloud computing. Procedia Comput Sci 2016 Dec 31;87:128–33.
[15] Manogaran G, Thota C, Lopez D, Vijayakumar V, Abbas KM, Sundarsekar R. Big data knowledge system in healthcare. In: Internet of things and big data technologies for next generation healthcare. Springer International Publishing; 2017. p. 133–57.
[16] Manogaran G, Lopez D. Disease surveillance system for big climate data processing and dengue transmission. Int J Ambient Comput Intell 2017;8(2):88–105.
[17] Thota C, Manogaran G, Lopez D, Vijayakumar V. Big data security framework for distributed cloud data centers. In: Cybersecurity breaches and issues surrounding online threat protection. USA: IGI Global; 2017. p. 288–310.
[18] Manogaran G, Thota C, Lopez D, Vijayakumar V, Abbas KM, Sundarsekar R. Big data knowledge system in healthcare. In: Internet of things and big data technologies for next generation healthcare. Springer International Publishing; 2017. p. 133–57.
[19] Katagi M, Moriai S. Lightweight cryptography for the internet of things. Sony Corporation; 2008. p. 7–10.
[20] Bi Y, Shamsi K, Yuan JS, Standaert FX, Jin Y. Leverage emerging technologies for dpa-resilient block cipher design. In: Proceedings of the 2016 conference on design, automation & test in Europe. EDA Consortium; 2016 Mar 14. p. 1538–43.
[21] Lian S. Multimedia content encryption: techniques and applications. CRC Press; 2008 Sep 17.
[22] Bi Y, Shamsi K, Yuan JS, Jin Y, Niemier M, Hu XS. Tunnel FET current mode logic for DPA-resilient circuit designs. IEEE Trans Emerg Topics Comput 2016 Apr 27.
[23] Bi Y, Shamsi K, Yuan JS, Gaillardon PE, Micheli GD, Yin X. Emerging technology-based design of primitives for hardware security. ACM J Emerg Technol Comput Syst (JETC) 2016 May 6;13(1):3.
[24] Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff 2014 Jul 1;33(7):1123–31.
[25] Murdoch TB, Detsky AS. The inevitable application of big data to health care. Jama 2013 Apr 3;309(13):1351–2.

Please cite this article as: G. Manogaran, D. Lopez, Spatial cumulative sum algorithm with big data analytics for climate change detection, Computers and Electrical Engineering (2017), http://dx.doi.org/10.1016/j.compeleceng.2017.04.006

Gunasekaran Manogaran is currently pursuing PhD in the Vellore Institute of Technology (VIT) University. He received his B.E and M.E from Anna University and VIT University respectively. He has worked as a Research Assistant for a project on spatial data mining funded by Indian Council of Medical Research, India. His current research interests include data mining and big data analytics.

Daphne Lopez is a professor in the School of Information Technology and Engineering, Vellore Institute of Technology University. Her research spans the fields of grid and cloud computing, data mining, and big data. She has vast experience in teaching and industry. Prior to this, she worked in the software industry as a consultant in data warehouse and business intelligence.

Please cite this article as: G. Manogaran, D. Lopez, Spatial cumulative sum algorithm with big data analytics for climate change detection, Computers and Electrical Engineering (2017), http://dx.doi.org/10.1016/j.compeleceng.2017.04.006