An ML-based study of extreme weather events incorporating seasonality factor

Peter Khaiter [a] and Masooma Suleman [b]

[a] School of Information Technology, York University, Toronto ON, Canada, [b] School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India

Email: masooma20suleman@gmail.com

Abstract: It has been recognized that one of the most visible manifestations of global climate change is an increase in the intensity and frequency of extreme weather events. The latter include droughts, heat waves, heavy downpours, tornados, typhoons, and major hurricanes. In this paper, we propose a conceptual framework for assessing extreme weather conditions incorporating the factor of seasonality. The study was based on historical meteorological data of Ahmedabad city, India (latitude: 22.9914, longitude: 72.6167) consisting of daily average temperature (°C), minimum temperature (°C), maximum temperature (°C), wind speed (m/s), surface pressure (kPa) and precipitation (mm) collected over the past 38 years, from 1st January 1982 to 31st December 2020 (Figure 1).

The main steps of the framework are shown in Figure 2. We used boxplot technique to visualize the dataset and determine the central tendency, range, symmetry, and the presence of outliers in data. Predicting extreme weather events based on fixed seasonal time frames may produce inaccurate or biased results. It is important to consider the seasonal variability across the years before detecting extreme weather events and predicting their trends. We conducted cluster analysis to group observed data points into distinct seasons based on the similarity in their meteorological features. Given an obvious time-oriented nature of data, K-means clustering algorithm with dynamic time warping metric to measure similarity have been applied. The resultant clusters (i.e., artificial seasons) have been used to study the variation in seasonality across contiguous years and to identify the long-term trends in extreme weather conditions (namely, temperature and precipitation) within a seasonal context over a 38-year period (1982-2020). Traditionally, the study of extreme weather events includes computation of 5th, 10th, 90th and 95th percentiles of observed meteorological data as thresholds across the time periods, and this approach is extensively applied and recommended in the literature. However, in the prior research, the thresholds have been computed across the whole periods, whereas we used these thresholds and computed them over derived seasonal clusters to analyze the extreme weather events pertaining to a given season. Additionally, we included 1st and 99th percentile thresholds as severe/extreme weather events. The magnitude trends in extreme hot and extreme cold events during each season and extreme rainfall events in the "Monsoon" season have been estimated and visualized.

It would be worthwhile to include the intensity of precipitation and humidity for a finer determination of seasonality. In combination, we can analyse the contribution of each meteorological feature to the formation of clusters (seasons) and compare our obtained results with different permutation of features in K-means.

Figure 1. Location of the study area (city of Ahmedabad in India).

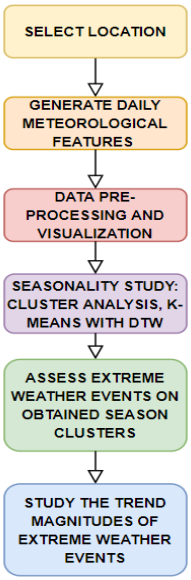


Figure 2. The proposed conceptual framework for studying seasonality and extreme weather events factoring seasonality.

Keywords: *Unsupervised learning, extreme weather event, dynamic time warping, seasonal clustering*

1. INTRODUCTION

Numerous efforts have been made to define the term "extreme weather" but no universal formalized definition quantifying the term "extreme" exists so far. There are multiple types of severe and extreme weather conditions, such as heat waves, thunderstorms, cold spells, hurricanes, tornadoes, hails, ice and dust storms, *etc*. which are determined by the combination of meteorological features. It is important to quantify "extreme" weather conditions and define absolute thresholds to the degree of severity based on daily meteorological observations. One such approach is to define severity based on percentiles which is convenient to produce seasonally aggregated scores and perform comparison between seasons and regions (Magnusson et al. 2014). The meteorological attributes used to define the extreme weather events for different seasons should also be considered chronologically, i.e., in their temporal order. In reality, however, seasons do not always start and end on fixed calendar dates but are variable and dependent on the daily meteorological features like air temperature, wind speed, precipitation, surface pressure, humidity, *etc*.

Predicting extreme weather events based on fixed seasonal time frames may produce inaccurate or biased results. For example, if we consider the winter season to span from 1st November to 31st January for a certain location it is not necessary that the winter season will occur every year exactly during the same time frame. Practically, we observe seasonal shift in either direction every 12 months. Therefore, it is important to take into account the seasonal variability across the years before detecting extreme weather events and predicting their trends.

Machine learning (ML) methods (such as clustering) can be applied on panel data to segregate observed data points into distinct seasons. This artificial grouping of multi-dimensional data points to form seasons based on the similarity in their meteorological attributes would help in the study of seasonal variability and detection of extreme weather events while taking the seasonality factor into consideration.

This paper describes a generic framework to group observed data points to form seasons based on the similarity in their meteorological features. The data used for this research is time-oriented and thus we use K-means clustering algorithm with dynamic time warping metric to measure similarity (Petitjean et al. 2010). The resultant clusters (i.e., seasons) are analysed to study the variation in seasonality across contiguous years. Furthermore, the derived seasonal clusters are used to analyse the long-term trends in extreme weather conditions (namely, temperature and precipitation) within a seasonal context over a 38-year period (1982-2020) using formalized definitions of "extreme" weather conditions from the literature.

2. METHODOLOGY

Unsupervised learning unlike other ML algorithms, such as regression, does not produce results based on characteristics or features but rather categorizes each data point based on similarity in its attributes or patterns discovered from its associated data (Mahesh 2020). It can be applied to classify data with no class labels. We use this technique to cluster our panel data containing time stamp and meteorological features into distinct seasons based on the similarity and sequential order of twelve calendar months within a year aiming at partitioning of contiguous months with the most similar weather patterns and labelling them as respective seasons. K-means and agglomerative hierarchical algorithms are the two most popular clustering techniques (Garima et al. 2005). K-means algorithm assigns each data point to its nearest centroid based on similarity, and the group of points particularly close to the centroid forms a cluster. This type of partitioning maximizes the measure of similarity within the cluster and minimizes similarity between the clusters (Badhiye et al. 2012). On the other hand, agglomerative hierarchical algorithm initially considers each data point as a single-element cluster (leaf) and iteratively merges closest pairs of clusters together until each data point is categorized into any one of the clusters. Both K-means and agglomerative methods are parametric algorithms; that is, they require a user to pre-define a fixed number (k) of clusters in a dataset (in our case – the number of seasons). We can identify the optimal number of clusters (seasons) using the elbow method or silhouette coefficient (Kodinariya et al. 2013). This automates the process of setting the optimal number of seasons based on the input dataset (meteorological time series) and establishes a generic approach as it depends only on the meteorological features and is universal relative to geographical location. The elbow technique uses the within cluster sum of squares (WCSS, the squared average distance of all the points within a cluster to the cluster centroid) as the performance indicator and calculates it for a series of k (number of clusters) values. On plotting WCSS vs k, the point of inflection (bend/elbow) gives the optimal value of k. The silhouette score is determined by the ratio of the difference of mean intra cluster distance and mean nearest cluster distance to the maximum of either of them. As a result, the maximum silhouette coefficient gives the ideal respective k value.

However, in clustering algorithms, the standard metric used for similarity measure is the Euclidean distance known to be very sensitive to distortion in time axis (Chu et al. 2002). Neglecting the temporal shifts and order of data can result in poor accuracy of clusters which would be propagated to biased grouping of months. This problem can be addressed by Dynamic Time Warping (DTW), an extremely efficient algorithm for time-series similarity measure which is sensitive to the temporal shifts and minimizes the effects of distortion in time (Ratanamahatana et al. 2004). DTW measures similarity between two time-series which do not align exactly in time, speed or length; that is, the sequences are similar but locally out of phase. For example, in Figure 3, both red and blue time series have similar shape but are not aligned in the same time axis. Here, the Euclidean metric calculates similarity based on the alignment of i^{th} point in red sequence to i^{th} point in blue sequence which would produce a pessimistic dissimilarity measure whereas DTW allows innate similarity calculations of non-linearly aligned sequences. It works best for clustering meteorological time series into different seasons ensuring accuracy and reducing time complexity as well as matching similar trends to each other.

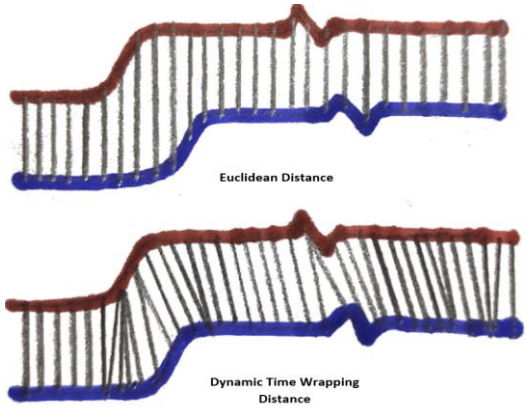


Figure 3. Euclidean distance measure and dynamic time warping distance measure.

After clustering our meteorological time-series into distinct seasons and grouping the respective seasonal months together using DTW, we use the derived season-based clusters to study the extreme weather events and analyze the shift of seasonal span across the years. Five seasons (i.e., winter, spring, summer, monsoon and autumn) are investigated. The study of extreme events includes computation of 5^{th}, 10^{th}, 90^{th} and 95^{th} percentiles as thresholds across the whole time periods which is extensilvely applied and recommended by STARDEX (Statistical and Regional Dynamical Downscaling of Extremes for European Regions; http://www.cru.uea.ac.uk/projects/stardex/) and the ETCCDI (Expert Team on Climate Change Detection and Indices; http: //cccma.seos.uvic.ca/ETCCDI/) projects (Khomsi et al. 2016). While in the above-mentioned projects, they compute the thresholds across the whole periods, we use these thresholds and compute them over derived seasonal clusters to analyze the extreme weather events within a given season. Additionally, we include 1^{st} and 99^{th} percentile thresholds as severe/extreme events. Tables 1 and 2 define the thresholds used to classify the weather events as extreme, heavy or exceptional.

Table 1. Percentile thresholds and their respective definition/meaning w.r.t. temperature events.

Percentile	Definition/Meaning
99^{th} (1^{st})	All the events whose maximum (minimum) temperature is greater (lower) than or equal to the given threshold is an extremely hot (cold) day.
95^{th} (5^{th})	All the events whose maximum (minimum) temperature is greater (lower) than or equal to the given threshold is a very hot (cold) day.
90^{th} (10^{th})	All the events whose maximum (minimum) temperature is greater (lower) than or equal to the given threshold is a hot (cold) day.

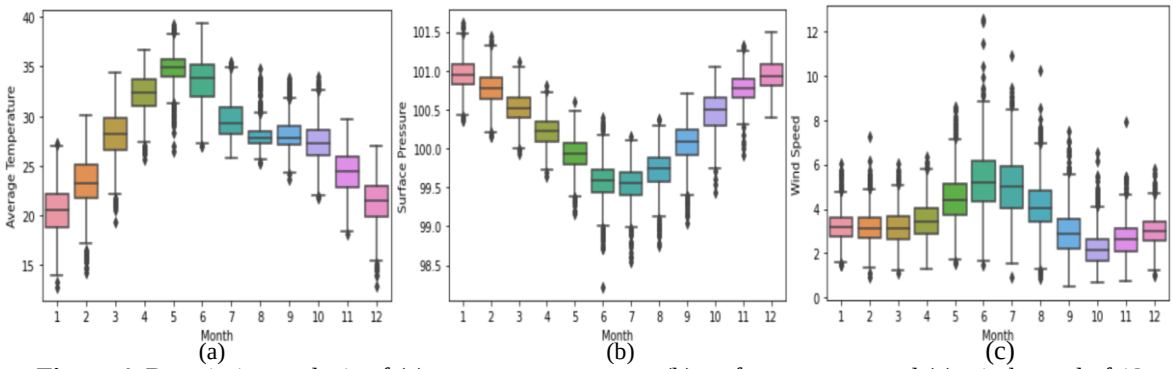
Table 2. Percentile thresholds and their respective definition/meaning w.r.t. precipitation events.

Percentile	Definition/Meaning
99^{th}	All the events whose daily precipitation value is greater than or equal to the given threshold is classified as an exceptional/extreme precipitation day.
95^{th}	All the events whose daily precipitation value is greater than or equal to the given threshold is classified as an intense precipitation day.
90^{th}	All the events whose daily precipitation value is greater than or equal to the given threshold is classified as a heavy precipitation day.

3. DATA SOURCE

We have captured historical meteorological data of Ahmedabad city, India (latitude: 22.9914, longitude: 72.6167) from NASA Prediction of Worldwide Energy Resources (https://power.larc.nasa.gov) and Power Data Access Viewer (https://power.larc.nasa.gov/data-access-viewer/). The dataset consists of daily average temperature (°C), minimum temperature (°C), maximum temperature (°C), wind speed (m/s), surface pressure

(kPa) and precipitation (mm) measured over the past 38 years, from 1st January 1982 to 31st December 2020, with no missing values. In order to visualize the dataset, we made use of boxplots as shown in Figure 4 which highlights the five most important descriptive statistics, i.e., mean, median, the minimum and maximum as well as the first and third quartiles. Boxplots also indicate the central tendency, range, symmetry and the presence of outliers in the dataset (Boslaugh 2012).



(a) (b) (c)

Figure 4. Descriptive analysis of (a) average temperature, (b) surface pressure and (c) wind speed of 12 months from 1982-2020 visualized as boxplots.

4. RESULTS

The results of this project are two-fold: the seasonality clustering is presented in section 4.1 and the study of extreme weather conditions incorporating the factor of seasonality is presented in section 4.2.

4.1. Seasonality study

Phenologically for Ahmedabad, the five seasons of the year are: winter (December – February), spring (March – April), summer (May – June), monsoon (July – September) and autumn (October - November), and the grouping of these months in the corresponding seasons can be confirmed by Figure 4. However, in reality, the starting and ending dates of the seasons fluctuate every year and do not necessarily align with the phenological norm. To analyse these seasonal shifts and possible trends, we apply clustering to group the data points into seasons based on the daily maximum temperature, minimum temperature, wind speed and surface pressure.

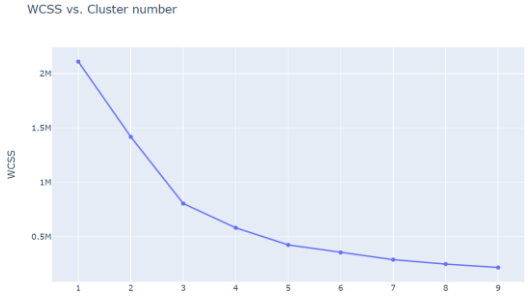


Figure 5. Elbow Plot to determine the optimal number of clusters (seasons).

The optimal number of clusters (seasons) for the input features (average temperature, minimum temperature, maximum temperature, wind speed and surface pressure) is calculated using elbow method and the resultant WCSS vs *k* plot is depicted in Figure 5. Since the graph flattens after 4 and 5, we can safely conclude that 4 is the optimal value for *k*. Using DTW metric for K-means, we perform time series clustering on our dataset, and the clusters obtained using above-mentioned input features are shown in Figure 6.

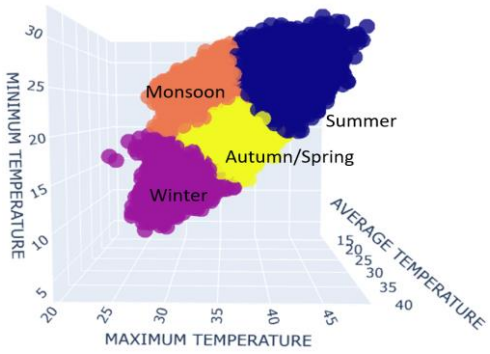


Figure 6. 3D scatter plot (minimum temperature, maximum temperature and average temperature) of data points with clusters obtained from DTW metric for K-means.

The labelling of clusters to their respective seasons can be explained using Figure 7 which depicts the average temperature curve for the year 1982 (over 360 days of the year) with colour encoded clusters. Clearly, the purple clustered data points fall in the lower temperature values and occur in the end of the year and beginning of the next year (i.e., approximately correspond to November – mid March period). Accordingly, we mark the purple cluster as "Winter" season. Similarly, the blue cluster occupies the highest temperature values and spreads from mid-April to June. We label the blue cluster as "Summer" season. The orange cluster essentially begins straight

after the "Summer" season and follows a negative slope (decline in temperature values) spreading across July to September months of the year. It is labelled as the "Monsoon" season. The yellow cluster includes two portions of the year, namely, pre-summer ("Spring" season) and post- monsoon ("Autumn" season). Data points belonging to "Autumn" and "Spring" seasons are clustered together because of their meteorological similarity in terms of minimum temperature and maximum temperature with the only distinguishing factor being the time-order of their occurrence within the year using which we label the pre-summer as "Spring" season and post-monsoon as "Autumn" season.

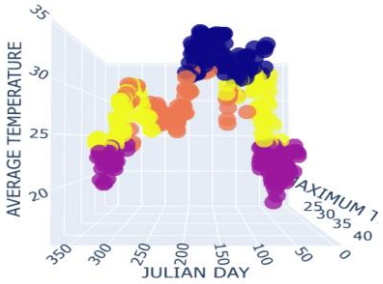


Figure 7. Average Temperature and Julian Day curve for the year 1982 with colour encoded seasons.

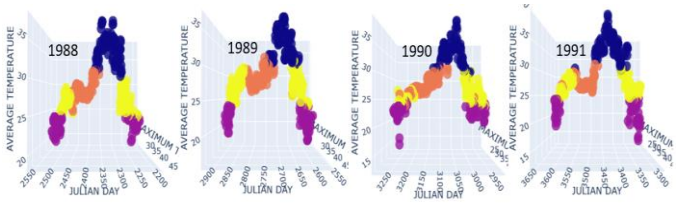


Figure 8. Seasonal shift across four contiguous years (1988 – 1991).

The derived clusters help us to analyse possible shifts in the seasons over the past 38 years. For example, Figure 8 and Table 3 (contains season end dates for 1988-1991) depict the shift and span of seasons across four contiguous years (1988-1991). These results are practical and informative to analyse the trend and shift in the seasonality based on historical data and contribute significantly to the forecasting of future variations.

Table 3. Season end dates for four contiguous years based on derived clusters (1988-1991).

Year	Winter	Spring	Summer	Monsoon	Autumn
1988	23rd February	5th April	26th June	16th September	5th November
1989	3rd March	16th April	25th June	17th September	17th November
1990	7th March	10th April	26th June	25th September	2nd November
1991	28th February	17th April	12th June	12th September	24th November

Table 4. Minimum (cold) and maximum (hot) temperature magnitudes estimated from the thresholds shown in Table 1 over 38 years (1982-2020) using the derived seasonal clusters.

Season	Extremely Hot (Daily Maximum temperature)	Extremely Cold (Daily Minimum Temperature)
Summer	>=46.45°C	<=21.27°C
Winter	>=34.82°C	<=6.88°C
Monsoon	>=36.87°C	<=20.66°C
Autumn/Spring	>=41.02°C	<=14.56°C

4.2. Extreme weather events factoring seasonality

We study the extreme weather events and their point of occurrence based on the results derived from seasonal clustering as described in section 4.1. Performing univariate analysis using the threshold percentiles described in Tables 1 and 2, we quantify the range of "extreme" weather points for all the four clusters ("Summer", "Winter", "Monsoon" and "Autumn"/"Spring"). The magnitude trends in extreme hot and extreme cold events during each of the four derived seasons as well as extreme rainfall events in the monsoon season have been estimated and the results are shown in Tables 4 and 5, respectively. We observe that the occurrence of extreme events (i.e., hot, cold and intense precipitation) is not contiguous throughout the studied years (1982-2020), meaning that there are either few years containing many extreme events or containing few extreme cold events but no extreme hot event or not containing any extreme event. This is illustrated in Figures 9–16 presenting the number of days in the year with extreme events and also showing the trend line over the period from 1982 to 2020. The trend magnitudes of extreme events (both hot and cold) are differently directed for every season. For the "Summer" season (Figures 9 and 10), we observe that the trend magnitude slightly increased for extreme cold events and slightly decreased for extreme hot events over the years. At the same time, the trend magnitude for all the other cases decreased (Figures 11-16).

Table 5. Precipitation magnitudes estimated from the thresholds mentioned in table 2 for 38 years (1982-2020) using the derived monsoon cluster.

	Daily Precipitation (mm)
Exceptional/Extreme Precipitation	>= 68.0389
Intense Precipitation	>=33.2829
Heavy Precipitation	>=20.4559

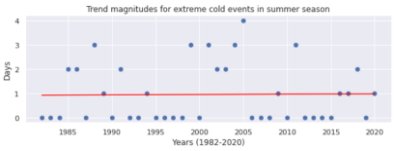


Figure 9. Trend magnitudes for extreme cold events in summer season.

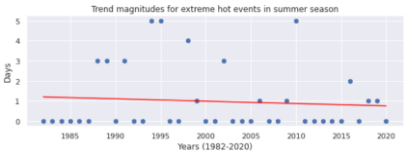


Figure 10. Trend magnitudes for extreme hot events in summer season.

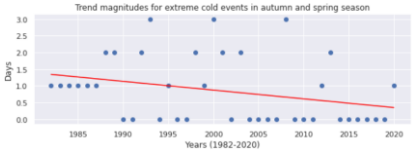


Figure 11. Trend magnitudes for extreme cold events in autumn and spring seasons.

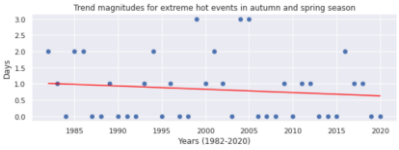


Figure 12. Trend magnitudes for extreme hot events in autumn and spring seasons.

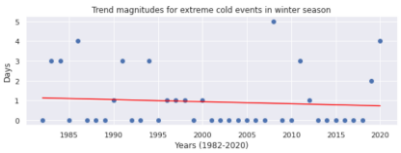


Figure 13. Trend magnitudes for extreme cold events in winter season.

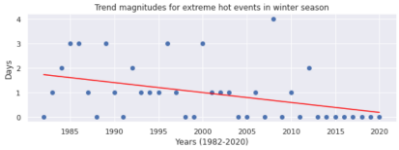


Figure 14. Trend magnitudes for extreme hot events in winter season.

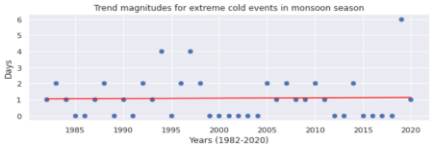


Figure 15. Trend magnitudes for extreme cold events in monsoon season.

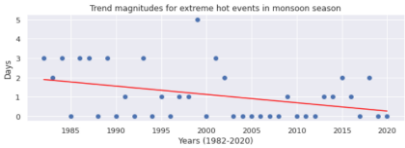


Figure 16. Trend magnitudes for extreme hot events in monsoon season.

Contrary to the extreme temperature events, the trend magnitude for extreme precipitations (Figure 17) demonstrates a steady increase over the studied 38 years.

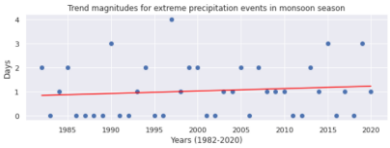


Figure 17. Trend magnitudes for extreme precipitation events in "Monsoon" season.

5. CONCLUSION AND FUTURE WORK

The results obtained above show the variability in seasonality over years and its effect and contribution to determine the extreme weather events. Over 38 years from 1982 to 2020, the extreme temperature (both maximum and minimum) trends for almost all the seasons are seen to decline in the studied area, while a statistically significant increasing trend is observed in the case of extreme precipitation in "Monsoon" season. This response depends on the clusters obtained from unsupervised K-means in an attempt to cluster data points to form seasons.

It would be worthwhile to include the intensity of precipitation and humidity as factors in determining the seasonality. Future study could also investigate the consistency of the obtained extreme weather events and their trends with the literature. Evaluation of the expansion/contraction of each season over a multi-year period could provide a significant dimension in the analysis of extreme weather events. In combination, we can assess the contribution of each meteorological feature to the formation of clusters (seasons) and compare obtained results with different permutations of features in K-means.

ACKNOWLEDGMENTS

This project is partially supported by the Mathematics of Information Technology and Complex Systems (MITACS NCE) Globalink Research Grant No. 91947-2021. The authors are greatly thankful to the anonymous reviewers for their valuable comments and suggestions which helped to improve the manuscript.

REFERENCES

Badhiye, S.S, Chatur, P.N., Wakode, B.V., 2012. Temperature and Humidity Data Analysis for Future Value Prediction using Clustering Technique: An Approach. International Journal of Emerging Technology and Advanced Engineering 2(1), 88-91.

Boslaugh, S., 2012. Statistics in a nutshell, 2nd Edition. O'Reilly Media, Inc. Chapter 4: Descriptive Statistics and Graphics, 71-73.

Chu, S., Keogh, E.J., Hart, D., Pazzani, M., 2002. Iterative deepening dynamic time warping for time series, 18pp. Second SIAM International Conference on Data Mining, Arlington, VA, USA, April 11-13, DOI: 10.1137/1.9781611972726.12.

Garima, Gulati, H., Singh, P.K., 2015. Clustering techniques in data mining: A comparison. 2nd International Conference on Computing for Sustainable Global Development (INDIACom), 410-415.

Khomsi, K., Mahe, G., Tramblay, Y., Sinan, M., Snoussi, M., 2016. Regional impacts of global change: seasonal trends in extreme rainfall, run-off and temperature in two contrasting regions of Morocco. Natural Hazards and Earth Systems Sciences 16, 1079-1090, https://doi.org/10.5194/nhess-16-1079-2016.

Magnusson L., Richardson, D., Haiden, T., 2014. Verification of extreme weather events: Discrete Predictands, 27pp. European Centre for Medium-Range Weather Forecasts, Technical Memoranda, 731, DOI: 10.21957/1iql31n2c.

Mahesh, B., 2020. Machine Learning Algorithms: A Review. International Journal of Science and Research 9(1), 381-386.

Petitjean, F., Ketterlin, A., Gançarski, P., 2011. A global averaging method for dynamic time warping, with applications to clustering. Pattern Recognition 44, 678-693, DOI: 10.1016/j.patcog.2010.09.013.

Kodinariya, T.M., Makwana, P.R., 2013. Review on determining number of clusters in K-Means Clustering. International Journal of Advance Research in Computer Science and Management Studies 6(1), 90-95.

Ratanamahatana, C.A., Keogh, E., 2004. Making Time-Series Classification More Accurate using Learned Constraints, 12pp. Fourth SIAM International Conference on Data Mining, Lake Buena Vista, Florida, USA, April 22-24, DOI: 10.1137/1.9781611972740.2.