Available at: www.researchpublications.org

Detection of Rare Patterns in Climate Change Using Data Mining Techniques

Mr. Parag N. Kolhe[1] Mr. Rahul M.Ugale[2] Miss Shraddha V Shingne[3]
[1]*SRTMU Nanded university, Nanded, India*
[2]*SGBAU Amravati university, Amravati, India*
[3]*RSTM Nagpur University, Nagpur, India*
[1]paragkolhe444@gmail.com
[2]rahulugale4@gmail.com
[3]s.shingne90@gmail.com

***Abstract*—Today the volume of data has been enormously increasing as a result of advances in data generation, collection and storage technologies. The effect of climate prediction on society, business, agriculture and almost all aspects of human life, force the scientist to give proper attention to the matter. Weather is a continuous, data-intensive, multidimensional, dynamic process that makes weather forecasting a formidable challenge. The prediction of rare events is a pressing scientific problem. We primarily concentrate to identify weather patterns in the long term while consistent with global climate change on weather patterns, identify rare/outlying patterns that coincide with rare events data mining. This paper propose an adaptive clustering pattern detection method for the detection of rare patterns in climate change using data mining techniques which uses k-means algorithm where an open number of states as clusters to accommodate the dynamic temporarily of data. By adding adaptive clustering property as a global restriction, the granular size of the clusters is determined for optimal performance. The global modeling result is presented which provides a base of data mining tasks. The metrological variables longitude, latitude, mer wind, zone wind, humidity, air temperature and sea surface temperature are analyzed to detect climate change patterns in this study. The result depicts different patterns of climate in the form of histogram based on the records of metrological variables of NOOA. Different distance measures are applied between the centers of the clusters formed for testing the sensitivity of the method. The robustness of our method is demonstrated by the results. Our method of detecting rare patterns in climate change will be very useful for weather and metrological research focusing on the trends in weather and the consequent changes. Adaptive clustering method uses an open number of states as clusters to accommodate the dynamic temporarily of data. By adding adaptive clustering property as a global restriction, the granular size of the clusters is determined for optimal performance. The global modeling result is presented which provides a base of data mining tasks. This adaptive climate change pattern detection algorithm will be proven to be of potential use for climatic and meteorological research as well as research focusing on temporal trends in weather and the consequent changes.**

Keywords*—Climate change, data mining, adaptive clustering, meteorological data, pattern detection, weather patterns.

I. INTRODUCTION AND RELATED WORK

Climate change is a widely recognized global environmental challenge [1]. A successful addressing of this challenge is essential to the sustainability of modern urban living, especially in domains such as environmental engineering, ecological management, human health, and global and regional economic systems. In this paper, we introduce techniques to understand and detect the patterns of climate change by using data from daily weather records.

This, in turn, may benefit studies on public health (e.g. spread of dengue disease and incidences of respiratory diseases), energy consumption, while they also indirectly have an effect on urban problems such as mobility (e.g. flood events and traffic slowdown) which are affected by weather patterns.

Pattern detection, especially anomaly pattern detection, as a data mining task, refers to disclosing patterns that do not conform to expected behaviors in databases. These unusual patterns are often referred to as different terms in different application domains, such as rare patterns [2], outliers [3], [4], faults [5], peculiarities or contaminants, and etc [6].

Most existing pattern detection techniques resolve specific formulations of a problem. The formulations are induced by various factors such as the nature of the data, availability of the labeled data, and the type of anomalies to be detected, and etc. A considerable amount of data mining research on pattern detection has been conducted, and this stream has gained considerable interest owing to the realization that anomaly patterns can be detected from very large databases by data mining [4]. In addition, pattern detection is normally carried out through fault detection methods which analyze the current signals obtained from induction machines [5]. High-dimensional current signals are transformed to low-dimensional data by the mapping of the original signals into different clusters according to their characteristics [5].

Pattern detection methods in the field of data mining typically pick out different patterns (clusters), their changes, and rare/outlying events and such changes are often the source of problems in impact studies [6]. They have been successfully applied to many fields. However, only a few

studies have been adapted for environment, weather or climate applications.

From a meteorological viewpoint, research on climate change to disclose typical seasonal patterns of weather, such as the seasonal variability of thermal conditions in Singapore is very important [7], [8]. Natural variability in the climate of Singapore is influenced primarily by the monsoonal influences in various months of the year. Studies on Singapore data have shown a long-term increase in temperature in the past few decades [9], [10], in line with global warming studies over the Southeast Asian region [11]. The anomaly pattern or rare events may also be attributed to anomalous conditions such as the El Nino or Southern Oscillation (ENSO) phenomena [12]–[13].

On a larger scale, climate change is an unprecedented environment change that is affecting our planet [14], [15]. It is already having significant impacts on many aspects of our lives [14], [16]–[19]. Climate change projections on both the global and regional scales are characterized by multiple sources of uncertainty [16]. In order to characterize such uncertainties, global and regional climate model projections need to be based on probabilistic approaches using multimodel ensembles of experiments [16].

Kyung Soo Jun et al. discussed the impact of climate change on spatial water resources. The study was a new attempt to quantify hydrologic vulnerability that included the impacts of climate change [17]. The long term impact of climate change on the carbon budget of Lake Simcoe, Ontario were discussed by analyzing the relationship between temperature and dissolved inorganic carbon in some tributaries [19]. Anna Augustsson et al. discussed how the climate effect can be inserted in a commonly used exposure model, and how the exposure then changes compared to present conditions [18]. The results indicated that changes in climate are likely to affect the speciation, mobility, and risks associated with metals.

As mentioned previously, many research activities have been carried out to improve the understanding of climate change patterns by means of different techniques and made considerable contributions [19]–[20]. However, the introduction of data mining techniques into this research field has been limited. Therefore, this paper aims to develop a detection method based on data mining techniques for detecting and classifying weather patterns through a case study on weather data.

In this paper, we propose an adaptive clustering pattern detection data mining method based on an incremental cluster chain. The proposed method, an adaptive clustering pattern detection method, has a flexible structure for allowing the pattern grows and the clusters are adjusted adaptively. The proposed method is applied to analyze the weather patterns through meteorological data mining and different weather patterns are disclosed. The results indicate that the early weather patterns disappear consistently across models and this suggests long-term climate changes. The proposed pattern detection algorithm will be of potential use for climatic and meteorological research as well as research focusing on pattern recognizes or knowledge discovery in other research field.

II. OBJECTIVE

The objective of the proposed method is to detect the climate change patterns through meteorological data mining [21]. Meteorological variables, including longitude, latitude, mer wind, zone wind, humidity, air temperature and sea surface temperature are simultaneously considered for identifying climate change patterns. Different scenarios with varied cluster distance thresholds are employed for testing the sensitivity of the proposed method. The robustness of the proposed method is demonstrated by the results. It is observed from the results that the early weather patterns that were present in past disappear consistently across models. Changes in temporal weather patterns suggest long-term changes to the global climate which may be attributed in part to urban development, and global climate change on a larger scale. Our climate change pattern detection algorithm is proven to be of potential use for climatic and meteorological research as well as research focusing on temporal trends in weather and the consequent changes. The research promotes the cause of advanced study and research and to elect coordination of research and investigations in all disciplines related to weather forecasting. Ultimately the Objective is to:

1. Detect rare weather patterns in the long-term while consistent with global climate change on weather patterns.
2. Predict future environmental changes based on historical data.

III. MATERIALS AND METHODS

A. Data collection

This is the most important part while implementing any of the data mining technique and thus for this purpose 10 channel midi-data logger system can be used. This system provides weather data in the form of excel sheets. Data Loggers are based on digital processor. It is an electronic device that record data over the time in relation to location either with a built in instrument or sensor or via external instruments and sensors. Data Logger can automatically collect data on a 24-hour basis; this is the primary and the most important benefit of using the data loggers [21]. It is used to capture the weather data from the local weather station to a dedicated PC located in the laboratory. The transmitted weather data was then copied to Excel spreadsheets and archived on daily basis as well as monthly basis to ease data identification.

Meteorological data in the form of daily summary data were extracted from National Climate Data Center (NCDC), National Oceanic and Atmospheric Administration (NOAA) [22]. Study of these metrological data gives us some sort of ideas about weather patterns and by processing such data and by applying data mining techniques we can properly detect the outliers in weather patterns and can predict future environmental changes which can affect the climate and day to day human life. Climate variables studied include

longitude, latitude, mer wind, zone wind, humidity, air temperature and sea surface temperature Other variables were not included due to the presence of large amount of missing data in data sets such as mean sea level pressure, mean station pressure, maximum wind gust and precipitation amount, etc.

B. Data Preprocessing

An important step in the data mining process is data preprocessing. One of the challenges that face the knowledge discovery process in meteorological data is poor data quality. For this reason we try to prepare our data carefully to obtain accurate and correct results. First we choose the most related attributes to our mining task. For this purpose we select all the data collected in the excel sheet which is then we try to fill the missing with appropriate values. Because we are working with weather data that is a form of time series, we must preserve the series smoothness and consistency. So we use imputation method [23]. This method is effective method to fill missing values in the case of time series where the missed value is strongly related to its previous and next values.

Obs	Yr	Month	Day	Latitude	longitude	Zone. Winds	mer. winds	Humidity	Air temp.	s.s. Temp.
1	80	3	7	.	-139.96	-5.8	-3.4	88.9	27.75	28.71
2	80	3	8	-2.03	-139.97	-4.5	-5.2	87.7	27.08	28.27
3	80	3	9	-2.02	-139.97	-3.6	-4.5	83.1	26.6	27.7
4	80	3	10	-2.02	-139.96	-2.6	-2.3	84.4	26.41	27.27
5	80	3	11	-2.02	-139.96	-1.2	-5.1	88.5	25.93	27
6	80	3	12	-2.01	-139.96	-4	-4.3	84.4	25.54	26.77
7	80	3	13	-2.01	-139.96	-4	-3.1	84.8	25.69	26.87
8	80	3	14	-2.02	-139.96	-2.8	-4.5	89.3	25.61	26.86
9	80	3	15	-2.02	-139.97	.	.	.	25.26	26.55
. . . . N										

Table .1. Metrological data in excel sheet

Imputation: Use the attribute mean to fill in the missing value, or use the attribute mean for all samples belonging to the same class to fill in the missing value: smarter

It is noteworthy that the daily summary data is not continuously available for every day.

From the logical point of view the missing values can be replaced easily by imputation [23].As it's a fact that sudden change in climate for each and every metrological variable is not possible, so missing values can be logically replaced. The number of missing values for each metrological variable can be one, two, three or its continuous and more than three values .If only single value is missing then it can be replace by the value of previous or next day for that particular variable as there will not much difference for every metrological value for that particular day. If continuous values are missing then each next missing value can be replaced by the mean of three consecutive, exactly previous values to the missing value. Thus the missing values can be efficiently replaced by imputation process using the attributes values belonging to that same class and we logically improve the imputation method to replace more proper values.

This historical metrological data can be organized in excel sheets to get it as an input data so as to analyze, preprocessed and getting the result. In order to make the data appropriate for data mining, the metrological data studied are converted to one vector [24] for each day.

$$X_k = (X_k(1)X_k(2) \quad \ldots\ldots \quad X_k(n)) \qquad (1)$$

Where $X_k(1)X_k(2) \quad X_k(n)$(n=7 for this proposed method) representthe values of longitude, latitude, mere wind, zone wind, humidity, air temperature and sea surface temperature ;k(k=1,2, N) is the input data index and Nis the total number of available days (N=700).

C. K-MEANS

K-means clustering is a partitioning based clustering technique of classifying/grouping items into k groups (where k is user specified number of clusters). The grouping is done by minimizing the sum of squared distances (Euclidean distances) between items and the corresponding centroid. A centroid (also called mean vector) is "the center of mass of a geometric object of uniform density". Although K-means is simple and can be used for a wide variety of data types, it is quite sensitive to initial positions of cluster centers. There are two simple approaches to cluster center initialization i.e. either to select the initial values randomly, or to choose the first k samples of the data points. As an alternative, different sets of initial values are chosen (out of the data points) and the set, which is closest to optimal, is chosen. However, testing different initial sets is considered impracticable criteria, especially for large number of clusters, Is mail et al (1989). Therefore, different methods have been proposed in literature by Pena et al. (1999). Also, the computational complexity of original K-means algorithm is very high, especially for large data sets. Computer science has been widely adopted in different fields like agriculture. One reason is that an enormous amount of data has to be gathered and analyzed which is very hard or even impossible without making use of computer systems. . The research of spatial data is in its infancy stage and there is a need for an accurate method for

rule mining. Association rule mining searches for interesting relationships among items in a given data set. This method enables us to extract pattern.

K-Means is one of the simplest unsupervised learning methods among all partitioning based clustering methods. It classifies a given set of n data objects in k clusters, where k is the number of desired clusters and it is required in advance. A centroid is defined for each cluster. All the data objects are placed in a cluster having centroid nearest (or most similar) to that data object. After processing all data objects, k-means, or centroids, are recalculated, and the entire process is repeated. All data objects are bound to the clusters based on the new centroids. In each iteration centroids change their location step by step. In other words, centroids move in each iteration. This process is continued until no any centroid move. As a result, k clusters are found representing a set of n data objects.

i. K-means Algorithm:

1) Specify k, the number of clusters to be generated
2) Choose k, points at random as cluster centers
3) Assign each instance to its closest cluster center using Euclidean distance
4) Calculate the centroid (mean) for each cluster; use it as a new cluster center
5) Reassign all instances to the closest cluster center
6) Iterate until the cluster centers don't change anymore.

For each of the initial clusters specified, the random centroid of thei^{th}cluster with m−1 number of input data is represented as following:

$$C_{m-1}^{(i)} = \frac{\sum_{j=1}^{m-1} x_j}{m-1} \qquad (2)$$

The similarity distances for the numeric values of attributes and similarity count for the number of attributes is set commonly for each initially formed clusters and if the new instance of data satisfy both conditions then each new instance of the data is assigned to its closest cluster center using Euclidean distance Formula . The Euclidean distance is employed to calculate the distance between the two input data vectors:

$X_m = (X_m(1)X_m(2) \ldots\ldots X_m(n))$ And $C_{(in)} = (C_{(in)}(1)C_{(i)}(2) \ldots\ldots C_{(i)}(n))$ (3)

in n-dimensional space:

$$D(X_m, C^{(i)}) = \sqrt{\sum_{j=1}^{n}(X_m(j) - C^{(i)}(j))}\,2 \qquad (4)$$

The centroid of the new formed cluster for each new instance of data is updated every time using adaptive clustering:

$$C_m^{(i)}$$

$$= \frac{\sum_{j=1}^{m} x_j}{m}$$

$$= \frac{\sum_{j=1}^{m-1} x_j}{m-1} * \frac{m-1}{m}$$

$$= \frac{\sum_{j=1}^{m-1} x_j + x_m}{m-1} * \frac{m-1}{m}$$

$$= \left(\frac{\sum_{j=1}^{m-1} x_j}{m-1} + \frac{x_m}{m-1}\right) * \frac{(m-1)}{m}$$

$$= \left(C_{m-1}^{(i)} + \frac{x_m}{m-1}\right) * \frac{(m-1)}{m}$$

$$= \frac{m-1}{m} C_{m-1}^{(i)} + \frac{x_m}{m_i} \qquad (5)$$

D. Adaptive cluster formation.

We describe the principal idea of the proposed method here. A simple example graph as shown in Fig. 1. If a new input is not sufficiently similar to any existing clusters, it will form a cluster which represents a new pattern. Through the procedure of data processing, each input data is either placed in existing clusters or treated as a new pattern according to its similarity to existing clusters. A simple example is shown in Fig. 1 where in the fourth cluster is treated as a new pattern. The advantage of the proposed detection method is that the similarity distance threshold can be tuned. More defined weather patterns can be discovered by setting a lower cluster distance threshold which represents a higher resolution.

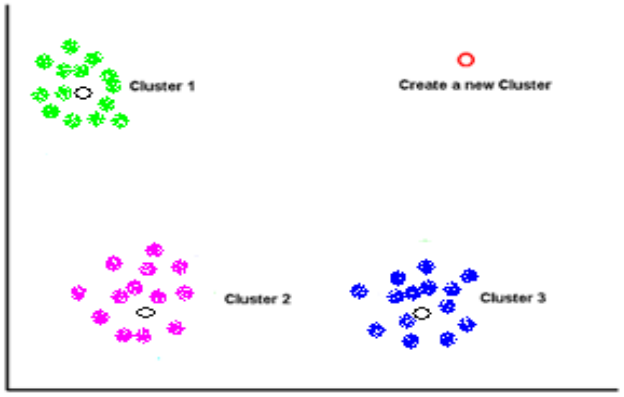

Fig .1. Explanation of the principle of the proposed method by using 2-dimensional data as an example

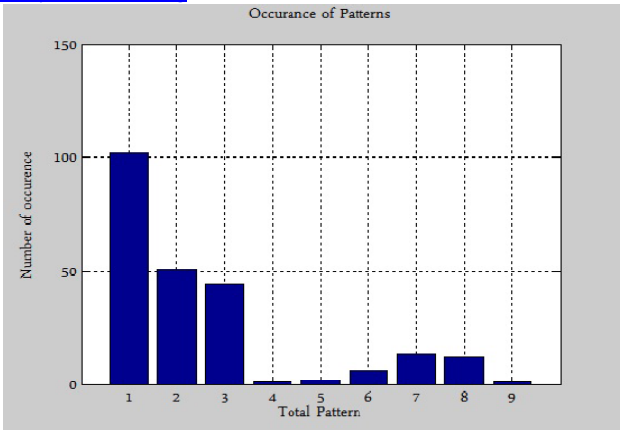

Fig .2. Histogram of 9 weather patterns

In the proposed method as the initial cluster formation is done on the training data selected using k means algorithm and Euclidean distance, further each instance of the testing data is selected and each numeric value of the attribute compared with the distance calculated with the centroids of the initial clusters. The similarity distance threshold is set for attributes values as well as attributes count for each initially formed cluster again. If the new instance satisfies the distance thresholds, it will be added to one of the initial clusters according to the distance with centroids. If not, the new cluster is formed. This process continues till the end of testing data available.

E. Adaptive Clustering method:

1. Specify No of new clusters other than initial clusters.
2. Choose each new instance from the testing data as an input
3. Set the different similarity threshold for distance with the centroids and attributes count.
4. Compare the new instance with existing clusters according to the distance similarity threshold.
5. If satisfies the similarity with the centroids of one of the initial clusters, add to that particular cluster.
6. If doesn't satisfies the similarity, a new pattern exists
7. till the end of data.

IV. RESULTS AND DISCUSSIONS

In this paper, we have proposed an efficient rare pattern detection method which generates histograms, the resultant histograms generated by this method is shown in Fig 2 below where pattern 4,5and 9 are clearly rare patterns as occurred less frequently.

By applying different distance thresholds we can generate different histograms which will be useful to detect weather patterns in more specified manner. Here we are showing different types of histograms on the basis of different distances thresholds set. The results of setting different thresholds generate histograms with more specified manner so as to detect detailed rare weather patterns. The results are shown in Fig 3 and Fig 4 below:

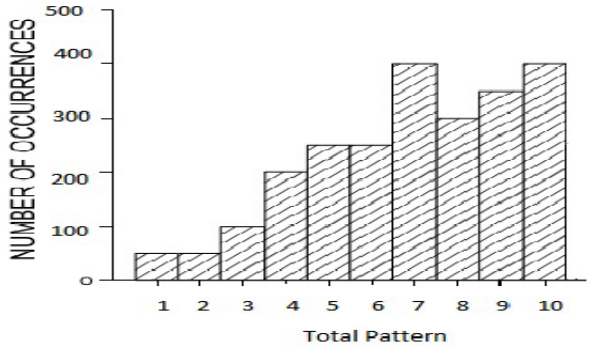

Fig .3. Histogram of 10 weather patterns

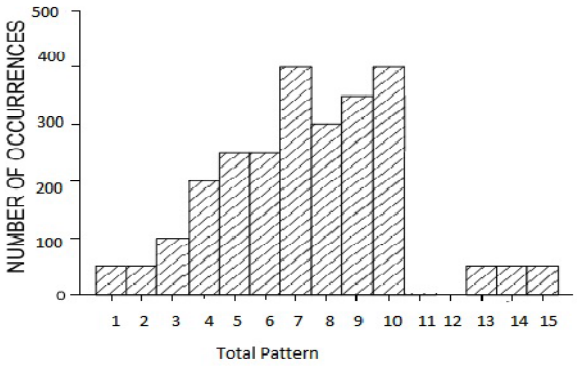

Fig .4. Histogram of 15 weather patterns

V. CONCLUSION AND SUGGESTIONS FOR FUTURE SCOPE

In this study, we proposed an adaptive clustering pattern detection method for disclosing weather patterns. The proposed method is able to deal with all meteorological or climate variables to detect hidden weather patterns. The proposed method is enabled to:

1. Detect weather patterns which are consistent with global climate change on weather patterns,
2. Detect rare weather patterns in greater detail by adjusting the resolution of the proposed detection model.

This method presents an efficient data processing technique used to enable the available data proper for the knowledge discovery as if the available metrological data is not proper the detection can be improper which leads to improper prediction of climate change. The goal of the proposed method is to implement an effective, efficient and adaptive rare pattern detection technique consisting of four parts missing data recognition, data preprocessing, initial clustering and adaptive clustering for dynamic detection. This method is implemented which uses metrological data in the form of daily summary, collected from NOAA, taken as input in excel sheet. The result is generated in the form of histogram. The results reveal the different weather patterns existing depending on the analysis of historical metrological data. With the help of these histograms we can easily identify the occurrence of rare patterns in weather which helps us to predict future environmental changes in the climate.

REFERENCES

[1] UNEP, *"United Nations Environment Programme, Climate Change,* "Available from: http://www.unep.org/climatechange; [cited 8 March 2013].

[2] Y. Meng, M. Dunham, F. Marchetti, and J. Huang, "*Rare event detection in a spatiotemporal environment,"* Proceedings of the Second IEEE International Conference on Granular Computing (GrC'06), pp. 10–12, 2006.

[3] Z. Yang, N. Meratnia, and P. Havinga, *"Outlier Detection Techniques for Wireless Sensor Networks: A Survey,"* IEEE Communications Surveys& Tutorials, vol. 12, no. 2, pp. 159–170, 2010.

[4] P. Gogoi, D. K. Bhattacharyya, B. Borah, and J. K. Kalita, "*A Survey of Outlier Detection Methods in Network Anomaly Identification," The Computer Journal*, vol. 54, no. 4, pp. 570–588, Mar. 2011.

[5] Z. Wang, C. S. Chang, and Y. Zhang, *"A feature based frequency domain analysis algorithm for fault detection of induction motors," Industrial Electronics and Applications (ICIEA), 2011 6th IEEE Conference on*, pp. 27–32, 2011.

[6] V. Chandola, A. Banerjee, and V. Kumar, *"Anomaly detection: A survey," ACM Computing Surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.

[7] R. M. Li, *Statistical Analysis of the Spatio-Temporal Variability of the Urban Heat Island in Singapore, Honours Thesis.* Department of Geography, National University of Singapore., 2009.

[8] [8] W. Chow and M. Roth, "*Temporal dynamics of the urban heat island of Singapore," International Journal of Climatology*, vol. 26, no. 15, pp. 2243–2260, 2006.

[9] [9] K. L. Ebi, N. D. Lewis, and C. Corvalan, *"Climate Variability and Change and their Potential Health Effects in Small Island States: Information for Adaptation Planning in the Health Sector," Environmental Health Perspectives*, vol. 114, no. 12, pp. 1957–1963, 2006.

[10] NEA, *"National Environmental Agency,Weather Wise Singapore," Available from:http://www.app2.NEA.gov.sg/data/cmsresource/2; [cited 29 July 2011]*.

[11] IPCC, *"Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change,"* Cambridge University Press, 2007.

[12] J. Harger, *"Air-temperature variations and ENSO effects in Indonesia, the Philippines and El Salvador. ENSO patterns and changes from 18661993,"* Atmospheric Environment, vol. 29, no. 16, pp. 1919–1942, Aug. 1995.

[13] NOAA, *"El Nino and La Nina, Climate Prediction Center,"* Available from: http://www.cpc.ncep.noaa.gov/products; [cited 5 July 2011].

[14] G. Stern, R. W. Macdonald, P. M. Outridge, S. Wilson, J. Ch´etelat, A. Cole, H. Hintelmann, L. L. Loseto, A. Steffen, F. Wang, and C. Zdanowicz, *"How does climate change influence Arctic mercury?"* Science of the total environment, vol. 414, pp. 22–42, Jan. 2012.

[15] S. Kovats, K. Ebi, B. Menne, D. Campbell-Lendrum, and U. N. E. Programme, *Methods of assessing human health vulnerability and public health adaptation to climate change.* Regional Office for Europe, World Health Organization, 2003, vol. 1, no. 1.

[16] F. Giorgi and E. Coppola, *"Does the model regional bias affect the projected regional climate change? An analysis of global model projections,"* Climatic Change, vol. 100, no. 3-4, pp. 787–795, May 2010.

[17] K. S. Jun, E.-S. Chung, J.-Y. Sung, and K. S. Lee, *"Development of spatial water resources vulnerability index considering climate change impacts."* Science of the total environment, vol. 409, no. 24, pp. 5228– 42, Nov. 2011.

[18] A. Augustsson, M. Filipsson, T. Oberg, and B. Bergb¨ack, *"Climate change - An uncertainty factor in risk analysis of contaminated land."* Science of the total environment, vol. 409, no. 22, pp. 4693–700, Oct. 2011.

[19] S. K. Oni, M. N. Futter, L. a. Molot, and P. J. Dillon, *"Modelling the long term impact of climate change on the carbon budget of Lake Simcoe, Ontario using*

INCA-C.” Science of the total environment, vol. 414, pp. 387–403, Jan. 2012.

[20] C. Tisseuil, M. Vrac, G. Grenouillet, a. J. Wade, M. Gevrey, T. Oberdorff, J.-B. Grodwohl, and S. Lek, *“Strengthening the link between climate, hydrological and species distribution modeling to assess the impacts of climate change on freshwater biodiversity.”* Science of the total environment, vol. 424, pp. 193–201, Mar. 2012.

[21] Meghali A. Kalyankar, Prof. S. J. Alaspurkar, *“Data Mining Technique to Analyse the Metrological Data".* International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, pp. 114-118, Feb 2013.

[22] *“WMO, NNDC Climate Data Online, National Climatic Data Center, NESDIS, NOAA".* Available from: http://www7.ncdc.noaa.gov/CDO/dataproduct; [cited 5 July 2011].

[23] Lakshminarayan K., S. Harp & T. Samad, *“Imputation of Missing Data in Industrial Databases".* Applied Intelligence 11, 259-275, 1999.

[24] Zhaoxia WANG, Gary LEE, Hoong Maeng CHAN,Reuben LI, Xiuju FU and Rick GOH, Pauline AWPoh Kimand Martin L. HIBBERD, Hoong Chor CHIN, *“Disclosing climate change patterns using an adaptive Markov chain pattern detection method".* International Conference on Social Intelligence and Technology, 2013.