Recognizing Rainfall Pattern for Pakistan using Computational Intelligence

M.Ali Aun, Abdul Ghani, M.Azeem, M. Adnan

Department of Computer Science,
University of Agriculture,
Faisalabad, Pakistan

M. Ahsan Latif

U.S.-Pakistan Center for Advanced Studies in Agriculture &
Food Security, Pakistan
Department of Computer Science, University of
Agriculture–Faisalabad, Pakistan

Abstract—**Over the world, rainfall patterns and seasons are shifting in new directions due to global warming. In the case of Pakistan, unusual rainfall events may outcome with droughts, floods and other natural disasters along with disturbance of economy, so the scientific understating of rainfall patterns will be very helpful to water management and for the economy. In this paper, we have attempted to recognize rainfall patterns over selected regions of the Pakistan. All the time series data of metrological stations are taken from the PMD (Pakistan Meteorological Department). Using PCA (Principal Component Analysis), monthly metrological observations of all the stations in Punjab have been analyzed which covers the area of 205,344 km² and includes monsoon-dominated regions. To tackle the problem of inter-annual variations, trend detection, and seasonality, rainfall data of Lahore, the Pakistan is taken that covers the period of 1976-2006. To obtain results, MASH (Moving Average over Shifting Horizon), PCA (Principal Component Analysis) along with other supporting techniques like bi-plots, the Pair-wise correlation has been applied. The results of this study successfully show seasonal patterns, variations and hidden information in complex precipitation data structure.**

Keywords—Rainfall patterns; trend detection; time-series analysis; principal component analysis; box-plot; moving average over shifting horizon; inter-annual variability

I. INTRODUCTION

In the previous two decades, the topic of weather change has appeared strongly over global in perspective to its expected inferences on the atmosphere. Extreme weather events and disturbed patterns of rainfall outcomes with the uncertain behavior of the seasonal phenomenon in many areas over the world due to global warming [17] that affects water resources and agriculture directly [14]. These deviations in the seasonal patterns are noteworthy in their consequences. In case of Pakistan, economy and human life significantly depend upon the seasonal behaviors [6]. Moreover, national task force on climate change (2010) reports many droughts, floods, earthquakes, natural disasters and heavy rainfall events in this country [4]. Recognition of rainfall shapes has gained importance and interest during last decades to identify changes in the seasonal phenomenon [13]. Principal Component Analysis (PCA) is a well-known technique of multi-dimensional scaling having valuable properties such as expository analysis and computational simplicity [8]. Moreover, this is especially helpful to recognize unknown nature of component patterns [15]. It retains maximum

information by the linear transformation that converts high dimensional variables into low dimensional space i.e. dimension reduction [19] where new coordinates known as principal components. PCA is a standard analysis technique and used in many fields to clarify total variance in the dataset [9].

MASH (Moving Average over Shifting Horizon) is an innovative technique to identify patterns in time-series data by visualizing the variabilities via graphical representation based on EDA (Exploratory data analysis) [1]. Graphical representation of time-series dataset can be helpful in interpreting, identifying the patterns [7] and associations among data [16]. In general, MASH aims to determine sequential variation in a seasonal pattern based on metrological observations over time. Rainfall data can be monthly, daily, yearly. In our case, statistic measurement is the mean, but one can be used others like median as well. Results of MASH make possible and easier to investigate about seasonal patterns along with trends that can be detected by visual inspection [1]. This paper is arranged as follows: This part contains the importance of rainfall and a short introduction of methods used in this study. The following section includes the picture of the study area. The material and methods part incorporate the techniques that are used to reveal patterns in rainfall of selected areas, while the section of results & discussion shows the major findings of this study. In the last SECTION, the conclusions of the research are presented in Software testing.

II. STUDY AREA

The Pakistan is positioned at 23°, 37° north latitudes, and 61°, 76° east longitude having a large variety in a seasonal and spatial variation of the weather. Coastal areas of the country are located along with the Arabian Sea and have little rainfall and extreme warm seasons and; its western areas are waterless and very hot cover with deserts. The northern areas have heavy rainfall and very low temperature. Moreover, the East-South areas have very low rainfall and remain very hot in the monsoon [18]. To analyze the behavior's and patterns of rainfall in the Punjab Pakistan, five years' monthly precipitation observations of different stations have been taken which covers the period of 2005-2010. All the metrological stations that are selected for the study have their own significance due to their geographical location. Selected stations are listed in Table 1.

TABLE I.   METROLOGICAL STATIONS

Station	Latitude (ºN)	Longitude (ºE)	Elevation (a.m.sl)
Islamabad	33.71	73.06	622
Lahore	31.54	74.22	214
Jhelum	32.93	73.73	287
Sialkot	32.31	74.32	255
Jhang	31.16	72.19	159
Sargodha	32.03	72.40	187
Multan	30.12	71.26	121.95
Mianwali	32.58554	71.5436	210
Khanpur	28.39	70.41	88.41
Faisalabad	31.26	73.08	185.6
Bahawalpur	29.2	71.49	110
Muree	33.9	73.39	2291
Bahawalngar	29.2	73.51	161.05

III. LITERATURE REVIEW

Ellouze et al. [5] made an investigation to recognize patterns of rainfall along with temporal and spatial variability in precipitation dataset for the area of southern Tunisia. Rainfall data was covering the period of 1930 to 2000. To discover governing variables related to their variability and the nature of precipitation distribution, they applied Principal Component Analysis (PCA) on metrological observations recorded in 12 stations. They found first three PCs (principal components) that were significant to explain the total variance of 90%. Rainfall variability was exposed to be dependent on the seasonal situation. Annual precipitation was significant over the area of south-eastern. [11]. This study was conducted by applying PCA (Principal Component Analysis) with the aim of identifying daily precipitation pattern over a period of twenty years i.e. 1994 to 2013 using records from eighty-nine (89) metrological stations located throughout the country of Malaysia. Six principal components were retained by using principal component analysis with the whole variance of 53.43 percent. The 1st and the 2nd component incorporated the areas that were showing attributes of south-west and north-east monsoon seasons respectively. The 4th principal component was covering the northern areas of the peninsular Malaysia along with two extreme points in precipitation amount occurred per year. They analyze the difference between regions in 3rd, 5th, and 6th principal components. At the end of the study, they suggest that PCA is a suitable technique to reduce the dimension in complexity dataset.

Alkan et al. [2] determine the precipitation patterns of monthly rainfall data of Turkey by computing PCA (Principal Component Analysis) biplot. To extract patterns of seasonal rainfall, they use rainfall data which was covering the period of 1970-2010 recorded from 81 metrological stations. Principal component analysis bi-plot was applied to inspect the relationship multidimensional variations between the metrological variables. The conclusion of this research revealed that bi-plots and PCA can be helpful graphical techniques to monitor rainfall [2].

Anghileri et al. [1] proposed and applied an approach named MASH (Moving Average over Shifting Horizon) based on EDA (Exploratory Data Analysis), which reveals the patterns of rainfall by computing daily precipitation observations for the time period of 1974-2010. They find out the inter-annual variability and seasonality. The results of this study were showing many important trends in considered time horizon. The results obtained by MASH were successfully visualized to observe seasonal behavior and to detect precipitation trends via visual examination.

IV. MATERIAL AND METHODS

Principle component analysis (PCA) is a statistical procedure and commonly used in various modern data processing fields [10]. The general algorithm for performing a principal component analysis is below.

➤ Consider the total data-set having d-dimensional observations and ignore the data labels.

➤ For the whole dataset, compute the means for each dimension i.e. (Calculate the mean vector of d-dimensional)

➤ Compute the covariance matrix of the whole dataset.

➤ Calculate the eigenvectors and corresponding eigenvalues.

➤ The computed eigenvectors are now orthogonal and can be used to project the actual data into the new coordinate system. The projection of actual data by a matrix of eigenvectors reveals the PCs (principal components) Y. [3].

Moreover, this calculation can be done by the equation i.e. $y = W^T \times x$. (where x is a d×1-dimensional vector representing one sample, and y is the transformed k×1-dimensional sample in the new subspace.)

We have applied Moving Average over Shifting Horizon (MASH) technique to explore rainfall patterns while handling the subject of inter-annual variability and seasonality in rainfall data. The objective of this technique is to evaluate changes in the seasonal pattern. In this method, seasonal patterns are represented through 12 values of the average monthly flow over the year. While averaging, rainfall data over successive months in the same year and over same months in successive years will consider. On the other hand, the horizon of successive years is increasingly shifted-ahead to take into account any pattern to develop. Thus, MASH is a matrix and is shown below

$$MASH = \begin{pmatrix} \mu1,1 & \mu1,2 & \mu1,3 & ... & \mu1,Nh \\ \mu2,1 & \mu2,2 & \mu2,3 & ... & \mu2,Nh \\ \mu3,1 & \mu3,2 & \mu3,3 & ... & \mu3,Nh \\ ... & ... & ... & ... & ... \\ \mu12,1 & \mu12,2 & \mu12,3 & ... & \mu12,Nh \end{pmatrix}$$

Columns in above MASH matrix are the mean of seasonal flow patterns that are calculated over dissimilar horizons (Nh). More accurately µt,h denotes the average monthly flow on the time of t-th i.e. month of the year in horizon (h-th) and calculated as in equation below

$$\mu_{t,h} = \underset{y\in[h,h+Y-1]}{\text{mean}} \left[\underset{d\in[t-w,t+w]}{\text{mean}} x_{d,y} \right]$$

In this equation, Xd,y denotes the inflow over specific d-th month of y-th year with respect to time series. Y is equal to total length of years of the shifting horizon, 1+2w is a scalar figure of months. Nh is actually total number of horizons and associated with Y i.e. (Nh = Ny − Y + 1) where Ny is total number of years in actual time-series data.

V. RESULTS AND DISCUSSION

Box-plotting or whisker-plotting is a commonly used statistical approach that provides graphic representation of distribution and patterns placed in quantitative data [12].
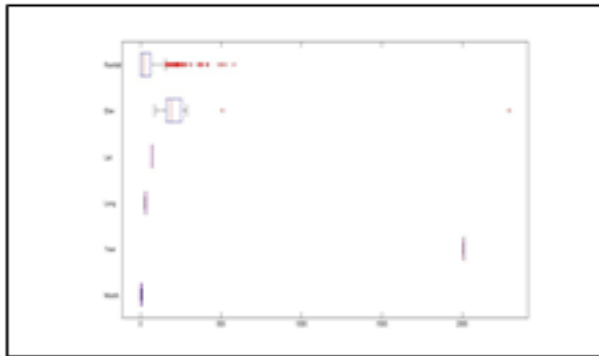

Fig. 1. Boxplot to analyze the distribution of rainfall data of the Punjab.

In the above Fig. 1, can see from a red dotted area that there is more variability in the variables of the Rainfall along with elevation than in other variables.

Before computing principle components of rainfall dataset, the pair-wise correlation between pairs of all variables has been checked and it was as high as 1 and 0.95. This shows the high correlation among some variables. However, PCA develops new independent variables which are the linear combination of actual variables.

PCA computes the rainfall dataset variables into scores known as component scores that are orthogonal (can be plotted on 2D graph).
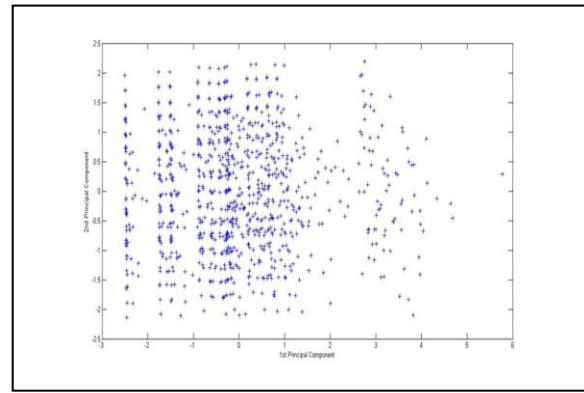

Fig. 2. Plot of the first two columns of scores generated.

Principle component analysis actually computes the all component scores to have mean zero. Above Fig. 2, plot shows the centred and scaled precipitation data plotted onto the initial stwo PC's i.e. PC1 and PC2.

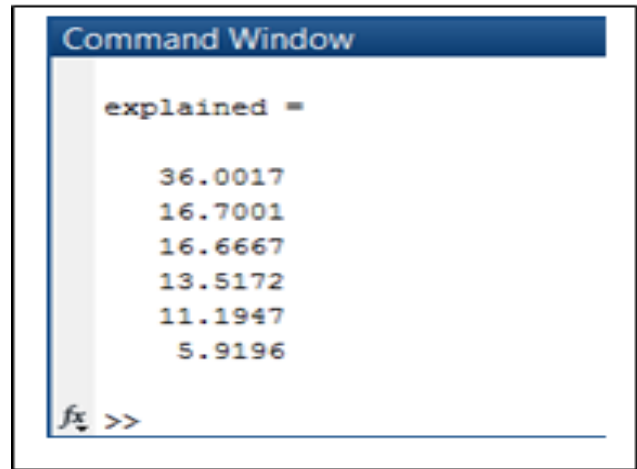Below is a vector that contains the values of variance in percentage by corresponding six PC's.

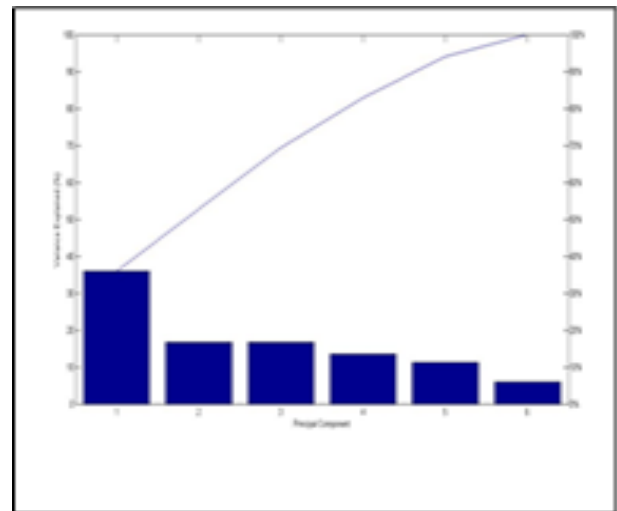Fig. 3. Observed variance in first six PC's.

Fig. 4. Variance in PC1 to PC6.

The above Fig. 3 shows variance (in percentage) with respect to PC1, PC2, PC3, PC4, PC4, PC5 and PC6. From this one can analyze that how much variance in the dataset is explained by which PC (as a bar) and how much variance is explained by the first 6 PCs. In Fig. 4, plot shows the six components that explain 100% of the total variance. The only clear break in the amount of variance accounted for by each component is between the first and second components. However, the first component by itself explains less than 40% of the variance. We can see that the first three principal components explain roughly two-thirds of the total variability in the standardized ratings, so that might be a reasonable way to reduce the dimensions. By removing the PCs that contribute little to the variance, we project the entire Data-set to a lower dimensional space but retain most of the information. So, we will take the PC1 and PC2 to visualize results. Hoteling's T2 is the last output that PCA gives and systematic approach to get most extreme points in the dataset. When we compute this value, it gives us index number of data that was locating the extreme point in rainfall dataset and from this, we observe that the rainfall values for Muree are the furthest from the average Punjab stations. All six variables are represented in Fig. 5 by a vector, and the direction and length of the vector indicate how each variable contributes to the two principal components in the plot. For the figure below, the first principal component, on the horizontal axis, has positive coefficients for all variables except Year (because year falls in negative side and have less effectiveness/variance and can be deduced in the PC1 to reduce dimensions). That is why the five vectors are directed into the right half of the plot. The major coefficients in the first principal component are the Longitude, elevation, and latitude having most positive effectiveness respectively. The second principal component has positive coefficients for rainfall to show his importance (positive effectiveness). Month variable has less positive variance and is difficult to see as it lies on the x-axis so we draw it in next 3D (Fig 6).
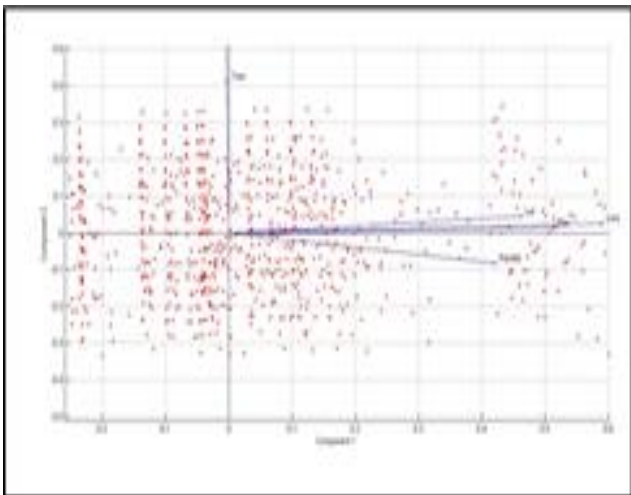

Fig. 5. 2D representation of PC1 vs. PC2.

This indicates that the first component distinguishes among metrological stations that have high values for the first set of variables and low for the second, and stations that have the opposite.
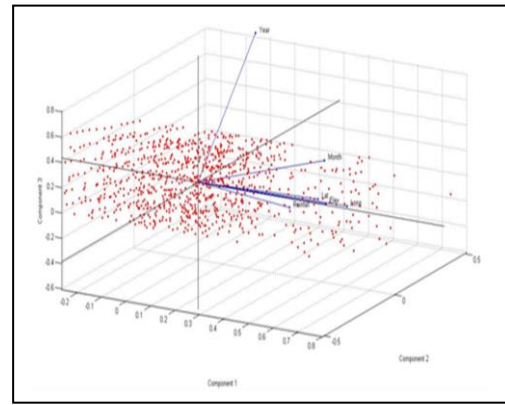

Fig. 6. 3D representation of PC1, PC2 and PC3.

The MASH technique has two numbers of tuning parameters used for averaging i.e. a number of months along with years. As other smoothing methods, there is also no general rule to fix these parameters.
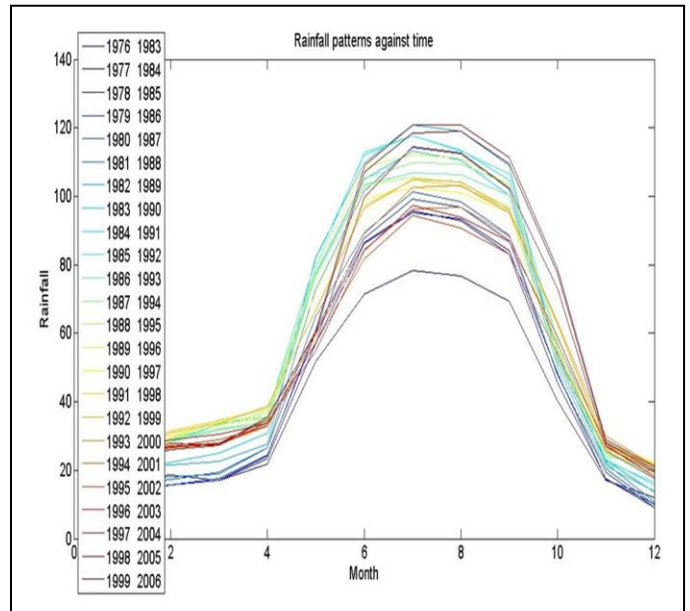

Fig. 7. Graphical representation of plotted MASH of the historical monthly inflows.

Above Fig. 7 provides a graphical illustration of a MASH of monthly inflows at the values of w = 2 months, Y = 7 years over the time horizon of 1976-2006. The first line in the color bar is actually the moving average calculated over the specific horizon of 1976-1983 and the second line represents the moving average computed over the period of 1977-1984, etc. Since the actual time-series of our monthly rainfall data covers a period of N_y = 30 years i.e. from 1976 to 2006 but the MASH is made of N_h = 24 flow monthly seasonal patterns due to the selected value of shifting horizon. In above figure, latest horizons are plotted through red-colored lines and older horizons are plotted through blue colored lines. In addition, above figure shows the rainfall phenomenon over 12 months along with variations in seasonal rainfall among different time horizons.
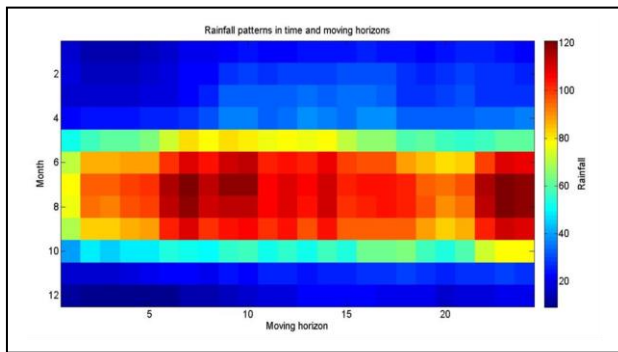
Fig. 8.   Visual Inspection of MASH: Rainfall patterns of hydrological
seasons.

The above Fig. 8 gives an informative and very concise representation of rainfall patterns along with variations during distinct hydrological seasons. The cool dry season has enlarged to 3rd month i.e. March, Monsoon season having red colored area starting from 6th and ending at 9th month comes with disturbed patterns and unusually heavy rainfall events. Moreover, one can also inspect the different seasons of the selected region i.e. Lahore, the Pakistan. These results were obtained using w = 2 which filter out the variation between months and Y = 7 which filter out the variability year to year.

VI. CONCLUSION

The study identifies the spatial and temporal characteristics of possible physical significance.

In this study, rainfall patterns along with seasonality and inter-annual variability's over selected regions of the Pakistan have been successfully extracted from time series data using PCA (Principle component analysis) and MASH (Moving average over shifting horizon). PCA has been put into work for the extraction of rainfall patterns in the Punjab, Pakistan. Using PCA, variability in the monthly rainfall of Jhelum, Sialkot, Jhang, Sargodha, Multan, Lahore, Mianwali, Khanpur, Islamabad, Faisalabad, Bahawalpur, Muree, and Bahawalnagar was examined. The supporting technique of PCA i.e. box-plot reveals that variability in the variables of the Rainfall and elevation was more than in other variables. Pair-wise correlation between pairs of all variables has been checked and it was as high as 1 and 0.95. Rainfall values for Muree were the furthest from the average Punjab stations. Six principal components were computed to check the Variance and first two PC's was considered to obtain results as they were explaining two-thirds of the total variability. The first eight principal component patterns explain for 96.70% of the total variance. The first principal component was showing positive coefficients for all variables except Year. The analysis of Lahore, Pakistan reveals some significant trends of rainfall with disturbed patterns and heavy rainfall events in a monsoon for the period of 1976-2006: Moreover, different seasons of the selected region have been projected. The results of this investigation suggest that PCA and MASH can be very useful techniques to inspect metrological data as well as for rainwater management.

REFERENCES

[1]   D. Anghileri, F. Pianosi, and R. Soncini-Sessa, "Trend detection in seasonal data: from hydrology to water resources," Journal of Hydrology, vol. 511, pp. 171-179, 2014.

[2]   B. B. Alkan, C. Atakan, and Y. Akdi, "Visual Analysis Using Biplot Techniques of Rainfall Changes over Turkey," MAPAN, vol. 30(1), pp. 25-30, 2015

[3]   D. Boardman, M. Reinhard, and A. Flynn, "Principal component analysis of gamma-ray spectra for radiation portal monitors," Nuclear Science, IEEE Transactions on, vol. 59(1) pp. 154-160, 2012.

[4]   M. A. Chaudhry, and G. Rasul, "The effect of Eurasian snow cover on the monsoon rainfall of Pakistan," Pakistan Journal of Meteorology, vol. 5:1-11, 2013.Ellouze, M., Azri, C., & Abida, H. (2009). Spatial variability of monthly and annual rainfall data over Southern Tunisia. Atmospheric Research, 93(4), 832-839.

[5]   M. Ellouze, C. Azri, and H. Abida," Spatial variability of monthly and annual rainfall data over Southern Tunisia," Atmospheric Research, vol. 93(4) pp. 832-839, 2009.

[6]   A. Imran, Q. Zaman, G. Rasul, and A. Mahmood, "An Analytical Study of Variations in the Monsoon Patterns over Pakistan," Pakistan Journal of Meteorology, Vol. 10(20), 2014.

[7]   B. W. Lane, "A time-series analysis of gasoline prices and public transportation in US metropolitan areas," Journal of Transport Geography, vol.22, pp.221–235, 2012.

[8]   M. Mahrooghy, N. H. Younan, V. G. Anantharaj and J. V. Aanstoos, "Enhancement of Satellite Precipitation Estimation via Unsupervised Dimensionality Reduction," Geoscience and Remote Sensing, IEEE Transactions on, vol. 50(10)pp. 3931-3940, 2012.

[9]   Z. A. Mohtar, A. S. Yahaya, F. Ahmad, S. Suri, and M.H. Halim, "Trends for daily rainfall in northern and southern region of peninsular Malaysia," Journal of Civil Engineering Research, vol. 4(3A), pp. 222-227, 2014.

[10] S. Ouyang, and Z. Bao, "Fast principal component extraction by a weighted information criterion," Signal Processing, IEEE Transactions on, vol. 50(8), pp. 1994-2002, 2002.

[11] M. Othman, Z. H. Ash'aari, and N. D. Mohamad, "Long-term Daily Rainfall Pattern Recognition: Application of Principal Component Analysis," Procedia Environmental Sciences, vol. 30, pp. 127-132,2015.

[12] M. Rong, S. Jian-sheng, and L. Ji-chao, "Assessing groundwater quality using the fuzzy multivariate statistical method," In Water Resource and Environmental Protection (ISWREP), 2011 International Symposium on IEEE Vol. 1, pp. 57-65., May 2011.

[13] C. Roug´e, Y. Ge, X. Cai, "Detecting gradual and abrupt changes in hydrological records.

[14] Advances in Water Resources, vol. 53, pp.33–44, 2013.

[15] C. V. Singh, "Pattern characteristics of Indian monsoon rainfall using principal component analysis (PCA)," Atmospheric research, vol. 79(3), pp.317-326, 2006.

[16] Y. Sadahiro, and T. Kobayashi, "Exploratory analysis of time series data: Detection of partial similarities, clustering, and visualization," Computers, Environment and Urban Systems, vol. 45, pp. 24-33, 2014

[17] S. Salma, S. Rehman, and M. A. Shah, "Rainfall trends in different climate zones of Pakistan," Pakistan Journal of Meteorology, Pakistan, 2012

[18] G. Spock, and I. " Hussain Spatial sampling design based on convex design ideas and using external drift variables for a rainfall monitoring network in Pakistan," Statistical Methodology, vol. 9(1),pp. 195-210,2012.

[19] H. Xu, C. Caramanis, and S. Sanghavi, "Robust PCA via outlier pursuit," In Advances in Neural Information Processing Systems, pp. 2496-2504, 2010.