Discovering Flood Recession Pattern in Hydrological Time Series Data Mining during the Post Monsoon Period

Satanand Mishra
IT & WM&RT Group,
CSIR-Advanced Material Process &Research Institute,
Bhopal- 462064, India

C. Saravanan,
Computer Centre
National Institute of Technology,
Durgapur-713209, India

V.K.Dwivedi,
Dept of Civil Engineering,
National Institute of Technology,
Durgapur-713209, India

K. K. Pathak
Dept of Civil and Environmental Engineering,
NITTTR, Bhopal-462002, India

ABSTRACT

This paper examines the flood recession pattern for the river discharge data in the river Brahmaputra basin. The months from October to December comes under the post monsoon season. In this paper, with the help of time series data mining techniques, the analysis has made for hydrological daily discharge time series data, measured at the Panchratna station during the post monsoon in the river Brahmaputra under Brahmaputra and Barak Basin Organization after the high flood. Statistical analysis has made for standardization of data. K-means clustering, Dynamic Time Warping(DTW), Agglomerative Hierarchical Clustering(AHC) and Ward's criterion are used to cluster and discover the discharge patterns in terms of the autoregressive model. A forecast model has been developed for the discharge process. For validation of the recession pattern, Gauge–Discharge Curve, Water Label Hydrographs, Rainfall Bar Graphs have been developed and also discharge recession coefficient has been calculated. This study gives the behavioral characteristics of rivers discharge during recession of high floods with the time series data mining.

Keywords

Clustering; agglomerative hierarchical clustering; data mining; runoff; hydrological time series; pattern discovery; post monsoon; recession patern; similarity search ; Ward criterion.

1. INTRODUCTION

In India, the months from October to December comes under the Post-Monsoon season. During these months, there occurs a different monsoon cycle called the North-east monsoon which brings dry and cool air masses. In the months of October, the south west monsoons begin to decrease, climate begins to be drier and the precipitation also decreases due to winter. The North East monsoons carry winds that have already lost their moisture while travelling across Central Asia. In the mean time, North-East India receives minor precipitation from the north east monsoon.

The Central Water Commission (CWC) is engaged on surface water management. CWC has its own networks for observation of water level, rainfall, discharge, sediment, evaporation, temperature and water quality data. Meteorological department is engaged in climatic and whether forecasting services and observing rainfall, temperature, humidity etc. These data are very useful in research, historical trend analysis and future forecasting. With the development of database technology, various techniques of data analysis and knowledge extraction are being used for discovering knowledge from such collected data in various organizations like hydrological, environmental, earthquake, meteorological etc.

The ability to build a successful predictive model depends on past data. Data mining is subjected to learn from past success and failures and will be able to predict what will happen next (future prediction) [12]. Data mining, also popularly referred to as Knowledge Discovery from Database (KDD), is defined as "Discovery of comprehensible, important and previously unknown rules or anything that is useful and non-trivial or unexpected from our collected data [29]. Data mining is a multidisciplinary field, drawing work from areas including database technology, machine learning, statistics, pattern recognition, information retrieval, neural networks, knowledge-based systems, artificial intelligence, high-performance computing, and data visualization. We present techniques for the discovery of patterns hidden in large data sets, focusing on issues relating to their feasibility, usefulness, effectiveness, and scalability.

Today, the development of information technology has generated huge amount of databases; covering in various areas science and technology. Data mining is being vastly applied in research and business field. Finding association rules, sequential patterns, classification and clustering of data are typical tasks involved in the process of data mining. Mainly, data mining is an iterative process in which data have to be critically selected and cleaned; parameters of the mining algorithms are familiarised. Hydrological time series are sets of various record values of hydrological data that vary with time.

In the past decade, many researches brought in scientific community, an artificial intelligence technique is emerging trends for extracting knowledge from hidden historical data. Worldwide, water resources organisation is one of among which is using this technique for hydrological forecasting. Real time hydrological forecasting is a great challenge in scientific community. Hydrological forecast saves life, infrastructure, and economy of any country. This is the presumption of future happening after the natural hazards. In the field of hydrological data mining various researches and techniques carried out for extraction of knowledge from historical data. Some of them which are relevant for this study are : Pattern discovery in hydrological time series data mining during the monsoon period of the high flood years in Brahmaputra river basin [23], A novel approach to the similarity analysis of multivariate time series and its application in hydrological data mining [41], River flow time series using least squares support vector machines [33], Similarity search and pattern discovery in hydrological time series data mining [28], Flood pattern detection using sliding window technique [32], Runoff forecasting using fuzzy support vector regression [39], Forecasting monthly runoff using wavelet neural network model [3], Neural network model for hydrological forecasting based on multivariate phase space reconstruction ([38], Mid-short term daily runoff forecasting by ANNs and multiple process based hydrological models, Research and application of data mining for runoff forecasting [18], Computational methods for temporal pattern discovery in biomedical genomic databases [31], an efficient k-Means clustering algorithm: analysis and implementation [16], The prediction algorithm based on fuzzy logic using time series data mining methods [4] , and a forecast Model of Hydrologic Single, Element Medium and Long-Period Based on Rough Set Theory [34].

Time series data mining combines the fields of time series analysis and data mining techniques [4]. This method, creates a set of process that reveal hidden temporal patterns that are characteristics and predictive of time series consequences. The main goal of this study is to develop a data mining application using modern information technology and discover the hidden information or patterns behind the historical hydrological data during the post monsoon of the river Brahmaputra under the hydrological process. The TSDM tools like similarity search, k-means clustering, AHC, ARIMA Model are used here. The hydrographs, rainfall bar graphs, GD curve are validating the results of this research.

2. STUDY AREA AND DATA SET

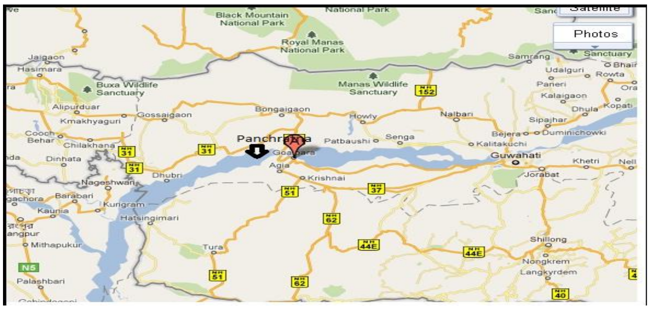


Fig 1: Google map view of the basin and catchment

The site Panchratna (Latitude 26o 11’ 55" and Longitude 900 34’ 38”) of the river Brahmaputra, located in the district of Goalpara in the state of Assam shown in Fig 1 is selected for the study. The length of the river upto the site is 2562 Km. The catchment area upto the site measures 468790 sq km. The site is located on the left bank of river. The type of site is HO (Hydrological Observation). For the study, daily discharge and water level data for the entire year were taken during the highly flooded years of 1988, 1991, 1998, 2004 and 2007. The average max temperature is 300C and minimum temperature is 150C, recorded during the post monsoon period in the Goalpara. Average humidity percentage is recorded as 82% during the months from October to December.

3. TIME SERIES DATA MINING

The time series data mining is commonly covered under classification, clustering, similarity analysis sequential pattern mining, forecasting, summarization, anomaly detection (Interestingness Detection), and segmentation. Runoff forecasting is a classical hydrological problem falls under the Hydrological time series analysis[36]. A new framework for analyzing time series data called Time Series Data Mining (TSDM) is introduced in this study. This framework adapts and innovates data mining concepts to analyzing time series data. In particular, it creates a set of methods that reveal hidden temporal patterns that are characteristic and predictive of time series events. When the time dimension added to real-world database, its produces Time Series Database (TSDB) and thus introduces a new aspects in the field of data mining and knowledge discovery. Time series data are the classes of data where the sequence of values changes during a period with time, for example the amount of sales, temperature changes, earthquake eruption, a patient’s heart rate changes, financial stock sales, and hydrological observations as river water level, river discharge in particular stations and so on [28,41]. Time series data mining algorithm can be used to predict continuous values of data. When the algorithm is skilled to predict a series of data, it can predict the output of the other series of data.

The algorithm generates a model that can predict trends, based on original dataset. New data can also be added that becomes part of the trend analysis. TSDM algorithms usage is restricted to cluster analysis, similarity search and pattern discovery. There are two main goals of time series analysis (i) identifying the nature of the phenomenon represented by the sequence of observations, and (ii) forecasting - future values of the time series variable. Both of these goals require that the pattern of observed time series data is identified and more or less formally described. Once the pattern is established, we can interpret and integrate it with other data. Regardless of the depth of our understanding and the validity of our interpretation of the phenomenon, we can extrapolate the identified pattern to predict future events [22, 23, 24].

3.1 Cluster analysis

Clustering is a job of assigning a set of objects into groups called clusters. Clustering is among one of the unsupervised learning methods. Its main goal is to identify structure in an unlabeled data set by objectively organizing data into homogeneous groups where the within- group-object similarity is minimized and the between-group-object dissimilarity is maximized [36]. Clustering is the process of grouping the data into classes or clusters so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. In recent years clustering of time series has received considerable attention because of its fundamental task in data mining. Clustering methods used in time series are K-means clustering, K-medoids clustering, nearest neighbour clustering, hierarchical clustering, self-organizing maps, and so on.

Among all clustering algorithms, K-means clustering is the most commonly used clustering algorithm [5] with the number of clusters K, specified by the user. K-means clustering is more useful for finding spherical-based clusters capability in small- to medium-sized databases [11]. K-means clustering algorithm applied as following steps. First, it selects k of the objects, each of which initially represents a cluster mean or centre. For each of the objects in the dataset that remain, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster means (Euclidean Distance or any other distance measures). As soon as the objects are assigned to the best fit cluster, the cluster means are updated. This process iterates until the cluster centres no longer make a movement [30].

3.2 Agglomerative hierarchal clustering

The another classical clustering algorithm is a hierarchical clustering method which generates nested hierarchy of similar groups of time series according to pair wise distance matrix of the series [17,2]. Hierarchical Clustering refers to the formation of a recursive clustering of the data point- a partition into two clusters, each of which is itself hierarchically clustered. It forms a cluster tree by grouping the data objects. Hierarchical methods can be classified as being either agglomerative or divisive, on the basis of how the hierarchical decomposition is being formed. The agglomerative approach or bottom–up approach starts by forming a separate group of each object. It successively merges the objects or groups close to one another, until all of the groups merge into one, or until a termination condition holds. Whereas in case of the divisive approach or top–down approach, starts with all the objects in the same cluster. Then a cluster is split up into smaller clusters, until eventually each object is in one cluster, or until a termination condition holds. The process involved in Agglomerative Hierarchical Clustering is as under -

1. Start with N clusters each containing a single entity, and an N × N symmetric matrix of distances (or similarities), Let dij = distance between item i and item j.
2. Search the distance matrix for the nearest pair clusters (i.e., the two clusters that are separated by the smallest distance). Denote the distance between these most similar clusters A and B by d_{AB}
3. Merge clusters A and B into new clusters labelled T. Update the entries in the distance matrix by

   (a). Deleting the rows and columns corresponding to clusters A and B, and

   (b). Adding a row and column giving the distances between the new cluster T and all the remaining clusters.
4. Repeat steps (2) and (3) a total of N-1 times.

AHC is illustrated using the bottom – up strategy to the data set having 5 objects {a, b, c, d, e}[23]. In this process, initially, cluster works as a cluster of its own. Clusters, which distance is minimum are merged with other cluster. Here the distance among all clusters is measured using average link approach.
Let us consider the cluster C_i and $C_{j,,}$ the distance between these two clusters is measured using average –link with the following formula-

$$d(C_i,C_j) = \frac{1}{n_i\, n_j}\sum_{x\in C_i}\sum_{x'\in C_j}|x - x'| \qquad (1)$$

Where |x-x'| is the distance between elements x and x', n_i and n_j are the number of objects in cluster $C_{i.}$ and $C_{j,}$ respectively.

3.3 Ward's criterion

The two methods k-Means and Ward, add on each other in that clusters, are cautiously built with Ward agglomeration, whereas k-Means allows overcoming the inflexibility of the agglomeration process over individual entities by rearranging them. There is a limitation with this scheme of Ward agglomeration, like k-Means, which is a computationally intensive method. It is not applicable to large sets of data.
The Ward's agglomeration starts with singletons whose variance is zero and proceeds by combining those clusters that effect as small increase in the square error criterion as possible at each agglomeration step.

Let us take a partition S = {S_1,S_2,S_3,…..S_k} arrived at on an agglomeration steps. As per the ward's rule the distance between two clusters, S_A, S_B is defined as the increase in the value of k-means criterion W(S,c) at the partition obtained from S by merging them into S_AU $S_{B.}$ where centroids c = {c_1,c_2,c_3….c_k}. The square error criterion is given as

$$W(S,c) = \sum_{k=1}^{k}\sum_{i\in S_k} d(i,S_k) \qquad (2)$$

Let us consider, the two clusters S_f, and S_g merged both so that the resulting partition is S(f,g) concurrent with s except for the merge cluster S_fUS_g . Let the new centroid is C_{fUg} and also let the N_f and N_g are cardinalities of clusters S_f and S_g. Therefore,

$$C_{fUg} = (N_f c_f + N_g c_g)/(N_f+N_g) \qquad (3)$$

The value of square error criterion on partition S(f,g) is greater than W(S,c) and given by following equation as:

$$W(S_{fUg}, c_{fUg}) – W(S,c) = \frac{N_f\, N_g}{N_f+N_g}\sum_{v\in V}\left(c_{fv} - c_{gv}\right)^2$$

$$= \frac{N_f\, N_g}{N_f+N_g}\, d(c_f,c_g) \qquad (4)$$

Here, the squared Euclidean distance between the centroids of the merged clusters S_f and S_g weighted by a factor proportional to the product of cardinalities of the merged clusters [26].

The increase can be calculated which is called Ward distance between centroids of the two clusters. The squared Euclidean distance scaled by a factor whose numerator is the product of cardinalities of the clusters and denominator is the sum of them. The ward distance between singletons is exactly half the squared Euclidean distance between the corresponding entities. This justifies the use of Ward's agglomeration results to get fair initial setting for k-Means where k is preset.

3.4 Dynamic time warping (DTW) algorithm

The dynamic time warping (DTW) algorithm is used for comparing two time series. The distance between the two series is computed, after stretching, by summing the distances of individual aligned elements ([10]. DTW is an algorithm for measuring optimal similarity between two time data sequences [14,10].The time series data varies not only on the time amplitudes but also in terms of time progression as the hydrological processes may reveal with different rates in response to the different environmental conditions. A non-linear alignment produces a similar measure, allowing similar shapes to match even if they are out of phase in time axis. The sequences are "warped" non-linearly in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension. To find the best alignment between times sequences Q1& Q2 one needs to find the path through the grid. In respect of hydrological time series similarity measures, the DTW algorithm is consider as follow [7]:

Let Q1 and Q_2 be the two time series discharge sequences of length *m* and *n* respectively and given as

$$Q_1 = x_1,x_2,\ldots,x_i,\ldots,x_m \qquad (5)$$

$$Q_2 = y_1,y_2,\ldots,y_j,\ldots,y_n \qquad (6)$$

An m-by-*n* matrix is constructed using DTW aligning for these two sequences. The (i^{th}, j^{th}) element of the matrix contains the distance $d(x_i,y_j)$ between the two points x_i and y_j , called the Euclidean distance and represented by -

$$d(x_i,y_j) = (x_i - y_j)^2 \qquad (7)$$

Each matrix element (*i*, *j*) corresponds to the alignment between the points x_i and y_j. This is illustrated in Figure 3. A warping path, W, is a contiguous set of matrix elements that defines a mapping between Q_1 and Q_2. The k^{th} element of W is defined as $w_k = (i,j)_k$, so we have:

$W = w_1, w_2, \ldots, w_k, \ldots, w_K \qquad \max(m,n) \leq K < m+n-1$ (8)

The warping path is typically subjected to several constraints.

Boundary conditions: $w_1 = (1, 1)$ and $w_K = (m, n)$. Simply stated, this requires the warping path to start and finish in diagonally opposite corner cells of the matrix.

Continuity: Given $w_k = (a, b)$ then $w_{k-1} = (a', b')$, where $a - a' \leq 1$ and b-$b' \geq 1$. This restricts the allowable steps in the warping path to adjacent cells (including diagonally adjacent cells).

Monotonicity: Given $w_k = (a, b)$ then $w_{k-1} = (a', b')$, where a–$a' \geq 0$ and b-$b' \leq 0$. This forces the points in W to be monotonically spaced in time.

There are exponentially many warping paths that satisfy the above conditions, however, which minimize the warping cost taken here:

$$\mathrm{DTW}(Q_1,Q_2) = \min\{\frac{1}{k}\sqrt{\sum_{k=1}^{k} w_k \ \}} \qquad (9)$$

The K in the denominator is used to compensate for the fact that warping paths may have different lengths. This path can be found very efficiently using dynamic programming to evaluate the following recurrence which defines the cumulative distance $\propto(i,j)$ as the distance $d(i,j)$ found in the current cell and the minimum of the cumulative distances of the adjacent elements:

$$\propto(i,j) = d(x_i,y_j) + \min\{\ (i\text{-}1, j\text{-}1),\ (i\text{-}1, j),\ (i, j\text{-}1)\ \} \qquad (10)$$

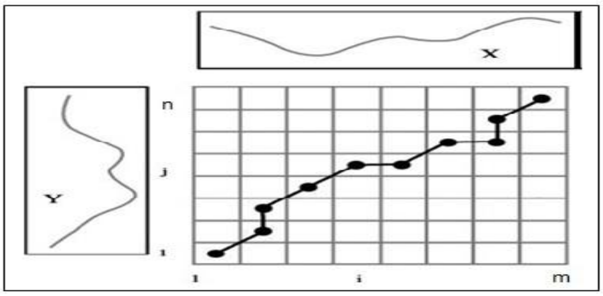


Fig2: DTW Warping Path

The Euclidean distance between two sequences is given as a special case of DTW, where the k^{th} element of W is constrained such that $w_k = (i,j)_k$, i = j = k. It is only possible where the two sequences have the same length. The time complexity of DTW is O(mn).

3.5 Similarity search

Euclidean distance is the most widely used distance measure for similarity search [1,6,8]. A similarity search finds data sequences in time series that differ only slightly from the given query sequence. It can be classified into two categories- i)Whole matching: In this kind of matching the time series data has to be of equal length and ii) Subsequence matching: In this mentioned category of matching we have a query sequence X and a longer sequence Y. The objective is to identify the subsequence in Y, beginning at Yi, which best matches X, and report its offset within Y [28]. For similarity analysis of hydrological time series data, Euclidian distance is typically used as a similarity measure. Given two sequences X = (x_1, . . . , x_n) and Y = (y_1, . . . , y_m) with m = n, their Euclidean distance is defined as follows:

$$\mathrm{D(X, Y)} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (11)$$

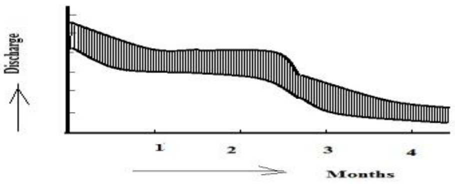


Fig 3 (a): Distance measure methods Euclidian

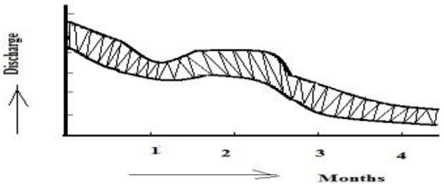


Fig 3 (b): Distance measure methods DTW

In this study, hydrological process for two discharge sequences have around the same shape, which are not aligned in the time axis. The Euclidean distance for the i^{th} point in first sequence is aligned with the i^{th} point in the another that produces a (10) pessimistic dissimilarity measures shown in Figure 4(a). DTW discovered the similar discharge process that is not aligned in the time axis which is non linear shown in Fig 4(b). Similarity search of the discharge time series process calculated DTW distance between every two discharge time series in each hydrological segmented period which is obtained from k-means clustering algorithm.

3.6 Pattern discovery

Discovery of patterns in data mining is a lucrative and highly demanding work. Data are sampled over time as X=X_1, X_2, X_3…X_t,..X_l (where l=length of data and the t denotes the sample). X is not independently and identically distributed. The X may come from different processes dependent on each other. Pattern discovery aims to find a subset of data from the available dataset, such that the subset represents a trend in the data. This trend when detected and modeled by an equation can be used in forecasting future responses of data. The problems that arise with detecting patterns are that, the data may contain multiple patterns, the data might be multidimensional, Even automated pattern discovery is difficult when the time series data is lengthy. In our case, discovered time series patterns, predict the future behavior of data that changes with time. This is a scope of trend analysis and prediction in time series data analysis [21].

3.7 Recession coefficient:

This study is based on post monsoon discharge TSDM. It is necessary to compute recession coefficient for validation of river discharge pattern after monsoon during the high floods years. In hydrology, the recession coefficient is usually expressed in the following exponential decay

$$\mathbf{Q = Q_0e^{-kt}} \qquad \mathbf{(12)}$$

Here, Q is the monthly average flow, Q_0 is the flow in the previous month and t is the time which is one month for the monthly data . The recession coefficient (k) is calculated from the monthly discharge data. Figure 4 depicts the recession flow of river Brahmaputra after monsoon. The value of k varies in a wide range according to the size and river course [25].

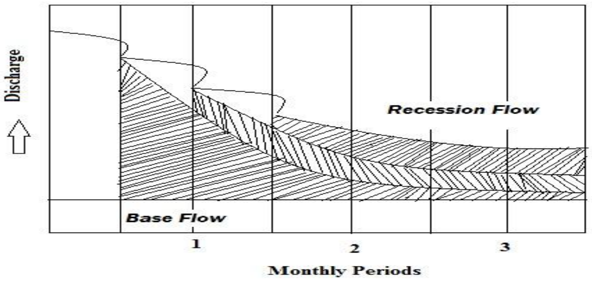


Fig 4: Component of River Brahmaputra Runoff Hydrograph

3.8 Auto regressive model (ARM)

An Auto Regressive Model is a forecast model. The autoregressive model is one of a group of linear prediction formulas that attempt to predict an output y[n] of a system based on the previous outputs (y[n-1],y[n-2]...) and inputs (x[n], x[n-1], x[n-2]...)

Deriving the linear prediction model involves determining the coefficients a1, a2, ...and b0, b1, b2,.. in the equation:

y[n] (estimated) = a1.y[n-1] + a2.y[n-2]... + b0.x[n] + b1.x[n-1] + ...) (13)

The remarkable similarity between the prediction formula and the difference equation used to describe discrete linear time invariant systems. Calculating a set of coefficients that give a good prediction y[n] is tantamount to determining what the system is, within the constraints of the order chosen. A model which depends only on the previous outputs of the system is called an autoregressive model (AR), while a model which depends only on the inputs to the system is called a moving average model (MA), and of course a model based on both inputs and outputs is an autoregressive-moving-average model (ARMA). By definition, the AR model has only poles while the MA model has only zeros. Several methods and algorithms exist for calculating the coefficients of the AR model [40].

4. DATA MINING IN HYDROLOGICAL TIME SERIES

The process of data mining in hydrological time series, for this study, is as follows:

1. Calculation of four statistical Characteristics eg. Qmean,Qmax,Qrange, and Qdev

2. Standardization of these characteristics using Z-scores methods

3. Clustering Monthly Discharge process using k –means clustering algorithms

4. Segmentation of hydrologic periods of the annual process

5. Use of dynamic time warping algorithm and detection of similarities in discharge process in Monsoon Period,

6. Application of agglomerative hierarchical clustering algorithms and Ward’S criterion to discover pattern of discharge process

7. Computing of recession coefficients

7. Establishing Gauge Discharge correlation

8. Analyzing the causal-effect relationship

9. Forecasting by Auto Regressive predictive model.

5. METHODOLOGY

5.1 Data Preparation and Segmentation

For the standardization of data, the four standard statistical characteristics (Q_{mean},Q_{max},Q_{range},Q_{dev}) has been calculated for each month in discharge data(Mishra et. al.,2013). The discharge data of river Brahmaputra has been taken for the years 1988, 1998, 2001, 2004, and 2007. In order to have an effective analysis of the data, the data were standardized using Z-scores technique so that the mean of the entire data range leads to 0 and the standard deviation is 1. The need for standardization was that to avoid affecting the study results by the wide variations in the data. The reason for such preference is that calculation of *z* requires, the discharge mean and the discharge standard deviation, not the sample mean or sample deviation.

The whole year data has been segmented in three periods, pre monsoon, monsoon and post monsoon and shown in . The graph shows the incline in discharge from the month of April in account of rainfall and it increases constantly as rainfall increases. In the mid monsoon periods, the graphs shows the peak discharge, accounting the heavy rainfall as reason. In the months of September the graphs shows decline due to decrease of rainfall which results in the recession of discharge, followed by Post-Monsoon period (Oct-Dec) [23]. The discharge also gradually changes with a decrease in temperature in September. The different climates cause different discharge processes, thus for a better study the discharge processes must be studied under same formation mechanism. In his paper, the Post Monsoon period is opted.

5.2 Similarity observation

For observation of similarities in the post monsoon discharge data of the high floods years 1988, 1998, 2001, 2004, and 2007 have been selected. The DTW search technique is applied for discovering similar discharge. This is because time series are expected to vary not only in terms of expression amplitudes, but also in terms of time progression, since flow of water may unfold with different rates in response to different natural conditions or within different locations in the basin in different times. A matrix (M) of size 5X5 is obtained, and the (ith, jth) element denotes the distance between the discharge processes for the ith and jth year.

For this study, a simulator named as DTW matrix viewer is designed which gives the DTW similarity matrix for the data as shown in [23]. User is allowed to select the corresponding hydrological period, which is obtained through previous work of clustering. The work of the simulator is to produce the DTW similarity matrix for the years with discharge data corresponding to the months in the hydrological period. The similarity matrix gave a comparison of the discharge processes in five years. In the above simulator, inputs are discharge data for years 1988, 1991,1998, 2004 and 2007 ranging from month of Oct-Dec. Repeating the process for one year against all other years for a period gives the distance values for discharge time series data for two years. The proper iteration and representation generates the matrix as given in Table 1.

Table 1 DTW Similarity Matrix for the Post Monsoon Period (Oct-Dec)

Year	1988	1991	1998	2004	2007
1988	0.0	5.74	4.43	1.30	3.72
1991	5.74	0.0	5.9	2.26	2.08
1998	4.43	5.9	0.0	4.62	1.36
2004	1.30	2.26	4.62	0.0	8.85
2007	3.72	2.08	1.36	8.85	0.0

The lowest value is for 1988 and 1998, which means in the hydrological periods they are the most similar years. Similarly, the matrices for other Hydrological Period can be obtained. For the Post Monsoon Oct-Dec the discharge pattern was similar in the years 1988-1998, 1998-2004, and 1991-2004.

On the basis of the above matrices the similarity graphs were plotted have an idea about the similar discharge processes in the corresponding two or three years are given in the Figure 5 and Figure 6.

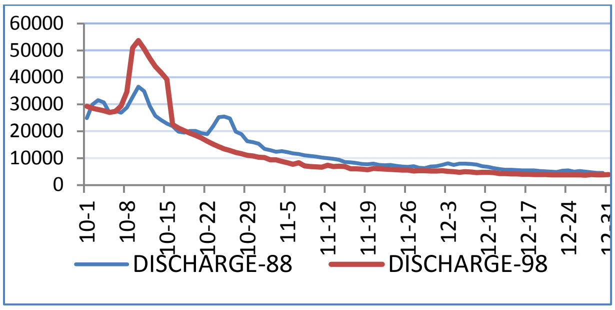


Fig 5: Similarity in discharge process of Post Monsoon (Oct-Dec) 1988, 1998

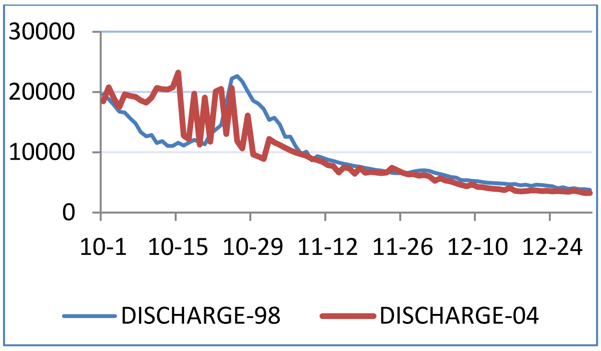


Fig 6: Similarity in discharge process of Post Monsoon (Oct-Dec) 1988, 2004

5.3 Pattern detection

The next step is to identify the discharge pattern from the corresponding discharge time series data. For this, each hydrological period obtained after the segmentation from the k-means clustering has been taken and then the discharge pattern in each of the periods has detected. The analysis is involved in the hierarchical clustering techniques and Wards criterion for the 5 years as the cases/samples and the attributes as observations of the average discharge data for the months which are in the hydrological period has done. For the analysis of patterns, post monsoon period data as the attribute have been taken for the period of 5 years (Oct-Dec).

AHC algorithm is particularly useful to find hidden patterns in the multidimensional data. As it is an unsupervised learning scheme, the number of clusters may be large or small at times. The lead role of AHC is to identify clusters or groups of related discharge time series natures that are similar to each other. Now the discharge time series data of the cluster center is the pattern because all other objects in a particular group then show similarity to the center only. Thus the cluster center can be taken as the pattern of discharge. On the basis of these five patterns, we have found the standard pattern for the discharge during the high floods. This is similar to all received pattern for a particular year.

In this analysis, the discharge pattern is cluster of a year into several clusters. But the year which formed the cluster center, formed the pattern with its discharge data for the months in the period. All other members (years) in the cluster attained membership of the cluster because there was similarity to the year representing the center, so they can be said to follow the pattern.

Now consider the center (the year) in the cluster and plotting of the discharge data of that year corresponding to daily discharge in the months, along the x-axis would give the pattern as shown Fig 8(a). In the Figure 8(b), the standard pattern has been detected during the high floods year at post monsoon periods. The standard pattern shows the future patterns of the floods during the post monsoon.

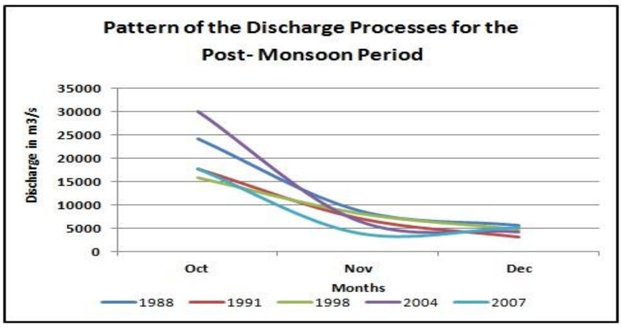


Fig 8(a): Patterns of discharge processes corresponding to all clusters obtained from AHC

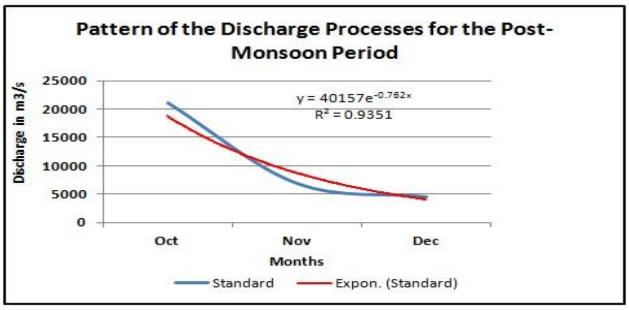


Fig 8(b): Patterns of discharge processes corresponding to all clusters obtained from AHC

Single product of cluster analysis is a tree diagram representing the entire process from individual points to one big cluster. This diagram is called a dendrogram, and is illustrated in Figure 12.

In the dendrograms the height of each U shaped line denote the distance between the objects being connected with following parameters:

(i) Proximity type: Dissimilarities, (ii) Distance: Euclidean

(iii) Agglomeration method: Ward's, and (IV) Cluster: along rows.

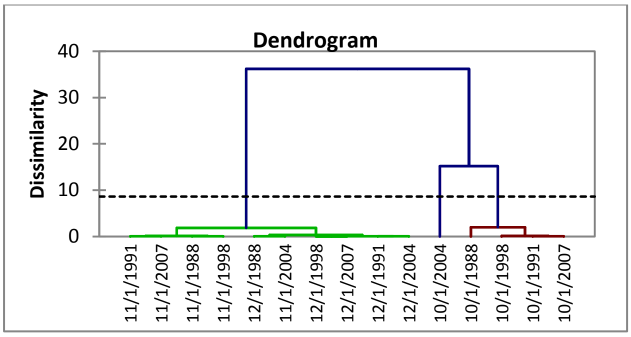


Fig 9: Dendrograms obtained after AHC of data for Monsoon, periods respectively.

5.4 Recession coefficient:

Equation (12) may be simplified for the calculation of k, as follows:-

$$\frac{Q}{Q_0} \quad = \mathbf{e}^{-\mathbf{kt}}$$

Applying natural logarithm on both sides, we get

$$\text{Log}(\ \frac{Q}{Q_0}) \ = \text{Log}\ (\mathbf{e}^{-\mathbf{kt}})$$

Since t = 1 (monthly discharge)

Therefore, the above equation is written as follows

$$\text{Log}(\ \frac{Q}{Q_0}) \ = \text{Log}\ (\text{e}^{-\text{k}}) = \ \text{-k Log(e)= -k}$$

Therefore, k = - Log($\frac{Q}{Q_0}$) = Log Q_0 – Log Q
(14)

Now, solving the above equation by substituting the values of Q and Q_0, the recession coefficient k is calculated as given in Tables 2.

In this table, the recession coefficient k is high in the month of November which shows the discharge of water is low. In the month of October the recession coefficient is less, its shows the discharge is high. In the month of December recession coefficient value decreases than the month of November, which shows the discharge is higher than previous month.

Table 2. Recession coefficient of Post Monsoon Period (Oct-Dec)

Year/Months	Oct	Nov	Dec
1988	0.17	0.42	0.19
1991	0.10	0.37	0.35
1998	0.41	0.23	0.26
2004	0.15	0.58	0.20
2007	0.29	0.32	0.29

6. ANALYSIS OF THE RESULTS

The hydrological processes show a causal-effect relationship with several happenings in nature as well as human interferences. The annual discharge processes is mainly affected by the rainfall and the temperature. The average monthly rainfall in mm to the upstream catchment for the month of October, November, and December are 90.9, 18.5 and 7. 3, respectively. The maximum monthly temperature for the months October, November and December of upstream catchment are 30, 28 and 25 in degree Celsius, respectively. Meanwhile, the discharge pattern is in decline mode due to upstream catchment rainfall contribution which affects the downstream catchment discharge. The rainfall distribution is taken and plotted in the Figures 10 shows the trend of average rainfall and temperature of upstream catchment station. The graphs shows similar pattern that conformed to the discharge pattern. This shows that rainfall is a strongly as a contributor to the discharge. The distribution of rainfall has shown decline order due to post monsoon season.

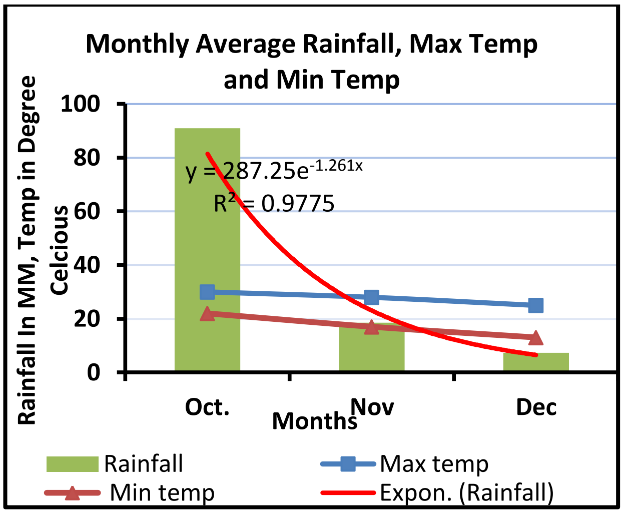


Fig 10: Monthly Average Rainfall, Max Temp and Min Temp of upstream Catchments

A relationship between the discharge and rainfall has been detected. In the analysis, month of October for the year 2004 is selected and correlation factor (Pearson correlation) between the discharge and rainfall has been obtained. The Pearson coefficient is independent of the scales and the units of variable measured. In this analysis the coefficient comes out to be 1 which is fairly good value to prove the association of discharge with the rainfall.

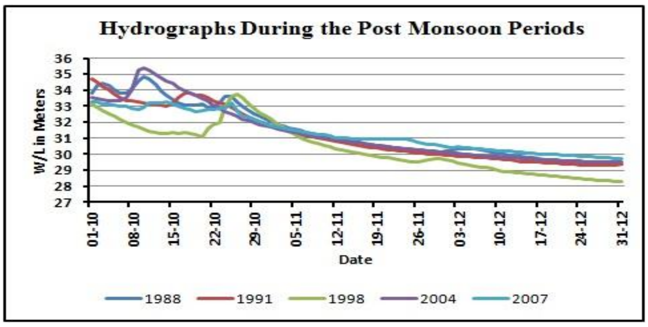


Fig 11 (a):Hydrograph exhibiting the variation in water level in Post Monsoon periods of the highly flooded years

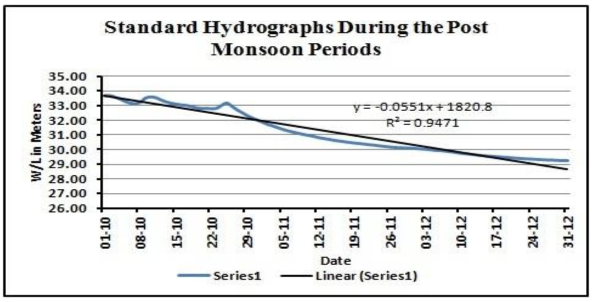


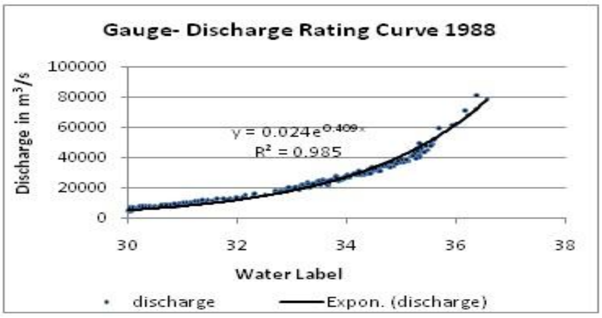
Fig 11 (b): Standard hydrograph on the basis of high flooded years

The analysis has been carried out for variation of the water level of the river during the post monsoon period as shown in Figure 11. For this, the hydrograph has been plotted for the water level data between October to December. Linear series of water level like to our discharge pattern and R^2 values varies between 0.830 to 0.952. The standard hydrograph shown in figure 11(b) and value of R^2 . These hydrographs shows the water level down fall during the high floods years in the post monsoon period.

The Gage –Discharge correlation curve has been established, shown in figures 12 and fitted the exponential regression. The R^2 value is greater than 0.95 which shows more accuracy of GD relation.

The above figures (Figure 9, 10, 11&12) show the very close relationship in the discharge pattern rainfall, hydrographs, and GD correlation curve during the same periods. Water level, rainfall variation graphs, and GD correlation curve validates our discharge pattern.

The pattern graph which is shown in Fig 9 gives expected pattern as a recession nature from the month of October to December. The horizontal line in deep blue gives the average water level in Figure 9. The discharge patterns show that the runoff of river is started to decline after the end of monsoon. The dotted line gives the trend which is an exponential ascending one.



y

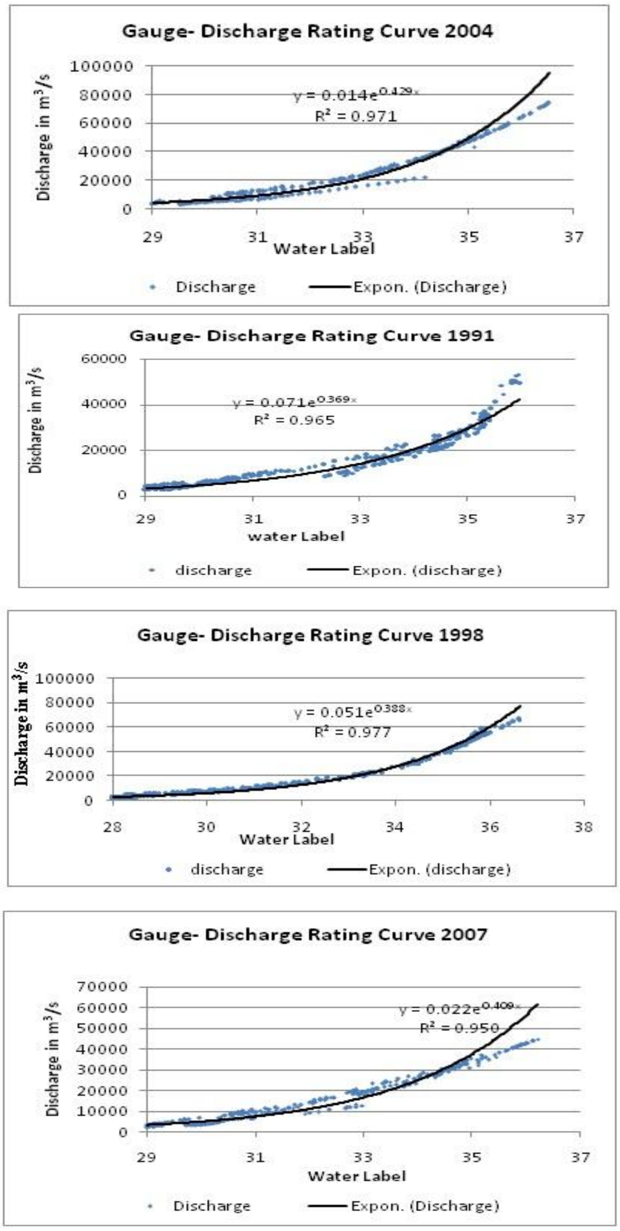


Fig 12: Gauge – Discharge Curve exhibiting the correlation during the Post Monsoon periods .

7. FORECASTING MODELS

Based on the obtained 5 patterns, the predictive models can be developed using Auto Regressive modelling technique. AR(3) model of degree 3 is developed, based on lag predictor variables Q_{n-1},Q_{n-2}, Q_{n-3}.The y-intercept or the constant was nullified and not taken into account. Table-3 enlists the models. The model is used for future prediction of discharged during the monsoon. The patterns extracted from hydrological data are valid for new hydrological data with some degree of certainty.

Table : 3. Models

Patern	Model
P1	$\mathbf{Q_n=Q_{n-1}-0.988Q_{n-2}+0.9924Q_{n-3}}$
P2	$\mathbf{Q_n=1.0174Q_{n-1}+1.005Q_{n-2}-1.0609Q_{n-3}}$
P3	$\mathbf{Q_n=0.998Q_{n-1}+0.99Q_{n-2}-0.99Q_{n-3}}$
P4	$\mathbf{Q_n=0.99Q_{n-1}+0.99Q_{n-2}-0.998Q_{n-3}}$
P5	$\mathbf{Q_n=0.99Q_{n-1}+0.99Q_{n-2}-0.99Q_{n-3}}$

8. CONCLUSIONS

In this paper, the data mining techniques like agglomerative hierarchical clustering algorithms and Ward's criterion, similarity search and pattern discovery is used in hydrological discharge time series data. The discovered patterns are more similar to discharge standard patterns. The comparison of hydrographs and rainfall during the same time period, proves that the discharge patterns one more similar under the same climatic periods. The patterns found by the AR Model to be used for the prediction of future value of discharge.

In Indian continent the whole river system is divided in three periods viz monsoon, post monsoon and pre-monsoon period. In this study, we have used only post monsoon data during the high floods year. The previous study was carried out during the monsoon and pre monsoon periods [23,24]. Our future study will focus on 20 years data a which will show complete study of hydrological behaviour of river during years for the particular station and whole catchment using ANN Model.

9. ACKNOWLEDGEMENT

The authors would like to thank the Central Water Commission, Ministry of Water Resources, India for providing Water Level and Discharge data.

10. REFERENCES

[1] Agrwal, R, Faloutsos, C. and Swami, A. (1993) 'Efficient Similarity Search in Sequence Data Bases' *International Conference on Foundations of Data Organization* (FODO),1993.

[2] Anuradha, K. and Sairam, N. (2011) 'Classification of images using JACCARD co-efficient and higher –order co-occurrences', *JATTI*, Vol. 34, No.1, pp.100-105.

[3] Aiyun, L. and Jiahai, L. (2011) 'Forecasting monthly runoff using weblet neural network model', *International conference on Mechatronic Science, Electronic Engineering and Compute*r, August 19-22,2011, *IEEE.* 978-1-61284-722-1, pp. 2177-2180.

[4] Aydin, I., Karakose and Akin, A. (2009) 'The prediction Algorithm based on Fuzzy logic using time series data mining method', *World Academy of Science, Engg. and Technology,* 27, pp. 91-98. www.waset.org/journals/waset/v27/v27-17.pdf

[5] Bradely, P. S. and Fayyad, U. M. (1998) 'Refining initial points for k-means clustering', *15th International Conference on Machine Learning, July 24-27,*1998, Madison, WI,USA, pp. 91-99.

[6] Chan, K. and Fu, A.W. (1999) 'Efficient time matching by weblets', proceeding of 15th *IEEE International Conference on Data Engineering,* 1999, Mar 23-26,Sydney,Australia, pp.126-133.

[7] Chu,S., Keogh, E., Hart, D. and Pazzani, M. (2002) 'Iterative Deepening Dynamic Time Warping for Time Series', *www.siam.org/proceedings/datamining/2002/dm02-12ChuS.pdf*

[8] Faloutsos, C., Ranganathan, M. and Manolopoulos, Y. (1994) 'Fast subsequence matching in time-series databases', *In proceedings of the ACM SIGMOD, International Conference on Management of Data,1994 May 25-27,* Minneapolis, MN, pp 419-429.

[9] Feng, L.H., and Zhang, J.Z. (2010) 'Application of ANN in Forecast of surface runoff', *Networked Computing (INC), 2010, 6th International Conference, 11-13 May 2010,* Gyeongju, Korea (South), *IEEE* (INC). pp. 1-3.

[10] Giorgino, T. (2009) 'Computing and Visualizing Dynamic Time Warping Alignments in R: The DTW Package', *Journal of Statistical Software*, Vol. 31, Issue 7 pp.1-24

[11] Han, J.W. and Kamber, M. (2001) 'Data Mining Concept and techniques', *Morgan Kaufman*: San Fransciko, CA.

[12] Jayanthi ,R. (2007) 'Application of data mining techniques in pharmaceutical industry', *JATIT*, pp61-67.

[13] Jingwen, XU. (2009) 'Mid-short term daily runoff forecasting by ANNs and multiple process based hydrological models', *IEEE*, Conference : YC-ICT, pp.526-529.

[14] Kadir, A. and Peker, (2005) Subsequence Time Series (STS) Clustering Techniques for Meaningful Pattern Discovery', KIMAS 2005, April 18-21, Waltham, MA,USA, *IEEE*,0-7803-9013.

[15] Kamali M, Nezhad, Chokmani K, TBMJ Ourda, Barbet, M. and Bruneau, P. (2010) "Regional Flood frequency Analysis using residual kriging in physiographical space. *Hydrological Processes, in Wiley*, InterScience, Vol. 24, pp.2045-2055.

[16] Kanungo,T., Mount, D.M., Netanyahu, N. S., Piatko, C.D., Silverman, R. and Wu, A.Y. (2002) 'An efficient k-Means clustering algorithm: analysis and implementation', *IEEE Transactions On Pattern Analysis And Machine Intelligence,* Vol. 24, NO. 7, pp. 881-892.

[17] Keogh , E., Lin, J. and Truppel, W. (2003), 'Clustering of Time Series Subsequences is Meaningless : Implictions for past and Future research', 3rd *IEEE International conference on Data mining, Melbourne,FL, USA.*

[18] Li, C. and Yuan, X. (2008) 'Research and Application of Data Mining for Runoff Forecasting , *IEEE,* 978-0-7695-3357-5/08.

[19] Liang, X., and Liang, Y. (2001), 'Applications of Data Mining in Hydrology', *ICDM0, Proceedings*

[20] *2001 IEEE International Conference on Data Mining, 29 November-2 December 2001,San Jose, California* . pp. 617-620.

[21] Liao, T. W. (2005) 'Clustering of time series data—a survey', *Pattern Recognition,* Vol. 38, pp. 1857–1874.

[22] Mishra S., Majumder S., and Dwivedi V.K. **,** pattern discovery in hydrological time series data mining in a Sustainable Water resources Management And Climate Change Adaptation, Vol.-II, pp.107-115, February 17-19, 2011, NIT Durgapur.

[23] Mishra, S., Dwivedi, V.K., Saravanan, C. and Pathak, K.K. (2013) 'Pattern discovery in hydrological time series data mining during the monsoon period of the high flood years in Brhamaputra river basin' *IJCA*, Vol.67,No.6,pp.7-14.

[24] Mishra S., Sarvanan C., Dwivedi V.K. , and Pathak K. K., Discovering Flood Rising Pattern in Hydrological Time Series Data Mining during the Pre Monsoon Period, Indian. J.of Zeo-Marine Science, Accepted on 12/01/2014.

[25] Martinec, J. (1970) 'Recession Coefficient in Glacier Runoff Studies', *bulletin of the International Association of Hydrology*, XV,1, 3/1970.

[26] Mirkin, B. (2011) 'Core Concepts in Data Analysis : Summarization, correlation and Visualization', *Springer –* Verlag London Limited.

[27] Ni, X. (2008) 'Research of Data mining based on neural Networks', *World Academy of Science, Engineering and Technology 15 2008*, pp.381-384.

[28] Ouyang, R., Ren, L., Cheng, W. and Zhou, C. (2010) 'Similarity search and pattern discovery in Hydrological time series data mining', *Wiley InterScience* , Hydrol. Process, Vol. 24, pp.1198-1210.

[29] Piatetsky-Shapiro, G. and Frawley, W. J. (1991). 'Knowledge Discovery in Databases', *AAAI/MIT* Press: Boston, MA,

[30] Pujari, A.K. (2006) ' Data Mining Techniques', *University press*.

[31] Rafiq, M.I., Martin, J., Connor, O. and Das, A.K. (2005) 'Computational method for Temporal Pattern Discovery in Biomedical Genomic Database', *Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference, IEEE* (CSB'05), pp. 362-365.

[32] Ruhana, K., Mohamud, K., Zakaria, N., Katuk, N., and Shbier M. 2009, Flood Pattern Detection Using Sliding Window Techniques, 978-0-7695-3648-4/09, *IEEE* DOI 10.1109/AMS,2009, 15.

[33] Samsudin, R., Saad, P., and Shabri , A. 2011, River flow time series using least square support vector machines, Hydrol,Earth Syst. Sci, 15, 1835-1852, 2011.

[34] Sihui, D. 2009, Forecast Model of Hydrologic Single, Element Medium and Long-Period Based on Rough Set Theory, *IEEE*, 978—07695-3735-1/09.

[35] Spate, J.M., Croke, B.F.W., and Jakeman, A.J. (2003) 'Data Mining in Hydrology', Conference: *MODSIM* , 2003.

[36] Shijin, L., Lingling, J., Yuelong, Z., and Ping, B. (2012),hybrid forecasting model of discharge based on support vector machine,Elseveir

[37] Tapas, K., David, M., Nathan, S., Christine, D., and Angela, Y. (2002) 'An efficient k-Means Clustering Algorithm: Analysis and Implementation', IEEE Vol. 24, No.7.pp. 881-892.

[38] Theodoridis, S. and Koutroumbas , K. (2006) 'Pattern Recognition', Third Edition, New York: *Academic Press*, Vol.39 No.5, pp. 776-788.

[39] Weilin, L. (2011) 'Neural network model for hydrological forecasting based on multivariate phase space reconstruction', *IEEE*, Vol.2, pp. 663-667.

[40] Wiriyarattanakul, S., Auephanwiriyakull, S., Theera, and Umpon, N, (2008) 'Runoff Forecasting using Fuzzy Support Vector Regression', ISPACS 2008, *IEEE*, pp.1-4.

[41] Wu, C.L., and Chaw, K.W. (2010), 'Data-driven models for monthly stream flow time series Prediction', *Engineering Applications of Artificial Intelligence*, Vol. 23, No.8, pp.1350-1367.

[42] Yuelong, Z., Shijin, L., Dingsheng, W. and Xiaohua, Z. (2008) 'A Novel Approach to the Similarity Analysis of Multivibrate Time series and its Application in Hydrological Data mining', International Conference on Computer Science and Software Engineering, 2008 , *IEEE*, Vol. 4, pp. 730-734.