Data mining for evolution of association rules for droughts and floods in India using climate inputs

C. T. Dhanya[1] and D. Nagesh Kumar[1]

Received 23 May 2008; revised 20 October 2008; accepted 10 November 2008; published 22 January 2009.

[1] An accurate prediction of extreme rainfall events can significantly aid in policy making and also in designing an effective risk management system. Frequent occurrences of droughts and floods in the past have severely affected the Indian economy, which depends primarily on agriculture. Data mining is a powerful new technology which helps in extracting hidden predictive information (future trends and behaviors) from large databases and thus allowing decision makers to make proactive knowledge-driven decisions. In this study, a data-mining algorithm making use of the concepts of minimal occurrences with constraints and time lags is used to discover association rules between extreme rainfall events and climatic indices. The algorithm considers only the extreme events as the target episodes (consequents) by separating these from the normal episodes, which are quite frequent, and finds the time-lagged relationships with the climatic indices, which are treated as the antecedents. Association rules are generated for all the five homogenous regions of India and also for All India by making use of the data from 1960 to 1982. The analysis of the rules shows that strong relationships exist between the climatic indices chosen, i.e., Darwin sea level pressure, North Atlantic Oscillation, Nino 3.4 and sea surface temperature values, and the extreme rainfall events. Validation of the rules using data for the period 1983–2005 clearly shows that most of the rules are repeating, and for some rules, even if they are not exactly the same, the combinations of the indices mentioned in these rules are the same during validation period, with slight variations in the classes taken by the indices.

Citation: Dhanya, C. T., and D. Nagesh Kumar (2009), Data mining for evolution of association rules for droughts and floods in India using climate inputs, *J. Geophys. Res.*, *114*, D02102, doi:10.1029/2008JD010485.

1. Introduction

[2] Asian monsoon greatly influences most of the tropics and subtropics of the eastern hemisphere and more than 60% of the earth's population [*Webster et al.*, 1998]. While the failure of the monsoon brings famine, an excess or strong monsoon will result in devastating floods, particularly if they are unanticipated. An accurate prediction of these two extremes (drought and flood) can help decision makers to improve planning to mitigate the adverse impacts of monsoon variability and to take advantage of beneficial conditions [*Webster et al.*, 1998]. From the early 1900s, various climatic and oceanic parameters had been used as predictors for monsoon rainfall prediction. Thus, if the association of the extremes with the climatic and oceanic parameters can be revealed, this can be used for designing an effective risk management system for facing the extremes.

[3] India receives major portion of its annual rainfall during the south west monsoon season (June–September). Even a small variation in this seasonal rainfall can have an adverse impact on Indian economy. As per the Indian Meteorological Department (IMD), an annual rainfall event is considered a drought (flood) if it is less (greater) than one standard deviation from the long-term average annual rainfall. According to this definition, in the past 50 years, India has experienced around 10 droughts and 9 floods with highest intensity of drought and flood in 1972 and 1959 respectively. Two multiyear droughts also occurred in the 1960s and 1980s. The frequency and intensity of drought is much more than of the flood.

[4] Recent studies in the variation of the Gross Domestic Product (GDP) and the monsoon [*Gadgil and Gadgil*, 2006] have showed that the impact of severe droughts is about 2 to 5% of the GDP throughout. This indicates the need for taking proactive steps to address the impacts of both the rainfall extremes which in turn demand for an accurate prediction of the occurrence and nonoccurrence of the extremes. It is also shown that the impact of deficit rainfall (drought) on GDP is larger than that of surplus rainfall (flood).

[5] Studies on the prediction of Indian Summer Monsoon Rainfall (ISMR) have used various empirical and physical (atmospheric and coupled) models. A brief history of these studies and the models and predictors used is shown in Table 1. A comparative study between empirical and physical models [*Goddard et al.*, 2001] has shown that

[1]Department of Civil Engineering, Indian Institute of Science, Bangalore, India.

0148-0227/09/2008JD010485

Table 1. Models Used for Prediction of Indian Summer Monsoon Rainfall

Serial Number	Predictors Used	Technique	Reference
1	Darwin sea level pressure, latitudinal position of 500-mb ridge along 75°E	Linear regression model	*Shukla and Mooley* [1987]
2	Arabian sea SST	Nonlinear gravity model	*Dube et al.* [1990]
3	Darwin sea level pressure, latitudinal position of 500-mb ridge along 75°E, May surface resultant wind speed	Neural network	*Navone and Ceccatto* [1994]
4	Northern Australia-Indonesia SST, Darwin pressure	Correlation analysis	*Nicholls* [1995]
5	Indian Ocean SST	Linear regression model	*Clarke et al.* [2000]
6	Quasi biennial oscillation, sea surface temperature anomalies over different Nino regions	Correlation analysis	*Chattopadhyay and Bhatla* [2002]
7	Darwin sea level pressure tendency, Nino 3.4, NAO, quasi biennial oscillation, western Pacific region SST, eastern Indian Ocean region SST, Arabian Sea region SST, Eurasian surface temperature, and Indian surface temperature	Linear regression model	*DelSole and Shukla* [2002]; *DelSole and Shukla* [2006]
8	Equatorial east Indian Ocean sea surface temperature	Correlation analysis	*Reddy and Salvekar* [2003]
9	Indian summer monsoon rainfall	Neural network + Linear regression	*Iyengar and Raghu Kanth* [2004]
10	Arabian Sea SST, Eurasian snow cover, northwest Europe temperature, Nino 3 SST anomaly (previous year), south Indian Ocean SST index, East Asia pressure, Northern Hemisphere 50-hPa wind pattern, Europe pressure gradient, south Indian Ocean 850-hPa zonal wind, Nino 3.4 SST tendency, North Indian Ocean-North Pacific Ocean 850-hPa zonal wind difference, North Atlantic Ocean SST	Power regression model	*Rajeevan et al.* [2004]
11	Nino 3.4 and Equatorial zonal Wind INdex (EQWIN)	Bayesian dynamic linear models	*Maity and Nagesh Kumar* [2006]
12	First stage predictors: North Atlantic SST anomaly, equatorial SE Indian Ocean anomaly, East Asia surface pressure anomaly, Europe land surface air temperature anomaly, northwest Europe surface pressure anomaly tendency, Equatorial Pacific Warm Water Volume (WWV) anomaly Second stage predictors: first three first-stage predictors, Nino 3.4 SST anomaly tendency, North Atlantic surface pressure anomaly, North Central Pacific zonal wind anomaly at 850 hPa	Ensemble multiple linear regression model and projection pursuit regression model	*Rajeevan et al.* [2006]
13	Arabian Sea SST and central equatorial Indian Ocean SST	Simple regression model	*Sadhuram* [2006]
14	Nino 3.4 and EQWIN	Correlation and phase plane analysis	*Gadgil et al.* [2007]
15	Nino 3.4 and EQWIN	Semiparametric, copula-based approach	*Maity and Nagesh Kumar* [2008]

empirical models continue to outperform physical models in prediction of ISMR, as most of the physical models are unable to simulate accurately the interannual variability of ISMR. However the skill of any of these models in predicting the extremes is not satisfactory [*Gadgil et al.*, 2005]. None of these models could successfully predict the droughts of 2002 and 2004. One of the reasons for the inability of these models to capture the relationship of the extremes with the predictors may be due to the infrequent occurrence of the extremes. Assuming the rainfall distribution as a normal fre-

quency curve, the occurrence of either drought or flood covers only 16% of the time (since only 16% of the distribution area is less than the mean − 1 × standard deviation).

[6] In this study, a time series data-mining algorithm is used to generate the association rules between oceanic and atmospheric parameters and rainfall extremes. In this attempt, attention is given to find the relationship between only the extremes and the predictors, without considering the normal rainfall which is quite frequent. By using such a data-mining algorithm in this context, one of the advantages is that there is no need to have a prior idea about the correlation and causal relationships between the variables. Unlike the empirical methods, this method takes into account the interrelationships between the predictor variables very well. The exact values of the model parameters such as coefficients in a regression model or weights in a neural network are of little importance in this approach. Thus the objective here is to unearth all the frequent patterns (episodes) of the predictors that precede the extreme episodes of rainfall using a time series data-mining algorithm.

2. Time Series Data Mining

[7] Data mining can be defined as a process in which specific algorithms are used for extracting some new nontrivial information from large databases. Data-mining techniques are widely applied in business activities and also in scientific and engineering scenarios. Various data-mining techniques can be broadly classified into two types [*Han and Kamber*, 2006]: descriptive data mining, in which the data in the database are characterized according to their general properties and predictive data mining, in which predictions are made by performing inference from the current data. Frequent patterns and association rules, clustering and deviation detection come under the first category while regression and classification come under the second one. Almost all the studies done so far on rainfall extremes are based on the predictive data-mining techniques. As mentioned earlier, these studies were unable to successfully predict the infrequent extreme episodes. Hence, in this study, a descriptive data-mining technique is used to capture especially the infrequent extreme episodes.

[8] Temporal data mining is concerned with data mining of large sequential sets (ordered data with respect to some index). Time series is a popular class of sequential data in which records are indexed by time. The possible objectives in the case of temporal data mining can be grouped as follows: (1) prediction, (2) classification, (3) clustering, (4) search and retrieval, and (5) pattern discovery [*Han and Kamber*, 2006]. Among these, algorithms of pattern interest are of most recent origin. The word "*pattern*" means a local structure in the data. The objective is to simply unearth all patterns of interest. One common measure to assess the value of a pattern is the frequency of the pattern. A frequent pattern is one that occurs many times in the data. The frequent patterns thus discovered can be used to discover the causal rules.

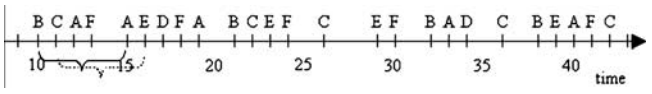
[9] A rule consists of a left-hand side proposition (antecedent) and a right hand side proposition (consequent). The rule states that when the antecedent occurs (is true), then the consequent also occurs (is true). Rule based approaches are often used to ascertain the relationships within the data set. For example, association rules determine the rules that indicate whether or how much the values of an attribute depend on the values of the other attributes in the data set. These are used to capture correlations between different attributes in the data. In such cases, the conditional probability of the occurrence of the consequent given the antecedent is referred to as the confidence of the rule. For example, if a pattern "B follows A" occurs n_1 times and the pattern "C follows B follows A" occurs n_2 times, then the temporal association rule "whenever B follows A, C will also follow" has a confidence of (n_2/n_1). The value of a rule is usually measured in terms of its confidence.

[10] There are two popular frameworks for frequent pattern discovery namely sequential patterns and episodes. In the sequential patterns framework, a collection of sequences are given and the task is to discover the order of sequences of the items (i.e., sequential patterns) that occurs in sufficiently good number of those sequences. In the frequent episodes framework, the data are given in a single long sequence and the task is to unearth temporal patterns (called episodes) that occur sufficiently often along that sequence. Frequent episodes framework is used in the present study, since one does not know in prior all the sequences to be searched in the time series as is required in sequential patterns framework. Also, concern is to extract the temporal patterns of the climatic indices and extreme events which can be done by applying frequent episodes framework. Several algorithms were formulated [*Mannila et al.*, 1997] for the discovery of frequent episodes within one sequence.

2.1. Framework of Frequent Episode Discovery

2.1.1. Event Sequence

[11] The data, referred to here as an *event sequence*, are denoted by $\langle (E_1, t_1), (E_2, t_2), \ldots \rangle$ where E_i takes values from a finite set of event types ε, and t_i is an integer denoting the time stamp of the ith event. The sequence is ordered with respect to the time stamps so that, $t_i \leq t_{i+1}$ for all $i = 1, 2, \ldots$. The following is a sample event sequence with six event types A, B, C, D, E and F in it:



[12] Any event sequence can be expressed as a triple element (s, T_B, T_D) where s is the time-ordered sequence of events from beginning to end, T_B is the beginning time and T_D is the ending time.The above sample event sequence can be expressed as S = (s, 9, 43) where $s = \langle (B, 10), (C, 11), (A, 12), (F, 13), (A, 15), \ldots (C, 42) \rangle$.

2.1.2. Episode

[13] An episode α is defined by a triple element $(V_\alpha, \leq_\alpha, g_\alpha)$, where V_α is a collection of nodes, $\leq_\alpha$ is a partial order on V_α and g_α: $V_\alpha \rightarrow \varepsilon$ is a map that associates each node in the episode with an event type. Thus an episode is a combination of events with a time-specified order. When there is a fixed order among the event types of an episode, it is called a *serial* episode and when there is no order at all, the episode is called a *parallel* episode.

[14] An episode is said to *occur* in an event sequence if there exist events in the sequence occurring in exactly the same order as that prescribed in the episode, within a given time bound. For example, in the above sample event sequence, the events (*A*, *19*), (*B*, *21*) and (*C*, *22*) constitute an occurrence of a *3-node serial episode* ($A \rightarrow B \rightarrow C$) while the events (*A*, *12*), (*B*, *10*) and (*C*, *11*) do not, because for this serial episode to occur, *A* must occur before *B* and *C*.

2.1.3. Window

[15] Now, to find all frequent episodes from a class of episodes, the user has to define how close is close enough by defining a time window width within which the episodes should appear. For an episode to be interesting, the events in an episode must occur close to each other in time span. A window can be defined as a slice of an event sequence and then the event is considered as a sequence of partially overlapping windows. A window on an event sequence (s, T_s, T_e) can also be expressed as a triple element w = (w, t_s, t_e)., where $t_s < T_e$, $t_e > T_s$ and w consists of those event pairs from s where $t_s \leq t_i \leq t_e$. The time span $t_e - t_s$ is called the width of the window w.

[16] Consider the example of event sequence given above. Two windows of width 5 are shown. The first window starting at time 10 is shown in solid line, followed by a second window shown in dashed line. First window can be represented as ($\langle$(*B*, 10), (*C*, 11), (*A*, 12), (*F*, 13)$\rangle$, 10, 15). Here the event (*A*, 15) occurred at the ending time is not included in the sequence. Similarly, the second window can be represented as ($\langle$(*C*, 11), (*A*, 12), (*F*, 13), (*A*, 15)$\rangle$, 11, 16).

[17] For a sequence S with a given window width "*win*", the total number of windows possible is given by $W(s, win) = T_e - T_s + win$. This is because the first and last windows extend outside the sequence, such that the first window contains only the first time stamp of the sequence and the last window contains only the last time stamp. Hence an event close to either end of a sequence is observed in equally many windows to an event in the middle of the sequence. For the sequence given above, totally there are 39 partially overlapping windows with first window (Φ, 5, 10) and last window (Φ, 43, 48).

[18] The frequency of an episode is defined as the number of windows in which the episode occurs divided by the total number of windows in the data set. For the *3-node serial episode* ($A \rightarrow B \rightarrow C$), there are only two occurrences i.e., in windows ($\langle$(*F*, 18), (*A*, 19), (*B*, 21), (*C*, 22), 18, 23) and ($\langle$(*A*, 19), (*B*, 21), (*C*, 22), (*E*, 23)$\rangle$, 19, 24). Thus the frequency of the episode is $(2/39) \times 100 = 5.13\%$. Now, given an event sequence, a window width and a frequency threshold, the task is to discover all frequent episodes in the event sequence.

[19] Once the frequent episodes are known, it is possible to generate rules that describe temporal correlations between events. However, there can be other ways to define episode frequency.

2.1.4. MINEPI Algorithm

[20] One such alternative proposed by *Mannila et al.* [1997] is MINEPI algorithm and is based on counting what are known as *minimal occurrences* of episodes. A minimal occurrence of an episode is defined as a window (or contiguous slice) of the input sequence in which the episode occurs, subject to the condition that no proper subwindow of this window contains an occurrence of the episode. The algorithm for counting minimal occurrences trades space efficiency for time efficiency as compared to the windows-based counting algorithm. In addition, since the algorithm locates and directly counts occurrences (as against counting the number of windows in which episodes occur), it facilitates the discovery of patterns with extra constraints (such as being able to discover rules of the form "if *A* and *B* occur within 10 seconds of one another, *C* follows within another 20 seconds").

[21] Minimal occurrences of episodes with their time intervals are identified in the following way. For a given episode α and an event sequence S, the minimal occurrence of α in S is the interval $[t_s, t_e]$, if (1) α occurs in the window $w = (w, t_s, t_e)$ on S, and if (2) α does not occur in any proper subwindow on w. A window $w' = (w', t'_s, t'_e)$ will be a proper subwindow of w if $t_s \leq t'_s$, $t'_e \leq t_e$, and $\text{width}(w') < \text{width}(w)$. The set of minimal occurrences of an episode α in a given event sequence is denoted by $mo(\alpha) = \{[t_s, t_e)|[t_s, t_e)\}$. For the example sequence given above, the *serial episode* $\alpha = B \rightarrow C$ has four minimal occurrences i.e., $mo(\alpha) = \{[10,11), [21,22), [32,36), [38,42)\}$.

[22] The concept of frequency of episodes explained in the previous section is not very useful in the case of minimal occurrences as there is no fixed window size and also a window may contain several minimal occurrences of an episode. Therefore *Mannila et al.* [1997] used the concept of *support* instead of frequency. The support of an episode α in a given event sequence S is $|mo(\alpha)|$. An episode α is frequent if $|mo(\alpha)| \geq$ user defined minimum support threshold.

2.1.5. MOWCATL Algorithm

[23] The above approach was modified to handle separate antecedent and consequent constraints and maximum window widths and also the time lags between the antecedent and consequent to find natural delays embedded within the episodal relationships by *Harms and Deogun* [2004] in Minimal Occurrences With Constraints And Time Lags (MOWCATL) algorithm. Although MINEPI and MOWCATL both use the concept of minimal occurrences to find the episodal relationships, MOWCATL has some additional mechanisms like (1) allowing a time lag between the antecedent and consequent of a discovered rule, and (2) working with episodes from across multiple sequences [*Harms et al.*, 2002]. Episodal rules are found out where the antecedent episode occurs within a given maximum window width win_a, the consequent episode occurs within a given maximum window width win_c, and the start of the consequent follows the start of the antecedent within a given maximum time *lag*. This algorithm allows to find rules of the form: "if A and B occur within 3 months, then within 2 months they will be followed by C and D occurring together within 4 months".

[24] This algorithm first goes through the data sequence and stores the occurrences of all single events for the antecedent and consequent separately. The algorithm only looks for the target episodes specified by the user. So it prunes the episodes that do not meet the user specified minimum support threshold. Then two event episodes are generated by pairing up the single events so that the pairs of events occur within the prescribed window width and the occurrences of these two event episodes in the data se-

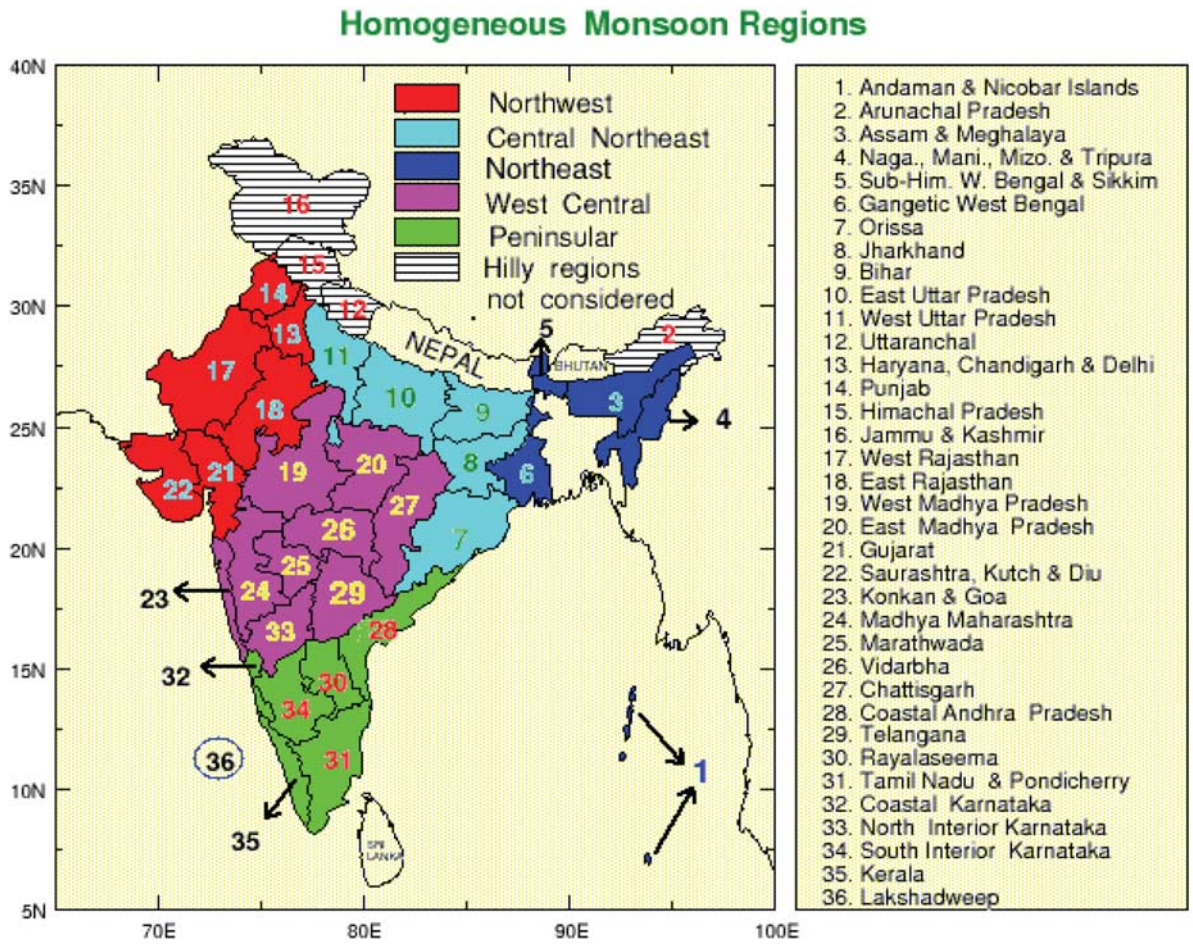


Figure 1. Homogenous monsoon regions of India, as defined by the Indian Institute of Tropical Meteorology.

quence are recorded. This is repeated until there are no more events to be paired up. The process repeats for three events, four events and so on until there are no episodes left to be combined that meet the minimum threshold. The frequent episodes for antecedent and consequent sequences are found independently. These frequent episodes are combined to form an episodal rule.

[25] An episodal rule is that in which an antecedent episode occurs within a given window width, a consequent episode occurs within a given window width and the start of the consequent follows the start of the antecedent within a user specified time lag. For example, let episode X is of the events A and B, and episode Y is of the events C and D. Also the user specified antecedent window width is 3 months, consequent window width is 2 months and the time lag is 3 months. Then the rule generated would indicate that if A and B occur within 3 months, then within 3 months they will be followed by C and D occurring together within 2 months. The support of the rule is the number of times the rule occurs in the data sequence. The confidence of the rule is the conditional probability that the consequent occurs, given the antecedent occurs. For the rule "X is followed by Y", the confidence is the ratio of the Support[X and Y] and Support[X]. Here X is a serial antecedent episode ($A \rightarrow B$) and Y is a serial consequent episode ($C \rightarrow D$).

[26] The support and confidence are the two measures used for measuring the value of the rule. The values of these are set high to prune the association rules. Even after setting the threshold of these measures high, there will be an adequate number of rules, making the user's task of rule selection difficult. The user needs some quantifying measures to select the most valuable rules in addition to the support and confidence measures. Several interestingness or goodness measures are used to compare and select better rules from the ones that are generated [*Bayardo and Agarwal*, 1999; *Das et al.*, 1998; *Harms et al.*, 2002]. In MOWCATL algorithm, J measure is used for rule ranking [*Smyth and Goodman*, 1991]. The J measure is given by

$$J(x;y) = p(x)\begin{bmatrix} p(y|x) \times \log[p(y|x)/p(y)] + \\ \\ [1-p(y|x)] \times \log\{[1-p(y|x)]/[1-p(y)]\} \end{bmatrix} \quad (1)$$

where p(x), $p(y)$ and $p(y|x)$ are the probabilities of occurrence of x, y and y given x respectively in the data sequence. The first term in the J measure is a bias toward rules which occur more frequently. The second term i.e., the term inside the square brackets is well known as cross-entropy, namely the information gained in going from the

Table 2. Threshold Values Used for the Categorization of Monthly Rainfall (mm) for Various Regions and Also for All India[a]

Region	Extreme Drought	Severe Drought	Moderate Drought	Normal Rainfall	Moderate Flood	Severe Flood	Extreme Flood
Northwest	$\leq$150	$150 < X \leq 500$	$500 < X \leq 850$	$850 < X < 1550$	$1550 \leq X < 1900$	$1900 \leq X < 2250$	$\geq$2250
West central	$\leq$1100	$1100 < X \leq 1450$	$1450 < X \leq 1900$	$1900 < X < 2600$	$2600 \leq X < 3000$	$3000 \leq X < 3400$	$\geq$3400
Central northeast	$\leq$1200	$1200 < X \leq 1600$	$1600 < X \leq 2000$	$2000 < X < 2850$	$2850 \leq X < 3300$	$3300 \leq X < 3700$	$\geq$3700
Northeast	$\leq$2200	$2200 < X \leq 2700$	$2700 < X \leq 3100$	$3100 < X < 3900$	$3900 \leq X < 4300$	$4300 \leq X < 4700$	$\geq$4700
Peninsular	$\leq$1000	$1000 < X \leq 1200$	$1200 < X \leq 1450$	$1450 < X < 1900$	$1900 \leq X < 2100$	$2100 \leq X < 2350$	$\geq$2350
All India	$\leq$1200	$1200 < X \leq 1500$	$1500 < X \leq 1800$	$1800 < X < 2400$	$2400 \leq X < 2700$	$2700 \leq X < 2900$	$\geq$2900

[a]Rainfall in millimeters.

prior probability $p(y)$ to a posterior probability $p(y|x)$ [*Das et al.*, 1998]. Compared to other measures which directly depend on the probabilities [*Piatetsky-Shapiro*, 1991], thereby assigning less weight to the rarer events, J measure is better suited to rarer events since it uses a log scale (information based). As shown by *Smyth and Goodman* [1991], J measure has the unique properties as a rule information measure and is a special case of Shannon's mutual information.

[27] The J values range from 0 to 1. The higher the J value the better it is. However, since drought and flood are so infrequent, the J values are so small that all values greater than 0.025 are to be considered.

[28] MOWCATL algorithm is used in the present study for extracting rules between extreme episodes and climatic indices, since this algorithm can be used for multiple sequences and also this will capture by itself the lag between the occurrences of climatic indices and rainfall events.

3. Data Used for the Study

[29] The time series data sets used in this study are of the monthly values for the period 1960 to 2005 and are defined as follows.

[30] 1. Summer monsoonal rainfall (June to September) for All India and also for the five homogeneous regions (as defined by Indian Institute of Tropical Meteorology), for the period 1960 to 2005 (http://www.tropmet.res.in).

[31] 2. Darwin sea level pressure (DSLP), (NCEP, ftp.ncep.noaa.gov/pub/cpc/wd52dg/data/indices).

[32] 3. Nino 3.4, east central tropical Pacific sea surface temperature (SST), 170°E–120°W, 5°S–5°N (http://www.cpc.ncep.noaa.gov/data/indices/sstoi.indices).

[33] 4. North Atlantic Oscillation (NAO), normalized sea level pressure difference between Gibraltor and southwest Iceland (http://www.cru.uea.ac.uk/cru/data/nao.htm).

[34] 5. 1 × 1 degree grid SST data over the region 40°E–120°E, 25°S–25°N (ICOADS, http://www.cdc.noaa.gov/icoads-las/servlets/datset).

4. Association Rules for Extremes

[35] The data-mining algorithm is applied to find the association rules of the extreme rainfall episodes with the climatic indices and thus to find the spatial and temporal patterns of extreme episodes throughout the country. The geographical locations of the homogenous regions: northwest, central northeast, northeast, west central and peninsular are shown in Figure 1.

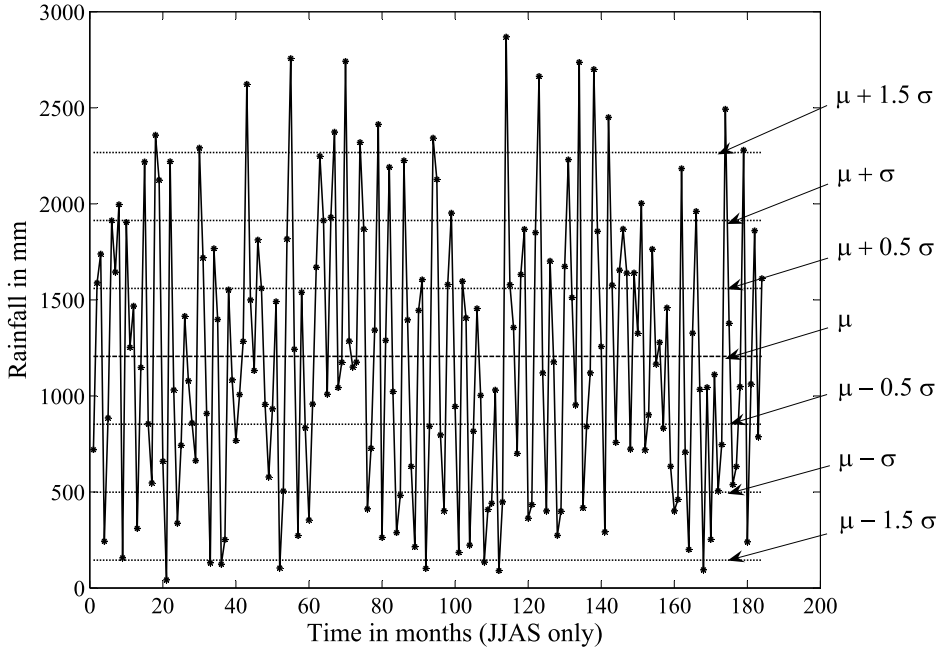


Figure 2. Summer monsoon rainfall for the northwest region for the period 1960–2005 indicating the threshold values to classify droughts and floods.

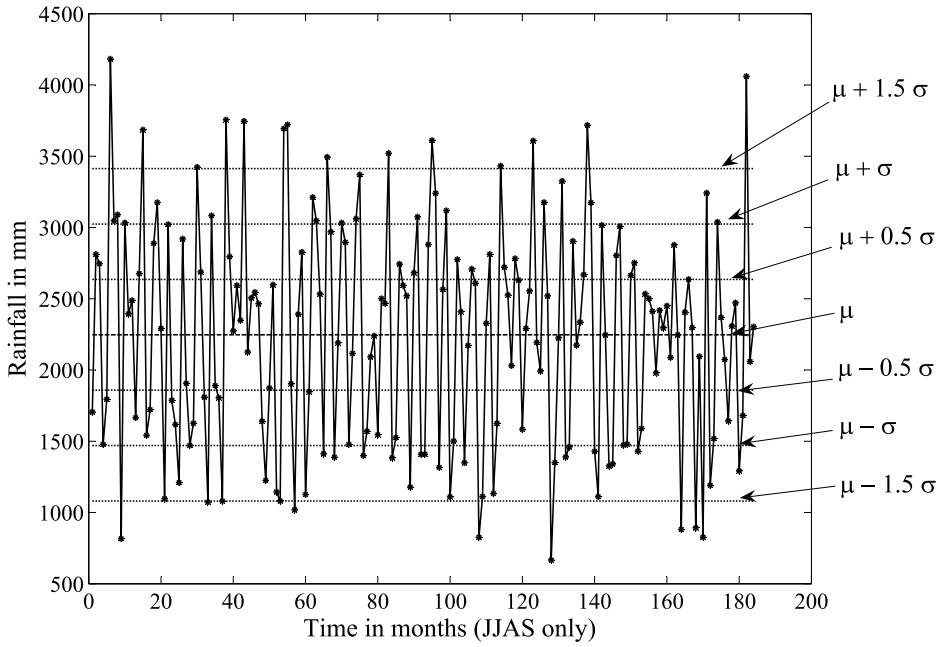


Figure 3. Summer monsoon rainfall for the west central region for the period 1960–2005 indicating the threshold values to classify droughts and floods.

4.1. Selection of Consequent Episodes

[36] In order to identify the extreme episodes, the rainfall for All India and also for the five homogenous regions is divided into seven categories. The threshold values are determined by identifying the values at ±1.5, ±1 and ±0.5 standard deviations from the average. Threshold values calculated for each region are given in Table 2. The seven classes thus identified are named as: moderate drought, severe drought, extreme drought, normal rainfall, moderate flood, severe flood and extreme flood. Although from a hydrologic point of view, greater than normal rainfall cannot be called as a flood, for a better classification, in this context, greater than normal rainfall are divided into 3 categories and are called as moderate, severe and extreme flood. Same is applicable for the classification of less than normal rainfall also. For example, while considering the northeast region, a rainfall value of less than or equal to 2200 mm/month is under the category of extreme drought although it will not result to any "real" drought.

[37] For application of the algorithm, only the extreme episodes (moderate drought, severe drought, extreme drought, moderate flood, severe flood and extreme flood)

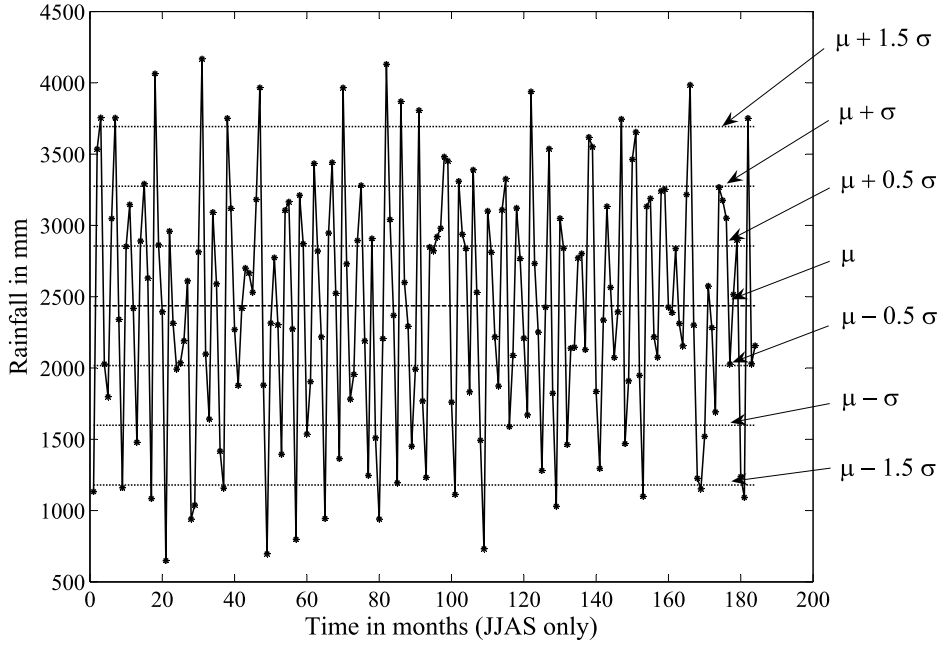


Figure 4. Summer monsoon rainfall for the central northeast region for the period 1960–2005 indicating the threshold values to classify droughts and floods.

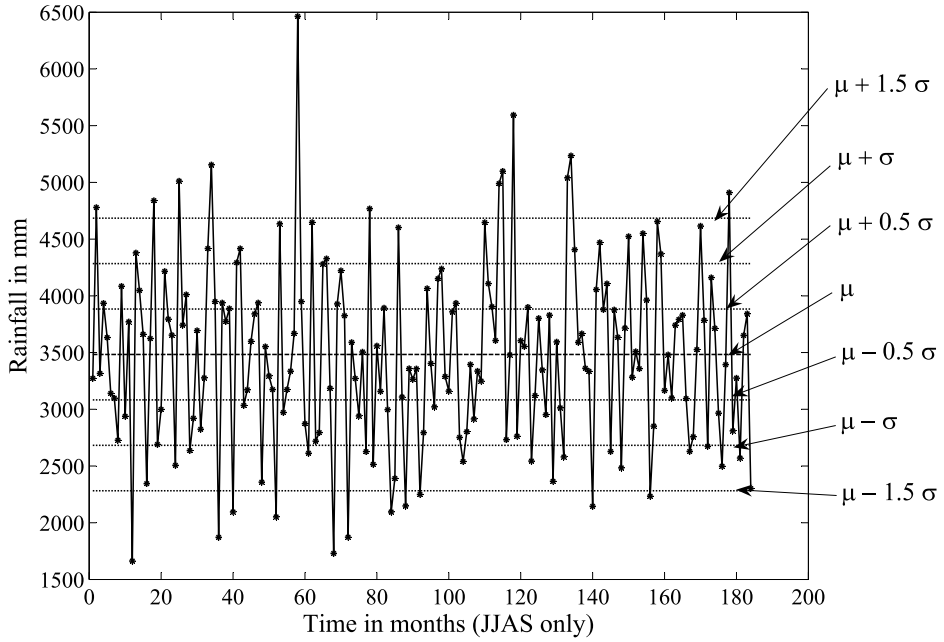


Figure 5. Summer monsoon rainfall for the northeast region for the period 1960–2005 indicating the threshold values to classify droughts and floods.

are specified as the target episodes. The summer monsoon rainfall (JJAS) time series of each region and also of All India for the period 1960–2005, indicating the threshold values are shown in Figures 2–7.

4.2. Selection of Antecedent Episodes

[38] 1 × 1 degree grid SST data over the region 40°E–120°E, 25°S–25°N are averaged to a 5 × 5 degree grid data, thus reducing to 127 grids (excluding the land area regions). Among these, the most influencing grids are selected by plotting the correlation contour plots considering different lags for each region. Grids used for correlation analysis (numbered 1 to 127) are shown in Figure 8. The maximum correlation of SST with the summer monsoon is achieved at lag 7 for all the regions. The correlation contours for northwest region for lag 7 is shown in Figure 9. The

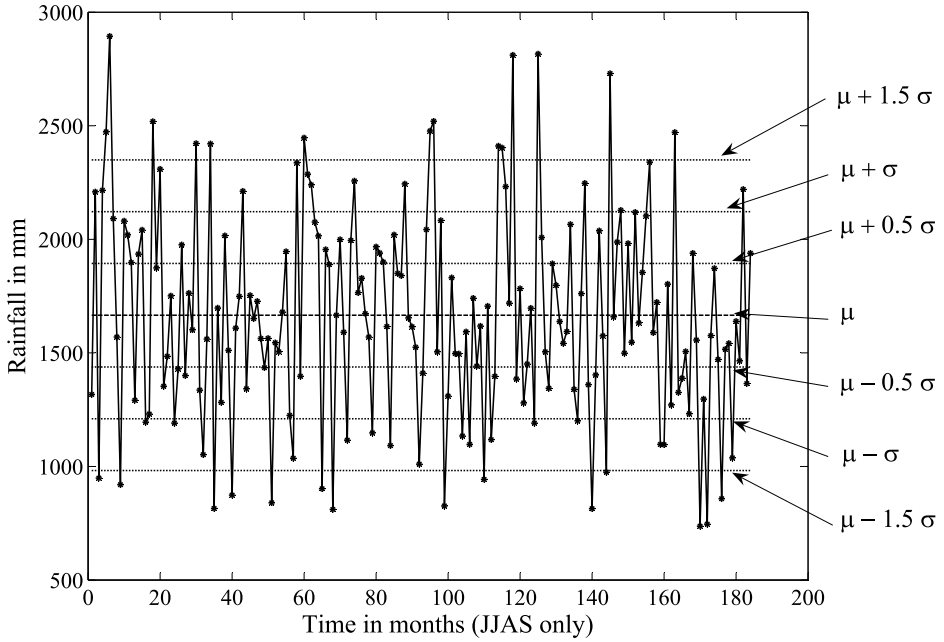


Figure 6. Summer monsoon rainfall for the peninsular region for the period 1960–2005 indicating the threshold values to classify droughts and floods.

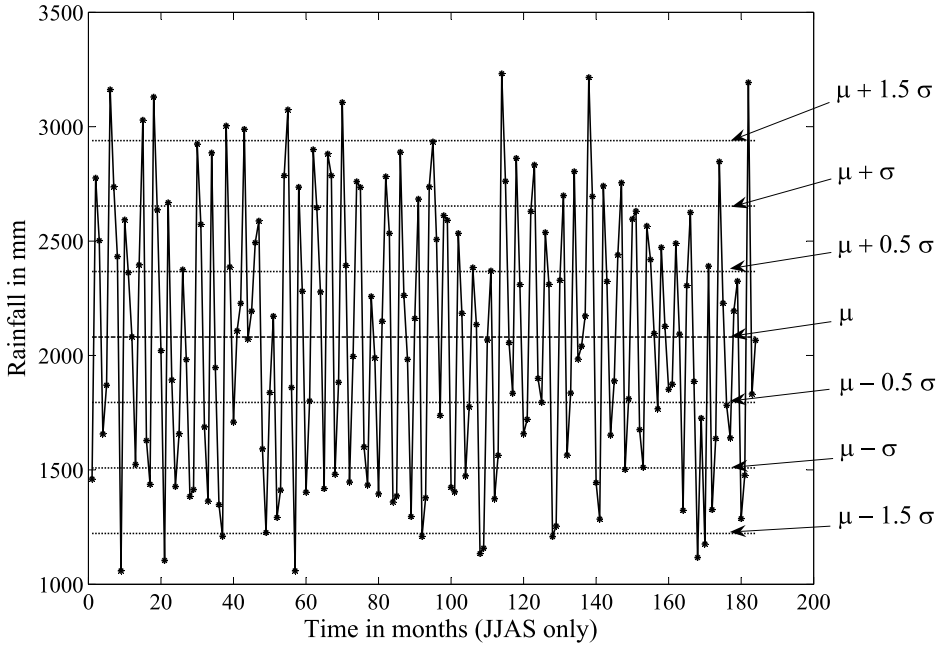


Figure 7. Summer monsoon rainfall for All India for the period 1960–2005 indicating the threshold values to classify droughts and floods.

variation of correlation versus lag for northwest region is shown in Figure 10 as an illustration.

[39] The climatic indices which are used as antecedents in rule generation are thus, DSLP, Nino 3.4, NAO and SST values of those grids which are showing maximum correlation with the summer monsoon rainfall of each region. The most influencing grids and the corresponding maximum correlation for each region are given in Table 3.

[40] The climatic indices are also categorized into seven categories by segregating at ±1.5, ±1 and ±0.5 standard deviations from the average. Threshold values for these indices (except the SST grids) are given in Table 4. The time series of the climatic indices (DSLP, NAO and Nino 3.4) for the period 1960–2005, indicating the threshold values are shown in Figures 11–13.

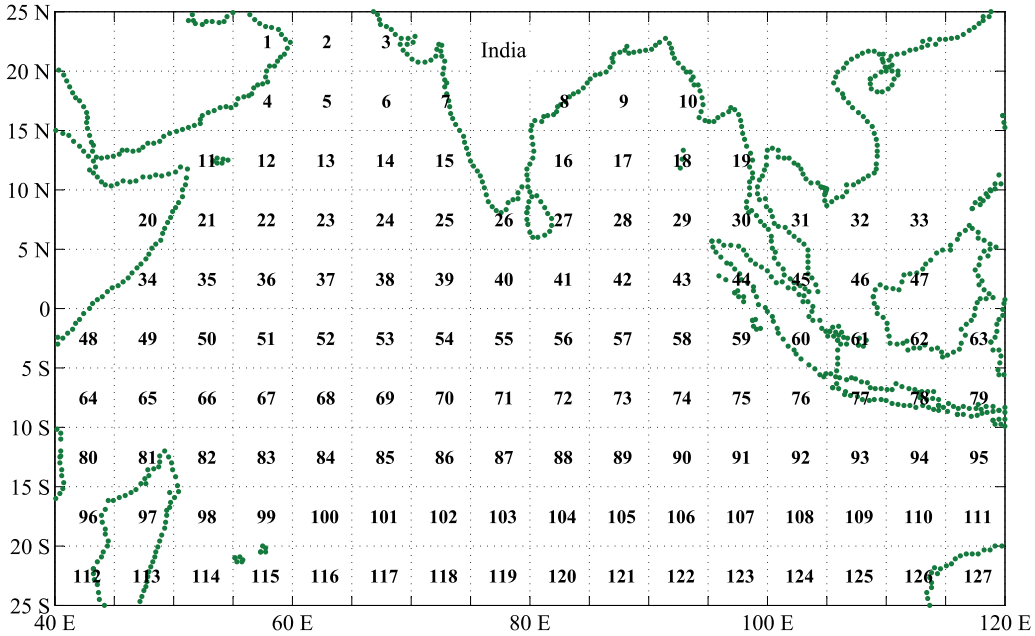


Figure 8. SST grids of size 5° × 5° over the region 40°E–120°E, 25°S–25°N (excluding the land regions) used for correlation analysis.

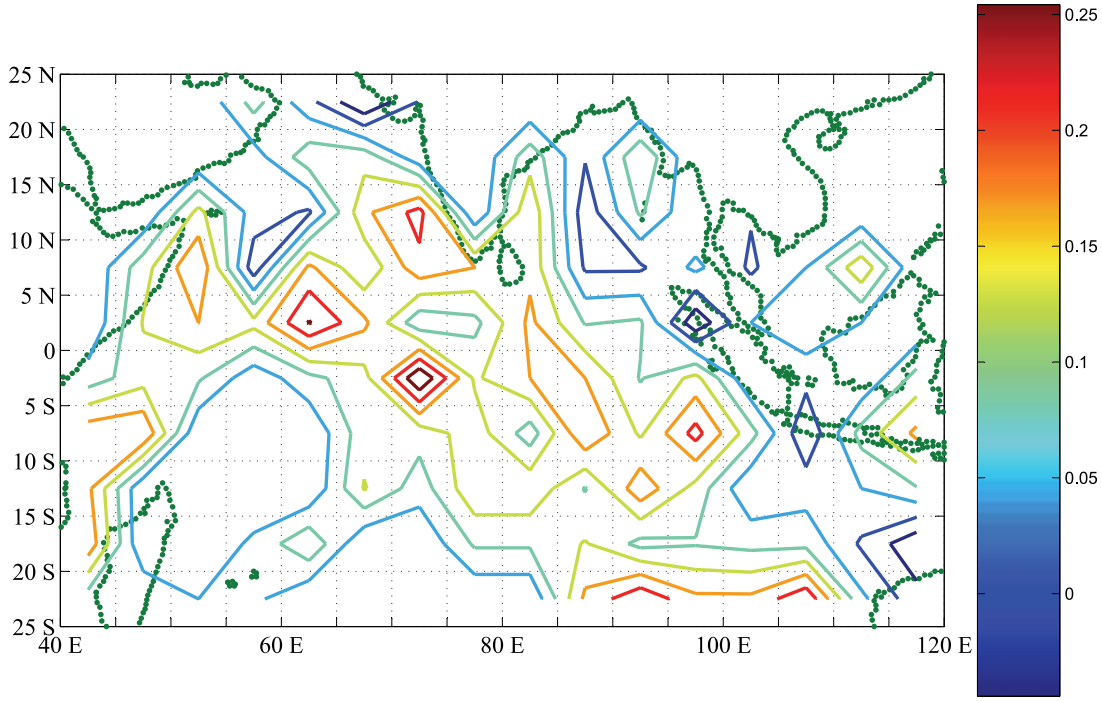


Figure 9. Correlation contours for the northwest region at lag 7.

[41] The time series algorithm employing the concepts of minimal occurrences with constraints and time lags was employed to find the associations between the antecedents and consequent. Climatic indices are considered as the antecedents and the target extreme episodes are considered as the consequents for generating the rules. A variety of window widths, time lags, frequency thresholds and confidence thresholds were tried to find the frequent episodes and rules. To assess the goodness (value) of a rule, both confidence and J measure were used.

4.3. Results and Discussions

4.3.1. Association Rules for Drought

[42] The data-mining algorithm is applied to find the association rules for all the regions and also for All India based on the data from 1960 to 1982 (23 years). A confidence threshold of 0.7 and a minimum J measure of

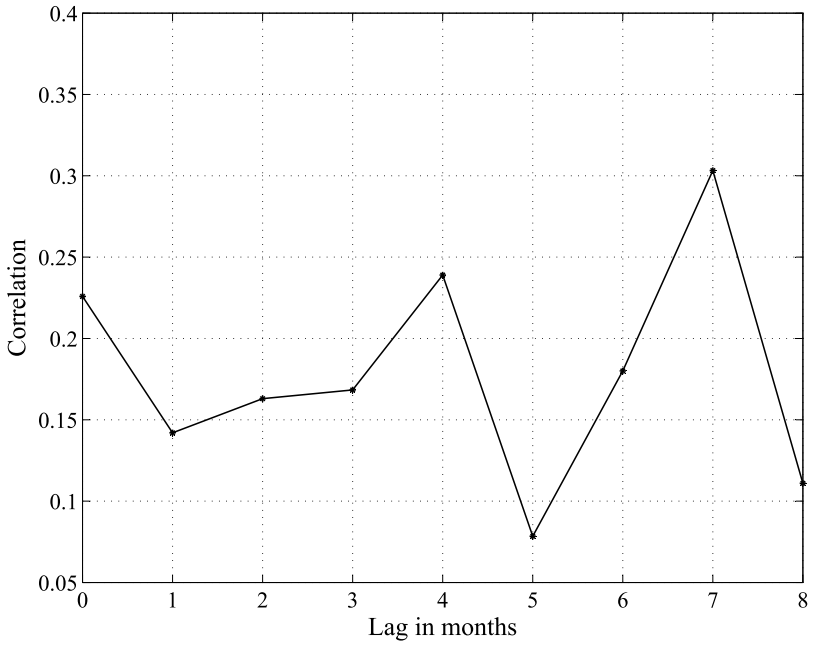


Figure 10. Variation of maximum correlation with respect to lag for the northwest region.

Table 3. Most Influencing Grids and Maximum Correlation for Each Region

Region	Grid	Correlation
Northwest	15, 37, 54, 75	0.2–0.27
West central	15, 21, 27, 35	0.25–0.31
Central northeast	21, 35, 56, 75	0.2–0.29
Northeast	105, 107, 123, 124	0.2–0.27
Peninsular	10, 21, 26, 31	0.2–0.24
All India	37, 72, 73, 74	0.25–0.32

0.025 were used for the extraction of frequent rules. It is found that rules with maximum confidence level and J measure were obtained for an antecedent window width of 4 months, consequent window width of 1 month and time lag of 7 months. The selected rules thus generated are shown in the Table 5.

[43] For almost all regions, a combination of DSLP, NAO, Nino 3.4 and SST values of the respective most influencing grids is causing drought episodes of varying intensities. However, the discrete states of the precursors are different for different regions. For example, a severe drought is occurring in west central region, if DSLP is between 12.5 and 14.0 (i.e., DSLP-6), NAO is between 1.0 and 1.5 (i.e., NAO-5), Nino 3.4 is between 28.0 and 28.5 (i.e., Nino-6) and lowest SST values (i.e., SST values which are less than 1.5 standard deviation from the average). Also, a combination of DSLP taking values between 12.5 and 14.0 (i.e., DSLP-6), NAO taking values between −1.5 and −1.0 (i.e., NAO-3), Nino 3.4 taking values between 27.5 and 28.0 (i.e., Nino-5) and lowest SST values are causing a moderate drought in the same station. Also, for Peninsular region, if DSLP is between 12.5 and 14.0 (i.e., DSLP-6), NAO is between 1.0 and 1.5 (i.e., NAO-5), Nino 3.4 is between 28.0 and 28.5 (i.e., Nino-6) and lowest SST values at two grids, then extreme drought is occurring with a confidence value of 1.0. Another most repeating rule in almost all regions is the combination of DSLP or NAO, Nino and SST as the precursors of drought. For All India, a severe drought is preceded by a combination of NAO, Nino 3.4, and low SST values and a moderate drought is preceded by a combination of DSLP, Nino 3.4 and low SST values. It can be noted that for all the repeating rules, the discrete states taken by the precursors are different for different regions.

[44] The rules generated are clearly showing a negative relation with DSLP and Nino 3.4 and also a positive relation with the SST. However, there is no such specific relation showing up with NAO. For example, rule 3 of central northeast region and rule 1 of northeast region are showing both positive and negative NAO values as the precursors of drought episodes.

[45] For some regions like northwest, west central and All India, no rules for extreme drought show up. The reason for this may be either no frequent episodes of antecedents are preceding the consequent or the rules for extreme drought are not above the given threshold for confidence and J measure.

4.3.2. Association Rules for Flood

[46] The data-mining algorithm is applied to find the association rules using the data from 1960 to 1982 specifying target episodes as moderate flood, severe flood and extreme flood. As in the previous case, rules with maximum confidence level and J measure were obtained for an antecedent window width of 4 months, consequent window width of 1 month and time lag of 7 months. The selected rules thus generated are shown in the Table 6.

[47] The combination of precursors is different for each region, with indices DSLP and NAO appearing in rules for almost all regions. A combination of DLSP, NAO and high SST values are causing flood of varying intensities in almost all the regions. Considering All India rainfall, higher SST conditions, DSLP value between 6.0 and 7.5 (i.e., DSLP-2) and NAO value less than −1.5 (i.e., NAO-1 and NAO-2) occurring within 4 months is succeeded by extreme flood at a lag of 7 months. Severe flood is preceded by a combination of the higher SST conditions, a DSLP value between 6.0 and 7.5 (i.e., DSLP-2) and NAO value between 1.5 and 2.5 (i.e., NAO-6) occurring within 4 months. Rules generated for flood also show a negative correlation of rainfall with DSLP and Nino 3.4. Here also, rules for extreme flood did not show up for northeast and peninsular regions.

4.4. Validation of the Rules

[48] In order to validate and to check the consistency of the rules generated, the data-mining algorithm is again used to generate rules for drought and flood using the data for the years 1983–2005 (23 years). The threshold values for confidence and J measure are kept same for rule extraction. As in training, rules with maximum confidence level and J measure were obtained for an antecedent window width of 4 months, consequent window width of 1 month and time lag of 7 months. The rules generated for drought and flood are shown in Tables 7 and 8 respectively.

[49] A comparison of the rules generated during the calibration period and the validation period shows that almost all the rules for both drought and flood are following the same combination of antecedents for the corresponding consequent with slight change in the values of confidence and J measure. The variations in these interestingness

Table 4. Threshold Values Used for the Categorization of Climatic Indices (Predictors)[a]

Index	1	2	3	4	5	6	7
DSLP (mb)	≤ 6.0	$6.0 < X \leq 7.5$	$7.5 < X \leq 8.5$	$8.5 < X < 11.5$	$11.5 \leq X < 12.5$	$12.5 \leq X < 14.0$	≥ 14.0
NAO	< -2.5	$-2.5 < X \leq -1.5$	$-1.5 < X \leq -1.0$	$-1.0 < X < 1.0$	$1.0 \leq X < 1.5$	$1.5 \leq X < 2.5$	≥ 2.5
Nino 3.4 (°C)	≤ 25.5	$25.5 < X \leq 26.0$	$26.0 < X \leq 26.5$	$26.5 < X < 27.5$	$27.5 \leq X < 28.0$	$28.0 \leq X < 28.5$	≥ 28.5

[a]DSLP, Darwin sea level pressure; NAO, North Atlantic Oscillation; Nino 3.4, Nino 3.4 SST.

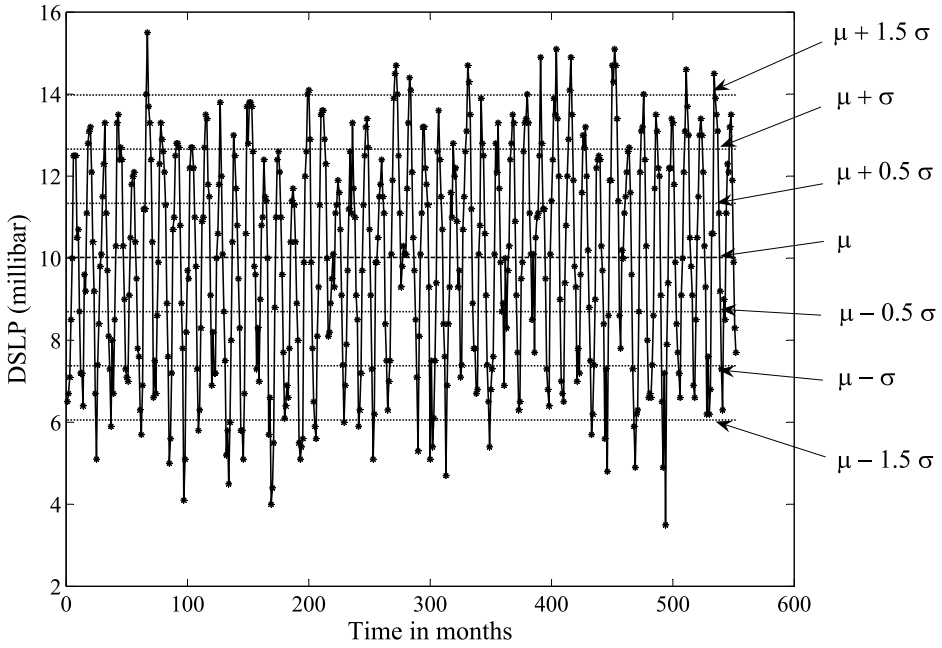


Figure 11. Darwin sea level pressure for the period 1960–2005 indicating the threshold values of discrete classes.

measures are mainly due to the difference in the number of consequent episodes occurring in the calibration and validation periods.

[50] A considerable deviation from the calibration rules is only in the association rules for drought for central northeast and peninsular regions, in which different SST grids are showing low SST values in the validation period. Instead of grid 56 showing a low SST value in rule 2 of central northeast, grid 35 is showing a low SST value during validation period. Also, for rule 3, grid 56 is replaced by grid 75. Similarly, in rule 1 of Peninsular region, grid 21 is replaced by low SST values of grid 31.

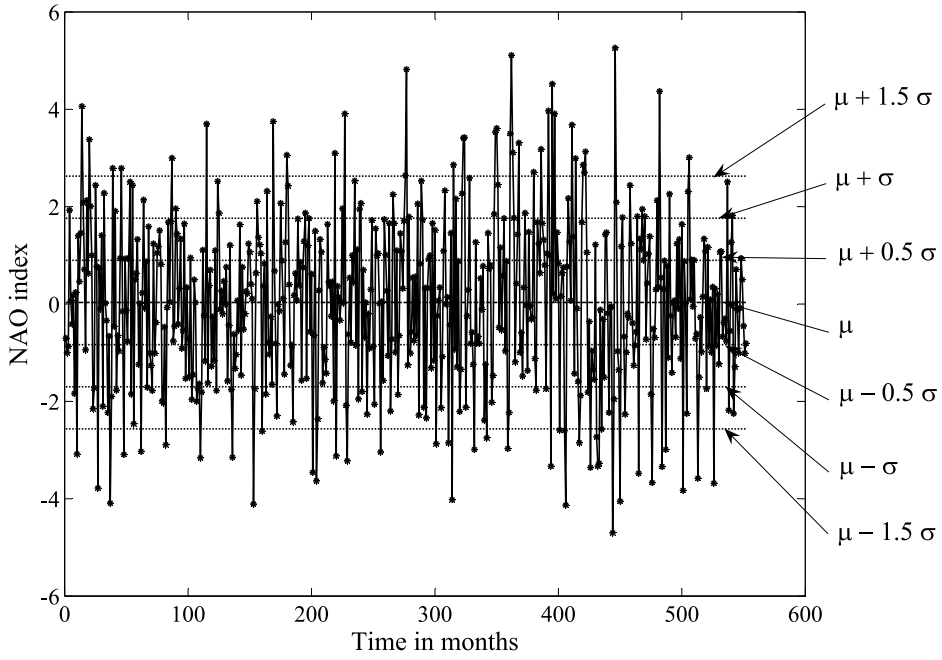


Figure 12. North Atlantic Oscillation for the period 1960–2005 indicating the threshold values of discrete classes.

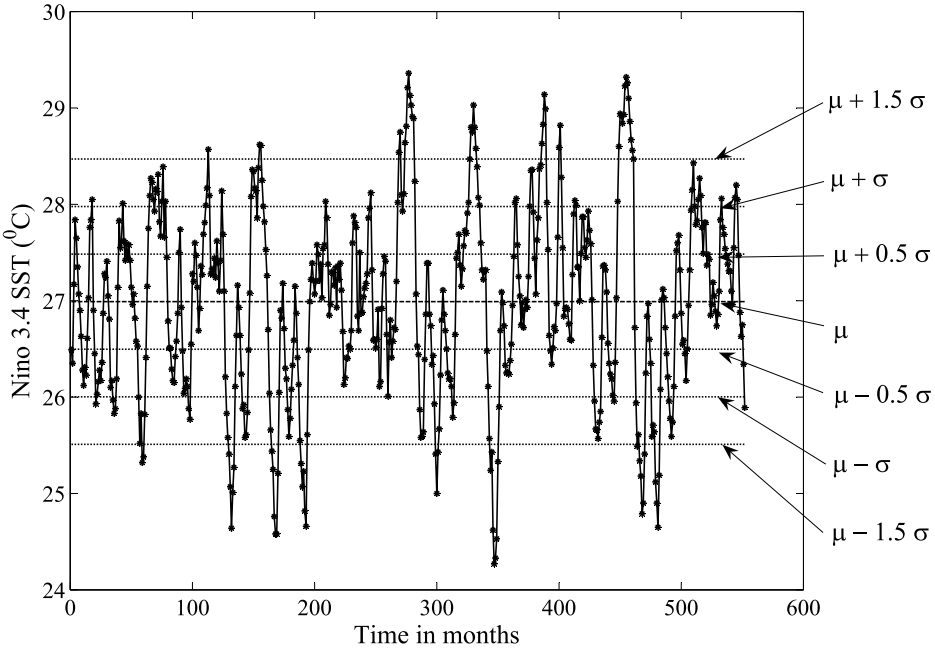


Figure 13. Nino 3.4 SST for the period 1960–2005 indicating the threshold values of discrete classes.

[51] For all other regions, the combinations are exactly the same for both drought and flood rules, with only a slight deviation in the discrete classes taken by the antecedents. For example, in northwest region, instead of the combination of NAO-6, Nino-5, SSTgrid37-3, SSTgrid54-3 as a precursor of severe drought, during validation period a combination of NAO-5, Nino-6, SSTgrid37-3, SSTgrid54-2 is causing severe drought. A comparison of the values taken by these indices reveals that they are taking nearby discrete classes during calibration and validation periods. Similarly, analyzing drought rules for other regions and also flood rules, it can be seen that even if the rules are not satisfied exactly, nearby classes of the indices specified in the rules are always preceding the target episodes. This necessitates the need for a flexible allotment of the classes for the indices.

[52] A further analysis is done by extracting rules using the data for the years 1960–2000 (41 years) to affirm the rules generated during the calibration period (1960–1982). It is found that these rules are also exactly following the same combination as that of the calibration period. The validation of these rules is conducted using the data for the years 2001–2005. Drought episodes occur almost in all regions during the years 2002 and 2004. Severe drought episodes in these years for All India are preceded by NAO-5, Nino-5, Nino-6 and lower SST values and are in accordance to the drought rules generated for All India. Similarly, analyzing drought rules for other regions and also

Table 5. Selected Association Rules for Drought[a]

Region	Rule	Antecedent	Consequent	Confidence	J measure
Northwest	1	NAO-6, Nino-5, SSTgrid37-3, SSTgrid54-3	Severe drought	0.75	0.038
	2	DSLP-5, Nino-5, SSTgrid54-1, SSTgrid54-3	Moderate drought	0.75	0.0394
West central	1	DSLP-6, NAO-5, Nino-6, SSTgrid15-2	Severe drought	0.75	0.038
	2	DSLP-6, NAO-3, Nino-5, SSTgrid21-2	Moderate drought	1.0	0.043
Central northeast	1	DSLP-6, Nino-6, SSTgrid21-2, SSTgrid56-2	Extreme drought	1.0	0.034
	2	DSLP-6, Nino-5, SSTgrid21-2, SSTgrid56-3	Severe drought	1.0	0.036
	3	NAO-3, NAO-6, Nino-6, SSTgrid56-3	Moderate drought	1.0	0.035
Northeast	1	DSLP-5, NAO-3, NAO-5, SSTgrid105-2, SSTgrid123-2	Extreme drought	1.0	0.036
	2	DSLP-6, NAO-2, Nino-6, SSTgrid107-2, SSTgrid123-2	Severe drought	1.0	0.038
	3	NAO-7, SSTgrid105-3, SSTgrid107-3, SSTgrid123-2	Moderate drought	1.0	0.0484
Peninsular	1	DSLP-6, NAO-5, Nino-6, SSTgrid21-2, SSTgrid26-2	Extreme drought	1.0	0.039
	2	DSLP-6, Nino-5, Nino-6, SSTgrid26-3, SSTgrid31-2	Severe drought	1.0	0.038
	3	NAO-5, Nino-5, Nino-6, SSTgrid21-3, SSTgrid26-2	Moderate drought	0.75	0.038
All India	1	NAO-5, Nino-5, Nino-6, SSTgrid72-3, SSTgrid74-3	Severe drought	.0.75	0.031
	2	DSLP-5, Nino-5, SSTgrid73-2, SSTgrid74-1	Moderate drought	1.0	0.056

[a]DSLP, Darwin sea level pressure; NAO, North Atlantic Oscillation; Nino, Nino 3.4 SST; SST, sea surface temperature.

Table 6. Selected Association Rules for Flood

Region	Rule	Antecedent	Consequent	Confidence	J measure
Northwest	1	DSLP-3, NAO-5, SSTgrid15-6, SSTgrid54-5, SSTgrid75-5	Extreme flood	1.0	0.038
	2	DSLP-1, NAO-6, Nino-2, Nino-3, SSTgrid75-5	Severe flood	1.0	0.0376
	3	DSLP-2, DSLP-3, NAO-3, NAO-6, SSTgrid37-5	Moderate flood	0.75	0.038
West central	1	NAO-3, NAO-6, SSTgrid21-5, SSTgrid27-6	Extreme flood	1.0	0.036
	2	DSLP-3, NAO-3, SSTgrid15-6, SSTgrid27-5, SSTgrid35-5	Severe flood	1.0	0.034
	3	DSLP-1, DSLP-2, NAO-1, SSTgrid35-5	Moderate flood	0.8	0.052
Central northeast	1	DSLP-3, NAO-3, SSTgrid21-5, SSTgrid56-5, SSTgrid75-7	Extreme flood	1.0	0.035
	2	DSLP-3, NAO-3, Nino-3, SSTgrid21-5, SSTgrid75-6	Severe flood	1.0	0.042
	3	DSLP-1, Nino-2, Nino-3, SSTgrid35-5	Moderate flood	1.0	0.060
Northeast	1	DSLP-2, NAO-6, Nino-3, SSTgrid107-5, SSTgrid123-6, SSTgrid124-5	Severe flood	1.0	0.038
	2	DSLP-1, DSLP-2, NAO-2, SSTgrid124-5	Moderate flood	0.75	0.036
Peninsular	1	NAO-3, Nino-3, SSTgrid10-5, SSTgrid21-5, SSTgrid26-6	Severe flood	0.75	0.043
	2	DSLP-1, NAO-6, SSTgrid10-5, SSTgrid26-6	Moderate flood	1.0	0.029
All India	1	DSLP-2, NAO-1, NAO-2, SSTgrid37-6	Extreme flood	1.0	0.058
	2	DSLP-2, NAO-6, SSTgrid37-5, SSTgrid72-5, SSTgrid73-5	Severe flood	0.83	0.063
	3	DSLP-2, NAO-7, Nino-3, SSTgrid37-5, SSTgrid72-5	Moderate flood	0.75	0.037

flood rules, it can be seen that almost all the drought and flood episodes are preceded by the exact combinations of the climatic indices shown by the respective rules. In all the cases, either the indices are taking the exact values mentioned in the rules or at least they are taking the nearby classes of the indices specified in the rules. This again demands for a flexible allotment of the classes for the indices. Instead of defining the classes with abrupt and well defined boundaries, a vague and ambiguous boundary by making use of the concept of fuzzy sets, can be used for classifying the indices into different sets.

5. Conclusions

[53] Data mining is a powerful technology to *extract the hidden predictive information from databases* thus helping in the prediction of future trends and behaviors. Data-mining tools scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. Implementing this technology in the extraction of association rules for the extreme conditions may help decision makers to improve their fundamental scientific understanding of drought, about its causes, predictability, impacts, mitigation actions, planning methodologies, and policy alternatives.

[54] Various rules generated for each region and also for All India clearly indicate a strong relationship with climatic indices chosen, i.e., DSLP, NAO, Nino 3.4 and SST values. From the rules extracted, it can be seen that almost all the climatic indices mentioned above are occurring as antecedents for drought episodes, with different combinations and confidence values. However, for rules extracted for flood episodes, the combinations with Nino 3.4 are confronted only a few times.

[55] The validation of the rules, using the data from 1983 to 2005, shows good consistency of the rules in the validation period. Almost all rules are exactly following the same combination as that of the calibration period rules. For some of the rules, although the combination of the indices mentioned is followed during the validation period, one or two climatic indices which are indicated as the precursors to the extremes in the rules are not falling in the same discrete range specified in the training period rules or in other words they are taking the nearby discrete states. Thus a better extraction of the rules may be possible if the classification of the indices is done in a fuzzy manner and not in a crisp manner. This fuzzy aspect can be taken up as

Table 7. Selected Association Rules for Drought for Validation Period

Region	Rule	Antecedent	Consequent	Confidence	J measure
Northwest	1	NAO-5, Nino-6, SSTgrid37-3, SSTgrid54-2	Severe drought	0.75	0.032
	2	DSLP-6, Nino-5, SSTgrid54-1, SSTgrid54-2	Moderate drought	0.75	0.034
West central	1	DSLP-7, NAO-5, Nino-6, SSTgrid15-3	Severe drought	1.0	0.060
	2	DSLP-5, NAO-3, Nino-6, SSTgrid21-2	Moderate drought	0.75	0.043
Central northeast	1	DSLP-5, Nino-6, SSTgrid21-1, SSTgrid56-3	Extreme drought	1.0	0.081
	2	DSLP-5, Nino-5, SSTgrid21-2, SSTgrid35-3	Severe drought	0.8	0.057
	3	NAO-3, NAO-7, Nino-6, SSTgrid75-1	Moderate drought	1.0	0.054
Northeast	1	DSLP-5, NAO-3, NAO-5, SSTgrid105-2, SSTgrid123-2	Extreme drought	1.0	0.052
	2	DSLP-6, NAO-3, Nino-5, SSTgrid107-2, SSTgrid123-2	Severe drought	1.0	0.034
	3	NAO-6, SSTgrid105-3, SSTgrid107-1, SSTgrid123-2	Moderate drought	0.75	0.037
Peninsular	1	DSLP-6, NAO-5, Nino-6, SSTgrid26-3, SSTgrid31-2	Extreme drought	1.0	0.039
	2	DSLP-6, Nino-5, Nino-6, SSTgrid26-2, SSTgrid31-2	Severe drought	1.0	0.056
	3	NAO-5, Nino-5, Nino-6, SSTgrid21-3, SSTgrid26-3	Moderate drought	1.0	0.031
All India	1	NAO-5, Nino-5, Nino-6, SSTgrid72-2, SSTgrid74-3	Severe drought	0.75	0.037
	2	DSLP-5, Nino-6, SSTgrid73-1, SSTgrid74-2	Moderate drought	1.0	0.047

Table 8. Selected Association Rules for Flood for Validation Period

Region	Rule	Antecedent	Consequent	Confidence	J measure
Northwest	1	DSLP-3, NAO-5, SSTgrid15-5, SSTgrid54-5, SSTgrid75-7	Extreme flood	1.0	0.038
	2	DSLP-1, NAO-6, Nino-2, Nino-3, SSTgrid75-6	Severe flood	1.0	0.041
	3	DSLP-2, DSLP-3, NAO-3, NAO-6, SSTgrid37-5	Moderate flood	0.71	0.031
West central	1	NAO-3, NAO-6, SSTgrid21-7, SSTgrid27-7	Extreme flood	1.0	0.042
	2	DSLP-3, NAO-2, SSTgrid15-5, SSTgrid27-5, SSTgrid35-7	Severe flood	1.0	0.039
	3	DSLP-1, DSLP-2, NAO-1, SSTgrid35-6	Moderate flood	0.8	0.046
Central northeast	1	DSLP-3, NAO-3, SSTgrid21-7, SSTgrid56-6, SSTgrid75-5	Extreme flood	1.0	0.045
	2	DSLP-3, NAO-3, Nino-3, SSTgrid21-5, SSTgrid75-6	Severe flood	1.0	0.035
	3	DSLP-2, Nino-2, Nino-3, SSTgrid35-7	Moderate flood	1.0	0.044
Northeast	1	DSLP-2, NAO-6, Nino-3, SSTgrid107-6, SSTgrid123-6, SSTgrid124-5	Severe flood	1.0	0.038
	2	DSLP-1, DSLP-2, NAO-2, SSTgrid124-6	Moderate flood	0.75	0.039
Peninsular	1	NAO-3, Nino-2, SSTgrid10-7, SSTgrid21-5, SSTgrid26-5	Severe flood	1.0	0.043
	2	DSLP-2, NAO-6, SSTgrid10-5, SSTgrid26-5	Moderate flood	1.0	0.034
All India	1	DSLP-2, NAO-2, NAO-3, SSTgrid37-7	Extreme flood	1.0	0.048
	2	DSLP-2, NAO-7, SSTgrid37-6, SSTgrid72-5, SSTgrid73-6	Severe flood	0.75	0.040
	3	DSLP-2, NAO-7, Nino-3, SSTgrid37-5, SSTgrid72-5	Moderate flood	1.0	0.030

further study. Introducing an uncertainty in the classes taken by the indices will help in improving the quality and confidence of the rules generated. Inclusion of other climatic and oceanic indices may also improve the quality of the rules in identifying the relationships with the extreme episodes.

References

Bayardo, R. J., and R. Agarwal (1999), Mining the most interesting rules, in *Proceedings of the Fifth ACM International Conference on Knowledge Discovery and Data Mining*, edited by S. Chaudhuri and D. Madigan, pp. 145–154, Assoc. for Comput. Mach., San Diego, Calif.

Chattopadhyay, J., and R. Bhatla (2002), Possible influence of QBO on teleconnections relating Indian summer monsoon rainfall and sea-surface temperature anomalies across the equatorial Pacific, *Int. J. Climatol.*, *22*(1), 121–127.

Clark, C. O., J. E. Cole, and P. J. Webster (2000), Indian ocean SST and Indian summer rainfall: Predictive relationships and their decadal variability, *J. Clim.*, *13*(14), 2503–2519.

Das, G., K. I. Lin, H. Mannila, G. Ranganathan, and P. Smyth (1998), Rule discovery from time series, in *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, edited by R. Agrawal and P. Stolorz, pp. 16–22, AAAI Press, Menlo Park, Calif.

DelSole, T., and J. Shukla (2002), Linear prediction of Indian monsoon rainfall, *J. Clim.*, *15*(24), 3645–3658.

DelSole, T., and J. Shukla (2006), Linear prediction of monsoon rainfall for Indian subdivisions, *COLA Tech. Rep. 208*, 19 pp., Center for Ocean-Land Atmosphere Studies, Calverton, Md.

Dube, S. K., M. E. Luther, and J. O'Brien (1990), Relationships between interannual variability in the Arabian Sea and Indian summer monsoon rainfall, *Meteorol. Atmos. Phys.*, *44*(1–4), 153–165.

Gadgil, S., and S. Gadgil (2006), The Indian monsoon—GDP and agriculture, *Econ. Polit. Weekly*, *41*(47), 4887–4895.

Gadgil, S., M. Rajeevan, and R. Nanjunduah (2005), Monsoon prediction: Why yet another failure, *Curr. Sci.*, *88*(9), 1389–1400.

Gadgil, S., M. Rajeevan, and P. A. Francis (2007), Monsoon variability: Links to major oscillations over the equatorial Pacific and Indian oceans, *Curr. Sci.*, *93*(2), 182–194.

Goddard, L., S. J. Mason, S. E. Zebiak, C. F. Ropelewski, R. Basher, and M. A. Cane (2001), Current approaches to seasonal to inter-annual climate predictions, *Int. J. Climatol.*, *21*(9), 1111–1152, doi:10.1002/joc636.

Han, J., and M. Kamber (2006), *Data Mining: Concepts and Techniques*, 770 pp., Elsevier, New York.

Harms, S. K., and J. S. Deogun (2004), Sequential association rule mining with time lags, *J. Intelligent Inf. Syst.*, *22*(1), 7–22.

Harms, S. K., J. Deogun, and T. Tadesse (2002), Discovering sequential rules with constraints and time lags in multiple sequences, in *Proc. 2002 Int. Symposium on Methodologies for Intelligent Systems*, edited by M. S. Hacid et al., pp. 432–441, ISMIS, Lyon, France.

Iyengar, R. N., and S. T. G. Raghu Kanth (2004), Intrinsic mode functions and a strategy for forecasting Indian monsoon rainfall, *Meteorol. Atmos. Phys.*, doi:10.1007/s00703-004-0089-4.

Maity, R., and D. Nagesh Kumar (2006), Bayesian dynamic modeling for monthly Indian summer monsoon rainfall using El Nino-Southern Oscillation (ENSO) and Equatorial Indian Ocean Oscillation (EQUINOO), *J. Geophys. Res.*, *111*, D07104, doi:10.1029/2005JD006539.

Maity, R., and D. Nagesh Kumar (2008), Probabilistic prediction of hydroclimatic variables with nonparametric quantification of uncertainty, *J. Geophys. Res.*, *113*, D14105, doi:10.1029/2008JD009856.

Mannila, H., H. Toivonen, and A. I. Verkamo (1997), Discovery of frequent episodes in event sequences, *Data Min. Knowledge Discovery*, *1*(3), 259–289.

Navone, H. D., and H. A. Ceccatto (1994), Predicting Indian monsoon rainfall: A neural network approach, *Clim. Dyn.*, *10*(6–7), 305–312, doi:10.1007/BF00228029.

Nicholls, N. (1995), All-India summer monsoon rainfall and sea surface temperatures around Northern Australia and Indonesia, *J. Clim.*, *8*(5), 1463–1467.

Piatetsky-Shapiro, G. (1991), Discovery, analysis, and presentation of strong rules, in *Knowledge Discovery in Databases*, edited by G. Piatetsky-Shapiro and W. J. Frawley, pp. 229–248, AAAI Press, Menlo Park, Calif.

Rajeevan, M., D. S. Pai, S. K. Diskhit, and R. R. Kelkar (2004), IMD's new operational models for long range forecast of southwest monsoon rainfall over India and their verification for 2003, *Curr. Sci.*, *86*(3), 422–430.

Rajeevan, M., D. S. Pai, R. Anil Kumar, and B. Lal (2006), New statistical models for long range forecasting of southwest monsoon rainfall over India, *Clim. Dyn.*, *28*(7–8), 813–828, doi:10.1007/s00382-006-0197-6.

Reddy, P. R., and P. S. Salvekar (2003), Equatorial east Indian Ocean sea surface temperature: A new predictor for seasonal and annual rainfall, *Curr. Sci.*, *85*(11), 1600–1604.

Sadhuram, Y. (2006), Long range forecast of southwest monthly rainfall over India during summer monsoon season using SST in the north Indian Ocean, *Curr. Sci.*, *91*(4), 425–428.

Shukla, J., and D. A. Mooley (1987), Empirical prediction of summer monsoon rainfall over India, *Mon. Weather Rev.*, *115*, 695–703.

Smyth, P., and R. M. Goodman (1991), Rule induction using information theory, in *Knowledge Discovery in Databases*, edited by G. Piatetsky-Shapiro and W. J. Frawley, pp. 159–176, MIT, Cambridge, Mass.

Webster, P. J., V. O. Magana, T. N. Palmer, and J. Shukla (1998), Monsoons: Processes, predictability, and the prospects for prediction, *J. Geophys. Res.*, *103*(C7), 14,451–14,510.

C. T. Dhanya and D. Nagesh Kumar, Department of Civil Engineering, Indian Institute of Science, Bangalore, Karnataka 560012, India. (dhanya@civil.iisc.ernet.in; nagesh@civil.iisc.ernet.in)