2005

Discovering Associations between Climatic and Oceanic Parameters to Monitor Drought in Nebraska Using Data-Mining Techniques

Tsegaye Tadesse
University of Nebraska-Lincoln, ttadesse2@unl.edu

Donald A. Wilhite
University of Nebraska - Lincoln, dwilhite2@unl.edu

Michael J. Hayes
University of Nebraska-Lincoln, mhayes2@unl.edu

Steve Goddard
University of Nebraska-Lincoln, goddard@cse.unl.edu

Discovering Associations between Climatic and Oceanic Parameters to Monitor Drought in Nebraska Using Data-Mining Techniques

TSEGAYE TADESSE, DONALD A. WILHITE, MICHAEL J. HAYES

National Drought Mitigation Center, University of Nebraska at Lincoln, Lincoln, Nebraska

SHERRI K. HARMS

Department of Computer Science and Information Systems, University of Nebraska at Kearney, Kearney, Nebraska

STEVE GODDARD

Department of Computer Science and Engineering, University of Nebraska at Lincoln, Lincoln, Nebraska

(Manuscript received 4 August 2003, in final form 12 October 2004)

ABSTRACT

Drought is a complex natural hazard that is best characterized by multiple climatological and hydrological parameters. Improving our understanding of the relationships between these parameters is necessary to reduce the impacts of drought. Data mining is a recently developed technique that can be used to interact with large databases and assist in the discovery of associations between drought and oceanic data by extracting information from massive and multiple data archives.

In this study, a new data-mining algorithm [i.e., Minimal Occurrences With Constraints and Time Lags (MOWCATL)] has been used to identify the relationships between oceanic parameters and drought indices. Rather than using traditional global statistical associations, the algorithm identifies drought episodes separate from normal and wet conditions and then uses drought episodes to find time-lagged relationships with oceanic parameters. As with all association-based data-mining algorithms, MOWCATL is used to find existing relationships in the data, and is not by itself a prediction tool.

Using the MOWCATL algorithm, the analyses of the rules generated for selected stations and state-averaged data for Nebraska from 1950 to 1999 indicate that most occurrences of drought are preceded by positive values of the Southern Oscillation index (SOI), negative values of the multivariate ENSO index (MEI), negative values of the Pacific–North American (PNA) index, negative values of the Pacific decadal oscillation (PDO), and negative values of the North Atlantic Oscillation (NAO). The frequency and confidence of the time-lagged relationships between oceanic indices and droughts at the selected stations in Nebraska indicate that oceanic parameters can be used as indicators of drought in Nebraska.

1. Introduction

Decision makers and planners need to understand the impacts related to various levels of drought severity and what conditions are associated with drought in order to take appropriate actions in proactive management of water and other natural resources during drought (Svoboda et al. 2002; Wilhite 2000a). For this reason, a thorough understanding of drought's associations with climatic, oceanic, and environmental parameters in a specific area is essential to combating the effects of drought in a proactive manner by addressing vulnerabilities through a risk management approach.

Monitoring drought involves considerations of past and present climatological conditions. Nebraska has experienced numerous drought episodes of various intensities and duration in the past 100 yr. Historical records show that Nebraska has sustained major multiyear droughts, including droughts in the 1930s and 1950s. During the "Dust Bowl" years, which refer to drought conditions from 1934 to 1941 over the Great Plains region, Nebraska had the longest period of hot summers ever recorded (Dewey 1996). In the 1950s, drought and heat waves peaked in the summers of 1953, 1954, and 1955. A multiyear drought also occurred in the 1970s, although it was not as severe or as long in duration as the 1950s. During the 1974 growing season, precipitation averaged 68% of normal throughout the state, resulting in a 28% overall crop production deficit. Below-normal climatic seasonal rainfall conditions continued through 1976 resulting in an annual precipitation deficit in Nebraska during that time.

Corresponding author address: Dr. Tsegaye Tadesse, National Drought Mitigation Center, University of Nebraska at Lincoln, Lincoln, NE 68583-0749.
E-mail: ttadesse2@unl.edu

Recent records show that 1988–89 and 1995–96 were drier than normal in many parts of the state. Drought frequency and intensity in the 1980s and 1990s were relatively low as compared to the 1930s, 1950s, and 1970s. However, during the 10-yr period from 1989 to 1998, the indemnity paid for drought losses to crops in Nebraska totaled more than $92 million [the U.S. Department of Agriculture Risk Management Agency (USDA RMA) 1999]. This implies that if droughts are more frequent in the future than those of the 1980s and 1990s, the losses could relatively increase. Thus, since droughts are recurrent natural phenomena in Nebraska, it is clear that proactive steps should be taken to address the impacts of future drought episodes.

Studying past and present droughts in relation to climatological, oceanic, and atmospheric parameters could help mitigate future drought impacts on society by improving our understanding of the drought hazard. Based on the observed data analysis, climate change researchers have indicated that in regions where precipitation decreases in midlatitudes during summer, or where snowfall decreases, the combined effect might result in substantial increases in drought frequency in the twenty-first century (Rind et al. 1990).

Drought, although it is a normal climatic phenomenon, is generally infrequent. For example, if one uses the standardized precipitation index (SPI) to identify dryness with values less than -0.99, it means that the occurrence of drought is less than one standard deviation from the climatic normal, which occurs less than 16% of the time (McKee et al. 1995). This is due to the fact that 68% of the area under a normal frequency distribution curve is within ± 1 standard deviation of the mean (i.e., 16% of the distribution area is less than the mean -1 standard deviation which shows the drought frequency). This infrequent occurrence of drought complicates the correlation or global association of drought with other atmospheric and oceanic parameters.

Temporal and spatial patterns of drought vary from one area to another (Wilhite 2000b; Hayes et al. 1999). For a specific area, if relationships to oceanic and atmospheric parameters can be identified with a certain time lag, taking appropriate action based on the associations may reduce the impacts of future droughts. To monitor drought and help in drought decision making, this study has attempted to identify relationships between drought and oceanic and climatic indices in Nebraska using time series data-mining algorithms.

Among the factors that determine droughts are atmospheric phenomena, such as the atmospheric circulation, and their relationship with ocean dynamics. Based on such relationships, it is important to consider the impacts of the variability of the oceanic parameters while monitoring drought. Generally, the variability of oceanic parameters is relatively slower than the variability of atmospheric parameters. For example, because of the large thermal inertia of oceans, sea surface temperatures change slowly compared to atmospheric temperatures. The slow variations make it relatively easier to monitor oceanic parameters compared to meteorological parameters. The development of models to predict oceanic parameters is also proving to be relatively easier and more efficient (Eden et al. 2002; Lu and Greatbatch 2002; Jacob et al. 2001; Covey et al. 2000). This implies that if the relationship of oceanic parameters and local drought is known, one can use these parameters to monitor developing drought conditions.

Many indices have been developed to measure the variability of oceanic and atmospheric parameters. These indices include the Southern Oscillation index (SOI), the multivariate ENSO index (MEI; Wolter and Timlin 1993), the Pacific–North American (PNA) index (Overland et al. 2002), the Pacific decadal oscillation (PDO; Bond and Harrison 2000; Mantua et al. 1997; Francis and Hare 1994), and the North Atlantic Oscillation (NAO; Hurrell 1995). Thus, this study intends to show that variability of the global oceanic and atmospheric indices can be used as a precursor to local drought by generating association rules relating to drought indices.

Data mining is one of the recent technologies used in handling very large amounts of data and discovering patterns and relationships among the parameters. Previous studies used data mining in business activities such as marketing and fraud detection (Berry and Linoff 2000; Cabena et al. 1998; Groth 1998). Analogous to the business community, the meteorological, climatological, and oceanic databases are growing at unprecedented rates using state-of-the-art instruments that are improved continually to take the most accurate and highest-resolution data at specific stations. These databases are increasingly overwhelmed by the massive amounts of different types of data that need an effective method to be transformed into interpretable knowledge. The handling of large amounts of data and extracting useful information such as drought patterns in space and time can be resolved with data-mining techniques. Although drought results from atmospheric phenomena that are complex and include many parameters, data-mining tools can help in modeling and understanding drought characteristics using the existing large dataset. It also has the potential to assist in proactive drought decision making including mitigation and drought risk management. For this purpose, a time series data-mining algorithm has been used to generate association rules discovering the relationships between drought and oceanic and atmospheric parameters to help in drought monitoring.

2. Data collection and preprocessing

The data used in this study are collected from various sources, including: precipitation, temperature, and soil

moisture data from the High Plains Regional Climate Center (HPRCC); SPI values from the National Drought Mitigation Center (NDMC); Palmer Drought Severity Index (PDSI) values at a station level calculated using station data from the HPRCC; NAO index values from the U.K. Climatic Research Unit, University of East Anglia; PDO and PNA indices from the Joint Institute for the Study of the Atmosphere and Ocean (JISAO), the National Oceanic and Atmospheric Administration (NOAA), and the University of Washington; and SOI and MEI data from the NOAA/Climate Diagnostics Center. All these data are collected and preprocessed to satisfy the format used to run the data-mining algorithm. The period of the study was 50 yr, 1950–99.

3. Time series data-mining and association rule algorithms to identify drought characteristics

Time series data-mining algorithms are being developed for many applications to identify hidden patterns within time series data (Berry and Linoff 2000; Klemettine 1999; Groth 1998). These algorithms are designed to characterize and predict complex, nonperiodic, irregular, chaotic phenomena (Povinelli 2000; Huang and Yu 1999; Keogh and Pazzani 1998). In a real-world application such as drought where time is an essential factor, it is important to study the relationships of the parameters that cause drought using their time series patterns.

Time series data-mining techniques organize data as a sequence of events, with each event having a time of occurrence. In data analysis applications on a sequence of events, one of the main challenges is finding similar situations. This is essential to predict future events and understand the dynamics of the process producing the sequence (Mannila and Seppänen 2001).

In monitoring drought, time series data analyses of meteorological, climatological, and oceanic parameters are used to generate rules that can identify the occurrence of drought. Based on the assumption that ocean–atmosphere relationships have a causal link to drought, global oceanic and atmospheric parameters are considered antecedents while droughts are considered consequents in finding their associations. This study attempts to mine the data by using time series association rules to reveal characteristics of drought for a given area. The data-mining concepts and keywords used in this study are briefly described in the following section.

a. Data discretization and event sequences

Discretizing or clustering refers to segmenting the data into groups of records or clusters that have similar characteristics. To apply algorithms on sequential datasets, the data are preprocessed by normalizing and discretizing to form a sequence of events.

An event sequence is a triple element (t_B, t_D, S) where t_B is the beginning time, t_D is the ending time, and S is the time-ordered sequence of events from beginning to end (Mannila and Toivonen 1995). For example, consider the event sequences of 1-month SPI values for Clay Center, Nebraska, from January to December 1998 shown in Fig. 1a. SPI values show the precipitation deviation from the mean for a given time frame and location. For this application the data was discretized into seven clusters: A is extremely dry (SPI value ≤ -2.0), B is severely dry ($-2.0 <$ SPI value ≤ -1.5), C is moderately dry ($-1.5 <$ SPI value ≤ -0.5), D is normal ($-0.5 <$ SPI value < 0.5), E is moderately wet ($0.5 \leq$ SPI value < 1.5), F is severely wet ($1.5 \leq$ SPI value < 2.0), and G is extremely wet (SPI value ≥ 2.0). The resulting sequences of cluster identifiers are shown in Fig. 1b. This is referred to as an event sequence. Figures 1a,b, respectively, show an example of event sequences before and after the data is categorized. Note that SPI data is precipitation data that is already in a normalized form. Event sequences have subsequences that are defined within a window. For example, in Fig. 1b two temporal windows of 4 months are highlighted. The first window is the subsequence[*E, D, C, D*] from 3 to 6 months, and the second window is the subsequence [*D, C, D, G*] from 4 to 7 months.

b. Episodes and frequency of episodes

An episode in an event sequence is a combination of events with a time-specified order (Mannila and Toivonen 1995). An episode occurs in a sequence if events are consistent with the given order, within a given time bound (window width). Thus, an episode P is a pair (V, *type*), where V is a collection of events. In this pair, the type of episode is called *parallel* if no order is specified in a window, and *serial* if the events of the episode have a fixed order. The frequency of an episode is defined as the number of windows in which the episode occurs divided by the number of all windows in the dataset. With the Minimal Occurrences With Constraints and Time Lags (MOWCATL) algorithm, a user can choose to specify the events and the minimum frequency to be considered. Only the events that meet the user-specified inclusion constraints (e.g., minimum frequency) are used to build episodes (Harms et al. 2002).

c. The MOWCATL algorithm

Using the MOWCATL algorithm, an association rule is defined as "if X then Y," where X is the rule antecedent and Y is its consequent. The support of the antecedent, denoted Support {X}, is the number of times that all the events in the antecedent episode X occur together within a user-defined window width. The support of the antecedent (Support {X}) is also called the rule coverage. An episode is considered frequent if its support meets or exceeds a user-specified minimum support threshold. The support and frequency of a consequent episode are defined similarly.

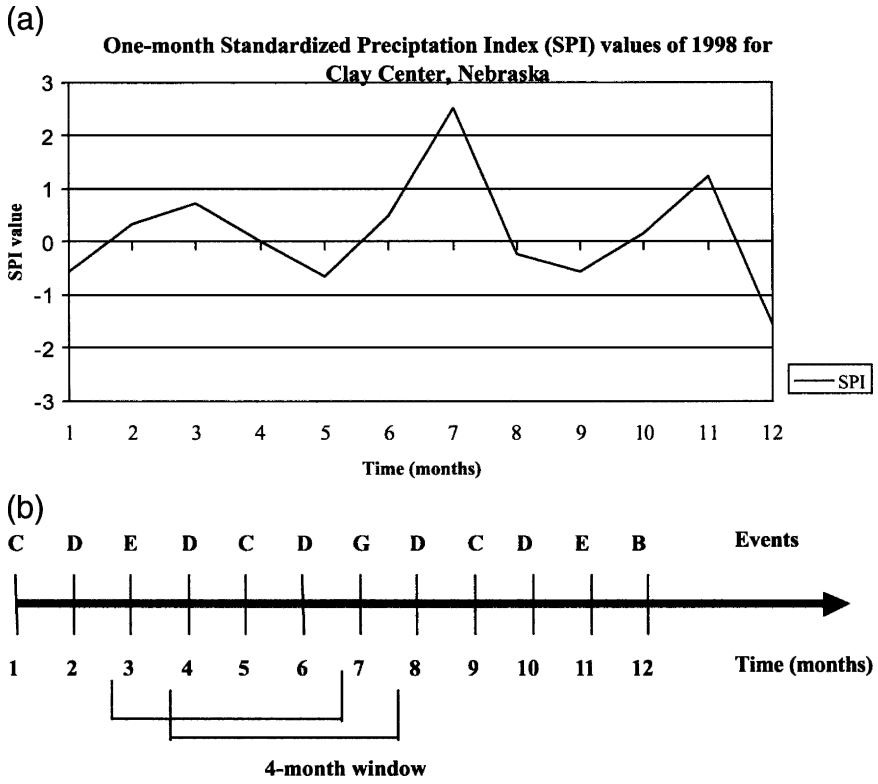

FIG. 1. (a) Example of an event sequence before discretization. (b) Example of an event sequence with two windows of 4-month window size after discretization. The letters represent discretized categories (clusters) while the numbers represent the time (month).

The MOWCATL algorithm first goes through the data file in a database storing the occurrences of the single events (in terms of time order in the case of serial episodes) for the antecedent and consequent events separately. Note that the algorithm only looks for occurrences of events that meet the inclusion constraints. Then it prunes the episodes that do not meet the user-specified minimum support threshold. It then pairs the single-event episodes into episodes of two events, for pairs of events that occur together within the prescribed window width, and records the occurrences of the episode in the dataset. When there are no more events to pair together, this step is finished. The process repeats for episodes of three events, four events, and so on, until there are no episodes left to be combined that meet the minimum support threshold.

After the frequent episodes are found for the antecedent and the consequent independently, MOWCATL combines the frequent episodes to form an episode rule (Harms et al. 2002). The algorithm generates episodal rules where the antecedent episode occurs within a given window width, the 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, assume that episode X is the events A and B, and episode Y is the events C and D. Also, assume that the prescribed antecedent window width is 2 months, the prescribed consequent window width is 3 months, and the time lag is 2 months. The parallel rule generated would indicate that if *A* and *B* occur within 2 months, then within 3 months they will be followed by *C* and *D* occurring together within 2 months.

For parallel rules, an occurrence of a particular rule is recorded when the starting time of the occurrence of the consequent episode follows the starting time of the occurrence of the antecedent episode and differs by at most the time lag. The order of the events in parallel episodes is not important. Thus, parallel rules can be used to see if the events in one episode occur "close" to the events in the other episode.

For serial rules, an occurrence of a particular rule is recorded when the starting time of the occurrence of the consequent episode is no less than the ending time of the occurrence of the antecedent episode, and when the starting time of the occurrence of the consequent episode follows the starting time of the occurrence of the antecedent episode and, at most, differs by the time lag. Moreover, the ending time of the occurrence of the consequent episode must be greater than the ending time of the antecedent.

The type of time lag described above is referred to as

the maximum time lag (Harms et al. 2002). If the maximum time lag is set to 4 months, it finds relationships of the parameters for 4-, 3-, 2-, and 1-month periods to generate rules. The advantage of this constraint in the MOWCATL algorithm is that it finds the hidden relationships between episodes that occur close together, but not always at exactly the same time difference.

Alternately, the user may specify a fixed time lag to count occurrences of a particular rule. With the fixed time lag, the start of the occurrence of the consequent episode must follow the start of the corresponding occurrence of the antecedent episode by exactly the fixed time lag. This is useful when finding fixed interval associations between the antecedent and consequent datasets. Thus, this may be used to find relationships when the specific oceanic index observed a fixed number of months before the associated occurrence of a drought.

The MOWCATL algorithm generates the rules and counts the occurrences of the individual rules as described above, based on the user-specified parameters. The support of the rule is the number of times the rule holds in the dataset. The algorithm prunes the rules that do not meet the predetermined minimum support (Harms et al. 2002).

The confidence of an episode rule with prescribed window widths for the antecedent and the consequent and a prescribed lag between the antecedent and the consequent is the conditional probability that the consequent occurs given that the antecedent occurs, under the time constraints specified. For example, if the events in episode X occurred together 16% of the time and the events in episodes X and Y occurred together 10% of the time, within the prescribed window widths and time lag, then the confidence is 10/16 (63%).

The confidence of the rule as defined in the data-mining context is the ratio of Support{X and Y} divided by the Support of {X} which depends on the occurrences of the X and Y while the traditional statistical confidence considers the global conditions, which considers the whole dataset including the events of X and Y. For example, if we consider a rule that has only MEI as X and PDSI as Y, then the confidence defined in this case is based on frequency of occurrences of these two parameters only. Thus, the confidence values that are generated for the rules may not be statistically significant in the traditional sense. For this reason, other interesting (goodness of the rule) measures are used to select the best rules in addition to confidence values.

If more than two datasets are used, the computations and generation of the association rules are more complicated. However, the MOWCATL algorithm can efficiently address this situation. In drought research, we may need many climatic and oceanic datasets to identify drought and associations of drought with specific values in these datasets. For example, the oceanic parameters represented by the SOI, MEI, NAO, PDO, and several oceanic indices can be built in a database so that the algorithm can generate the associations with the drought indices to monitor drought. After the generation of rules, selection of the most important rules is key to the efficient use of the data-mining algorithms.

d. Selection of rules and interestingness measure

In the data-mining context, selecting the best rule in a certain application depends on the "interestingness measure." This concept of "interestingness" or "goodness" of the rules is defined to compare and select the better rules among the ones that are generated (Bayardo and Agrawal 1999; Silberschatz and Tuzhilin 1995). Although support and confidence values are important in selecting rules, there are several other methods and algorithms to use in quantifying interestingness measures (Padmanabhan and Tuzhilin 1999; Das et al. 1998). Among these methods, Smyth and Goodman's J measure (Smyth and Goodman 1992) is used in this study to quantify the interestingness of the rules in the MOWCATL algorithm.

Smyth and Goodman's J measure is defined as the average information content of a probabilistic classification rule and is used to find the best rules relating discrete-valued attributes. A probabilistic classification rule is a logical implication of "if X then Y" with some probability p (Hilderman and Hamilton 1999). The J measure is given by

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

where $p(x)$, $p(y)$, and $p(y/x)$ are the probabilities of occurrence of x, y, and y given x, respectively, within the dataset. The term inside the largecurved brackets is defined as the similarity (or goodness of fit) of two probability distributions (Hilderman and Hamilton 1999). Klemettine (1999) used the J measure by replacing the probabilities with confidence values of the rules. The J values range between 0 and 1. The higher the J value is the better. However, for infrequent occurrences of episodes such as drought within historical records of a database, the J values are so small so that values greater than 0.04 are considered. This method has been adapted to the algorithms that are used in this study.

In selecting rules, high values for $J(x;y)$ are desirable, but are not necessarily associated with the best rule. For example, rare conditions may be associated with the highest values for $J(x;y)$, but the resulting rule is insufficiently general to provide any new information. Consequently, analysis may be required in which the accuracy of a rule is traded for some level of generality or goodness of fit (Hilderman and Hamilton 1999).

The advantage of this method is that it takes into consideration both frequencies of the left-hand side (X) and right-hand side (Y) of the rules (Smyth and Goodman 1992). It also favors rules that occur more

TABLE 1a. Drought episode classification threshold values used in association and rule generation.

Drought category	Extremely dry (ed)	Severely dry (sd)	Moderately dry (md)	Normal (n)	Moderately wet (mw)	Severely wet (sw)	Extremely wet (ew)
SPI	$<= -2$	$-2 < \times <= -1.5$	$-1.5 < \times <= -1$	$-1 < \times < 1$	$1 =< \times < 1.5$	$1.5 =< \times < 2$	$>= 2$
PDSI	$<= -4$	$-4 < \times <= -3$	$-3 < \times <= -2$	$-2 < \times < 2$	$2 =< \times < 3$	$3 =< \times < 4$	$>= 4$

frequently and provides a more complex metric for ranking in such a way that user can trade off rule support and rule confidence (Harms et al. 2002). The definition of J measure works for a single or multiple antecedents. When the window defines the X that includes the multiple incidents, the probability of occurrences of X is the probability of occurrences of the multiple antecedents defined in that episode defined in the window as X.

4. Discovering association rules to monitor drought in Nebraska

The MOWCATL algorithm can be used to find the relationships between oceanic indices and drought episodes. To demonstrate the use of this algorithm, five stations were selected in Nebraska. These stations were Alliance in Box Butte County (northwest), Ainsworth in Brown County (north-central), Hayes Center in Hayes County (southwest), Clay Center in Clay County (southeast), and West Point in Cuming County (northeast). The selected stations represent different geographical locations in Nebraska. In addition to these stations, the statewide average data were used to compare results between the stations and the state.

The drought episodes were identified using two climatic drought indices, the SPI and PDSI. One of the advantages of using the SPI is that it designed to quantify the precipitation deficit for multiple time scales. Thus, we have used four time scales of the SPI data that include monthly values of a 3-, 6-, 9-, and a 12-month SPI (McKee et al. 1995). The time scale values were considered separately and independently in the data metrics to define drought episodes. These time scales reflect the impact of drought on the availability of the different water resources (Hayes et al. 1999). Although it does not have the same flexibility for multiple time scales, the PDSI is one of the most popular drought indices used in the United States. It quantifies the precipitation deficit using a water balance method of the soil that depicts cumulative effects of drought that are caused as the result of available soil moisture, temperature, and precipitation conditions (Palmer 1965). Thus, we have used these two indices to identify drought for the study period, 1950–99.

As the first step, these two drought indices were categorized into seven categories. These categories were the following: extremely dry, severely dry, moderately dry, normal, moderately wet, severely wet, and extremely wet (Table 1a). For this classification purpose, the seven SPI drought categories and thresholds were adopted from that of used by the National Drought Mitigation Center (Hayes et al. 1999). However, in the PDSI classification, we aggregated the values of Palmer's "incipient" and "mild dry" categories within the normal category to make it consistent with seven classification categories. Then, the drought episodes (i.e., moderately dry, severely dry, and extremely dry categories) were specified as target episodes for generated rules.

The other oceanic and atmospheric indices (i.e., the SOI, MEI, PDO, NAO, and PNA) were also clustered into seven categories based on their historical data frequency distribution using the thresholds shown in Table 1b. These thresholds were determined by assuming a normal frequency distribution over the 50 yr of data, and each oceanic and atmospheric parameter value was divided into 0.5, 1, and 1.5 standard deviations.

In generating rules, the oceanic and atmospheric indices were considered to be the antecedent events while the target drought episodes were consequents. Tables 1a and 1b show the thresholds and keys of the indices that are used in generated rules.

For each station and the state-average data, the monthly PDSI and SPI drought indices were superimposed on the monthly SOI, MEI, PDO, and PNA indices to compare their variations. Figure 2 illustrates the temporal variations of the monthly PDSI and MEI indices from 1950 to 1999 for the state-averaged data. This graph also shows one example of the real-world problem such as precipitation deficit with irregular patterns of the time series data.

The data-mining algorithm (MOWCATL) was executed to find the associations of these temporal patterns to one another with a given confidence level. A variety of window widths, time lag values, and minimum support constraints were used. Because we wanted to take the time order (i.e., time of occurrences of the parameters) into account in our analyses, serial episodes were used in this study. Both confidence and J-measure values were used to indicate the goodness of the rule.

a. Discovering association rules using the MOWCATL algorithm for selected stations

Using the MOWCATL maximum and fixed lag data-mining algorithms, selected association rules that were generated for the selected five stations and the state average of Nebraska are shown in Tables 2 and 3. It can be shown that for serial episodes of MOWCATL

TABLE 1b. Oceanic and climatic indices classification threshold values used in association and rule generation.

Indices category	1	2	3	4	5	6	7
SOI	$>=1.5$	$1=<\times<1.5$	$0.5=<\times<1$	$-0.5<\times<0.5$	$-1<\times<=-0.5$	$-1.5<\times<=-1$	$<=-1.5$
MEI	$<=-1.5$	$-1.5<\times<=-1$	$-1<\times<=-0.5$	$-0.5<\times<0.5$	$0.5=<\times<1$	$1=<\times<1.5$	$>=1.5$
NAO	$<=-4$	$-4<\times<=-3$	$-3<\times<=-2$	$-2<\times<2$	$2=<\times<3$	$3=<\times<4$	$>=4$
PDO and PNA	$<=-2$	$-2<\times<=-1.5$	$-1.5<\times<=-1$	$-1<\times<1$	$1=<\times<1.5$	$1.5=<\times<2$	$>=2$

with an antecedent and consequent window size of two months, and a time lag of three months, the most repeated rule for Ainsworth, Clay Center, Hayes Center, West Point, and the state of Nebraska was if MEI was less than −1.5 and PDO was less than −2, then drought episodes occurred (Table 2). When the MEI is negative (or the SOI is positive), a La Niña event is taking place in the Pacific Ocean. It can be concluded that this condition may be considered as a precursor to drought.

Other important precursors to drought can be seen in the occurrence of negative values of the MEI and the PNA. This rule was generated for all selected stations. However, the values (or categories) of the MEI and the PNA were different. For example, Clay Center and Alliance were in drought when the MEI was less than −1.5 and the PNA was between −2 and −1.5. Ainsworth, Hayes Center, West Point, and the state of Nebraska were in drought with similar conditions with different categories of MEI and PNA but still negative values of both indices (Table 2). This implies that when the Pacific Ocean is colder (La Niña) followed by the negative values of the 500-hPa geopotential height in the Northern Hemisphere (PNA), there is a higher probability of occurrence of drought in Nebraska. It can be concluded that the rules generated (Table 2) show that occurrences of the MEI less than −0.5 followed by the PNA less than −1 with a 3-month window size implied drought particularly at longer time scales (i.e., a 6–12-month period) in all five selected stations, as well as for the state of Nebraska.

The rules that are generated using the MOWCATL algorithm with a fixed time lag are shown in Table 3 for serial episodes. The advantage of MOWCATL with the fixed time lag is that it provides the rules that can occur after a specified fixed number of months, whereas the MOWCATL algorithm with a maximum time lag provides the rules that occur within the specified time, which include the rules that are generated using a fixed time lag. Thus, the rules generated for the fixed time lag are subsets of the maximum time lag outputs.

The rules generated using the MOWCATL algorithm for serial episodes indicate that there are strong relationships between drought episodes in Nebraska and the MEI, SOI, NAO, and PDO with different combinations of the indices and confidence values for each station selected as well as for the state-averaged Nebraska data.

b. Assessment of the rules with past and recent drought years

Using the association rules based on 1950–99 historical data, the past (1950–99) and recent (2000–03) years were considered to assess and validate the relationships between drought and the oceanic parameters.

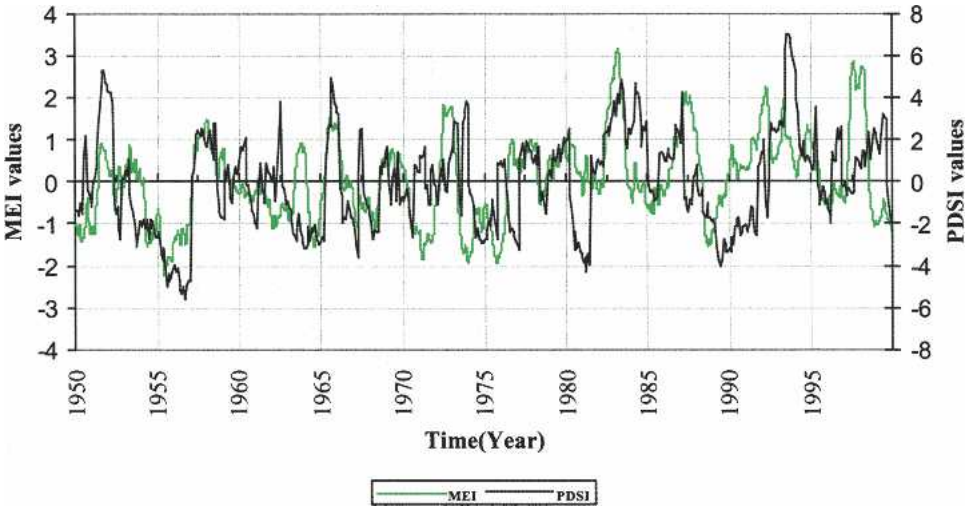

FIG. 2. Temporal patterns of MEI and PDSI values for state-averaged data of Nebraska from 1950 to 1999.

TABLE 2. Frequent minimal occurrence episodes generated with constraints and maximum time lags: serial episodes, antecedent window = 2 months, consequent window = 2 months, and the maximum time lag between the start of the antecedent and the start of the consequent = 3 months.

Location	Selected serial rules	Confidence	J measure	Location	Selected serial rules	Confidence	J measure
Clay Center	MEI1, PDO1 ⇒ PDSIed SP12sd	0.88	0.08	Hayes Center	MEI1, PDO1 ⇒ PDSIed	0.88	0.09
					MEI1, PDO1 ⇒ SP12md	0.75	0.08
	MEI3, PNA2 ⇒ SPI9sd, PDSIed	0.75	0.04		MEI3, PNA3 ⇒ SPI9md	0.75	0.08
	PDO3, SOI3 ⇒ SPI9md	0.71	0.07	West Point	MEI1, PDO1 ⇒ PDSIed	0.88	0.09
Ainsworth	MEI1, PDO1 ⇒ PDSIed	0.88	0.09		MEI3, PNA3 ⇒ PDSImd	0.75	0.08
	MEI1, PDO3 ⇒ SPI6sd	0.86	0.08		SOI2, NAO3 ⇒ SPI3md	0.71	0.07
	NAO2, MEI3 ⇒ SPI3md	0.80	0.06	Nebraska (state avg)	MEI1, PDO1 ⇒ PDSIed	0.88	0.09
	MEI1, PNA3 ⇒ SP12sd, PDSIed	0.75	0.04				
Alliance	PDO3, PDO3 ⇒ SPI3md	0.86	0.08		MEI3, PNA3 ⇒ PDSImd	0.88	0.09
	MEI3, PNA2 ⇒ SPI9md	0.75	0.05		MEI1, PDO1⇒ SP12sd	0.75	0.08

The 1950s droughts were typical and in agreement with the generated rules for all selected stations. The droughts in the mid-1960s, early and mid-1970s, and late 1980s were also associated with the rules that were generated with the MOWCATL algorithm. For example, severe droughts in 1954, 1955, 1956, 1965, 1976, 1983, 1988, and 1989 in Alliance, Ainsworth, Hayes Center, Clay Center, and Nebraska state-average data confirm these associations.

Considering the recent drought years, in 2000, the MEI, PDO, and NAO were dominantly negative throughout the year. These negative values were the continuation of the La Niña condition in 1999. This situation corresponded with the drier-than-normal condition in four of the five selected stations. Only Alliance had normal precipitation in 2000. In 2001, the values of MEI, PNA, and NAO were in normal category (Table 1b), while the PDO was negative in the second half of the year. The drought indices also showed a near-normal condition in 2001.

In the year 2002, four out of five selected stations recorded drier-than-normal conditions. Only West Point was near normal. Unlike the other drought years, the drought in 2002 was not related to the negative values of MEI, PDO, PNA, and NAO. This implies that not all drought periods are necessarily associated with La Niña. This confirms that drought is a complex phenomenon that no single method can solve.

Thus, since the technique that is developed does not apply in all cases, it should be used as a complement to other existing techniques for better drought monitoring. In the future, the inclusion of other local and ecological parameters such as the available land-cover type and soil moisture may improve the quality of the generated rules in identifying the relationships with drought that may be used in drought monitoring.

5. Summary and conclusions

Discovering association rules in sequences is useful in many scientific and commercial domains (Hipp et al. 2000). Identifying sequential rules that are inherent in the data helps drought researchers to learn from past data and make informed decisions about the future. Real-life applications include identifying patterns and finding relationships between global oceanic events (e.g., El Niño) and local weather events, such as precipitation.

The rules generated using the MOWCATL algorithm indicate that there are relatively strong associations between the oceanic indices and precipitation deficits in Nebraska. In most cases, MEI, PNA, SOI, and PDO occurred as an antecedent combination to consequent drought episodes, with different combinations of the indices and confidence values for the selected stations as well as for the state-average data.

TABLE 3. Frequent minimal occurrence episodes generated with constraints and fixed time lags: serial episodes, antecedent window = 2 months, consequent window = 2 months, and a fixed time lag between the start of the antecedent and the start of the consequent = 3 months.

Location	Selected serial rules	Confidence	J measure	Location	Selected serial rules	Confidence	J measure
Clay Center	MEI1, PDO1 ⇒ PDSIed, SP12sd	0.88	0.08	Hayes Center	MEI1, PDO1 ⇒ PDSIed	0.88	0.09
Ainsworth	MEI1, PDO3 ⇒ SPI6sd	0.57	0.05		MEI1, PDO1 ⇒ PDSIed, SP12md	0.75	0.06
	MEI1, PDO1 ⇒ PDSIed	0.50	0.05	West Point	MEI1, PDO1 ⇒ SP12sd	0.75	0.08
Alliance	PDO1, PDO3 ⇒ SPI3md	0.71	0.07	Nebraska (state avg)	MEI1, PDO1 ⇒ SP12sd	0.75	0.08

Generally, most rules in this study indicate that the oceanic and atmospheric parameters are precursors to long-term drought defined by the PDSI and the SPI. The rules also indicated that the 9- and the 12-month SPI are better associated with the SOI, MEI, PNA, and PDO than the 3- and the 6-month SPI values. The experimental results show that the more associated values among different time periods of the SPI were the 9-month SPI values. This may be comparable with PDSI drought values, which are based mainly on the water balance concept (Palmer 1965). This indicates that the oceanic parameters are more associated with long-term (more than 6 months) drought than with short-term (less than 3 months) drought values. The result can be justified by the fact that the oceanic parameters are relatively more stable and can be used to indicate longer-term drought conditions. This study has also identified three advantages of data mining as compared to the previous traditional methods.

1) Instead of global statistical correlation of the climatic and oceanic data, target episodes such as droughts can be specified separately from normal and wet conditions as an alternative discovery method of the relationships.
2) Data-mining algorithms give flexibility in time series analyses, allowing the discovery of relationships between the parameters with time lags with a defined window size. Because of this flexibility in time, the MOWCATL data-mining algorithm identifies a better association of the oceanic and climatic parameters. This information may be used for drought early warnings.
3) The algorithms allow the analysis of large amounts of data and complicated computations to be executed within a reasonable period of time.

Acknowledgments. This research was supported in part by the National Science Foundation Digital Government Grant EIA-0091530.

REFERENCES

Bayardo, R. J., and R. Agrawal, 1999: Mining the most interesting rules. *Proc. Fifth ACM Int. Conf. on Knowledge Discovery and Data Mining*, San Diego, CA, Association for Computing Machinery, 145–154.

Berry, J. A., and G. Linoff, 2000: *Mastering Data Mining: The Art and Science of Customer Relationship Management*. John Wiley & Sons, 494 pp.

Bond, N. A., and D. E. Harrison, 2000: The Pacific Decadal Oscillation, air–sea interaction and central north Pacific winter atmospheric regimes. *Geophys. Res. Lett.,* **27,** 731–734.

Cabena, P. H., R. Stadler, J. Verhees, and A. Zanasi, 1998: *Discovering Data Mining: From Concept to Implementation*. Prentice Hall, 195 pp.

Covey, C., and Coauthors, 2000: The seasonal cycle in coupled ocean–atmosphere general circulation models. *Climate Dyn.,* **16,** 775–787.

Das, G., K. I. Lin, and H. Mannila, 1998: Rule discovery from time-series. *Proc. Fourth Int. Conf. on Knowledge Discovery and Data Mining,* New York, NY, American Association for Artificial Intelligence, 16–22.

Dewey, K., 1996: Summer: The season of the sun. *NEBRASKAland Magazine's Wea. Climate Nebraska,* **74,** 57–62.

Eden, C., R. J. Greatbatch, and J. Lu, 2002: Prospects for decadal prediction of the North Atlantic Oscillation (NAO). *Geophys. Res. Lett.,* **29,** 1466, doi:10.1029/2001GL014069.

Francis, R. C., and S. R. Hare, 1994: Decadal-scale regime shifts in the large marine ecosystems of the Northeast Pacific: A case for historical science. *Fish. Oceanogr.,* **3,** 279–291.

Groth, R., 1998: *Data Mining: A Hands-on Approach for Business Professionals*. Prentice Hall Inc., 264 pp.

Harms, S. K., J. Deogun, and T. Tadesse, 2002: Discovering sequential rules with constraints and time lags in multiple sequences. *Proc. 2002 Int. Symp. on Methodologies for Intelligent Systems,* Lyon, France, ISMIS, 432–441.

Hayes, M., M. Svoboda, D. A. Wilhite, and O. Vanyarkho, 1999: Monitoring the 1996 drought using the Standardized Precipitation Index. *Bull. Amer. Meteor. Soc.,* **80,** 429–438.

Hilderman, R. J., and H. J. Hamilton, 1999: Knowledge discovery and interestingness measures: A survey. Tech. Rep. CS-99-04, Department of Computer Science, University of Regina, 28 pp.

Hipp, J., U. Gunter, and G. Nakhaeizdeh, 2000: Algorithms for association rule mining: A general survey and comparison. *Proc. Int. Conf. in Management of Data, ACM SIGKDD,* Dallas, TX, Association for Computing Machinery, 58–64.

Huang, Y., and P. S. Yu, 1999: Adaptive query processing for time-series data. *Proc. Fifth Int. Conf. on Knowledge Discovery and Data Mining,* San Diego, CA, Association for Computing Machinery, 282–286.

Hurrell, J. W., 1995: Decadal trends in the North Atlantic Oscillation: Regional temperatures and precipitation. *Science,* **269,** 676–679.

Jacob, R., C. Schafer, I. Foster, M. Tobis, and J. Anderson, 2001: Computational design and performance of the fast ocean atmosphere model, version one. *Proceedings, International Conference on Computational Science*, V. N. Alexandrov, J. J. Dongarra, and C. J. K. Tan, Eds., Springer-Verlag, 175–184.

Keogh, E. J., and M. J. Pazzani, 1998: An enhanced representation of time-series which allows fast and accurate classification, clustering and relevance feedback. *Proc. AAAI-98 Workshop on Predicting the Future: AI Approaches to Time-Series Analysis,* Madison, WI, American Association for Artificial Intelligence, 44–51.

Klemettine, M., 1999: A knowledge discovery methodology for telecommunication network alarm databases. Ph.D. thesis, Department of Computer Science, University of Helsinki, Helsinki, Finland, 137 pp.

Lu, J., and R.J. Greatbatch, 2002: The changing relationship between the NAO and northern hemisphere climate variability. *Geophys. Res. Lett.*, **29,** 1148, doi:10.1029/2001GL014052.

Mannila, H., and J. Seppänen, 2001: Recognizing similar situations from event sequences. *Proc. First SIAM Conf. on Data Mining,* Chicago, IL, Society for Industrial and Applied Mathematics, 1–16. [Available online at http://www.siam.org/meetings/sdm01/pdf/sdm01_03.pdf.]

——, and H. Toivonen, 1995: Discovering frequent episodes in sequences. *Proc. First Int. Conf. on Knowledge Discovery and Data Mining,* Montreal, Canada, Association for Computing Machinery, 210–215.

Mantua, N. J., S. R. Hare, Y. Zhang, J. M. Wallace, and R. C. Francis, 1997: A Pacific interdecadal climate oscillation with impacts on salmon production. *Bull. Amer. Meteor. Soc.,* **78,** 1069–1079.

McKee, T. B., N. J. Doesken, and J. Kleist, 1995: Drought monitoring with multiple time scales. Preprints, *Ninth Conf. on Applied Climatology*, Dallas, TX, Amer. Meteor. Soc., 233–236.

Overland, J. E., H. J. Niebauer, J. M. Adams, N. A. Bond, and S.

L. McNutt, cited 2002: Causes of variability in the Aleutian low: A project for the Arctic Research Initiative. [Available online at http://www.pmel.noaa.gov/~miletta/web/page1.html.]

Padmanabhan, B., and A. Tuzhilin, 1999: Unexpectedness as a measure of interestingness in knowledge discovery. *Decis. Support Syst.,* **27,** 303–318.

Palmer, W. C., 1965: Meteorological drought. U.S. Department of Commerce Weather Bureau Research Paper No. 45, 58 pp.

Povinelli, R. J., 2000: Using genetic algorithms to find temporal patterns indicative of time-series events. *Proc. Genetic and Evolutionary Computation Conf. (GECCO) Workshop: Data Mining with Evolutionary Algorithms,* Las Vegas, NV, American Association for Artificial Intelligence, 80–84.

Rind, D., R. Goldberg, J. Hansen, C. Rosenzweig, and R. Ruedy, 1990: Potential evapotranspiration and the likelihood of future drought. *J. Geophys. Res.,* **95,** 9983–10 004.

Silberschatz, A., and A. Tuzhilin, 1995: On subjective measures of interestingness in knowledge discovery. *Proc. First Int. Conf. on Knowledge Discovery and Data Mining,* Montreal, Canada, Association for Computing Machinery, 275–281.

Smyth, P., and R. M. Goodman, 1992: An information theoretic approach to rule induction from databases. *IEEE Trans. Knowledge Data Eng.,* **4,** 301–316.

Svoboda, M., and Coauthors, 2002: The drought monitor. *Bull. Amer. Meteor. Soc.,* **83,** 1181–1190.

USDA RMA, 1999: USDA National Risk Management Agency database. Risk Management Office, Billings, Montana.

Wilhite, D. A., 2000a: Preparing for drought: A methodology. *Drought: A Global Assessment*, D.A. Wilhite, Ed., Vol. 2, Routledge, 89–104.

——, 2000b: Drought as a natural hazard: Concepts and definitions. *Drought: A Global Assessment*, D.A. Wilhite, Ed., Vol. 1, Routledge, 3–18.

Wolter, K., and M. S. Timlin, 1993: Monitoring ENSO in COADS with a seasonally adjusted principal component index. *Proc. Seventh Annual Climate Diagnostic Workshop*, Norman, OK, NOAA, 52–57.