Global NEST Journal, Vol 14, No 3, pp 354-361, 2012

ASSOCIATIONS BETWEEN STREAM FLOW AND CLIMATIC VARIABLES AT KIZILIRMAK RIVER BASIN IN TURKEY

F. DADASER-CELIK[1*]
M. CELIK[2]
A.S. DOKUZ[2]

[1] *Department of Environmental Engineering*
Erciyes University, Kayseri, Turkey
[2] *Department of Computer Engineering*
Erciyes University, Kayseri, Turkey

Received: 05/12/11
Accepted: 27/07/2012
*to whom all correspondence should be addressed:
e-mail: fdadaser@erciyes.edu.tr

ABSTRACT
This study aims to demonstrate the use of association analysis for discovering the relationships between stream flow and climatic variables in the Kızılırmak River Basin in Turkey. Association analysis is a data mining technique that aims to discover rules in the form of A→B that may occur in large datasets with frequency above a given threshold. A and B can be defined as events of a certain type, with the rule if A occurs then B occurs. In this study, A refers to climatic variable(s) (i.e., precipitation, temperature, wind speed, relative humidity) of certain magnitude, and B refers to the magnitude of stream flow. The interesting rules were quantified using support and confidence measures. Stream-flow data from three gauging stations in the Kızılırmak River Basin and climate data from three weather stations in the same basin were included in the analyses. All data were first segregated into three groups that were named as low, medium, and high. Low and high ranges of stream-flow data were further divided into three to increase our focus on extreme events. The analyses were conducted at the annual and seasonal timescales. The analyses indicated that the relationships between precipitation and temperature and stream flow are most prevalent but, relative humidity and wind speed are also important determinants of stream flow in the Kızılırmak River Basin.

KEYWORDS: stream flow, climate, data mining, association analysis, Kızılırmak River Basin.

1. INTRODUCTION

Identification of the relationships between hydrologic and climatic variables is very important for many hydrologic applications, such as prediction of missing records, analysis of climate change impacts, and estimation of hydrologic responses in ungauged basins. Unfortunately, identification of these relationships is not a straight-forward process due to the characteristics of the data (Shekhar *et al*., 2009; Ganguly and Steinhaeuser, 2008) and complexity of hydroclimatic relationships: (1) hydrologic and climatic data are geographical data and have spatial and temporal correlations; (2) hydrologic and climatic data have nonlinear dependences, they have long memory in time and they have long-range or tele-connections in space; (3) the linkages between hydrologic and climatic data are based on complex physical processes that are difficult to conceptualize. Hydrological models have been developed to improve our understanding of hydrologic and climatic linkages, but they need local level information on hydrogeology, soils, topography, land-use, etc. This information is often hard to get and even more difficult to obtain when multiple locations are of interest, e.g., when a regional study is to be conducted; (4) hydrologic and climatic datasets in many areas include gaps and missing records, which poses a major problem in statistical analysis. In this study, the goal is to develop and apply a data mining technique, called association analysis, for discovering the relationships between hydrologic (i.e. stream flow) and climatic variables.

Data mining aims to develop automatic or semi-automatic methods for discovering unforeseen, interesting, and meaningful relationships from heterogeneous and large datasets, which cannot be analyzed manually. With this approach, it is possible to extract cause-effect relationships, determine which variables have the strongest relationships to the problems of interest, and develop models that

predict future outcomes. Data mining can be divided into four major methods or research areas (Tan *et al*., 2005; Han and Kamber, 2001): 1) classification, 2) clustering, 3) anomaly detection, and 4) association analysis. The first two methods can be used for grouping of data into classes/clusters. The third method is used for identification of anomalous data. The fourth method is used for identifying the relationships between various variables. This study aims to develop association analysis techniques to discover the interesting and non-trivial associations of hydrologic and climatic variables.

Other techniques such as regression or correlation analysis can also be used to find the relationships between various variables (e.g., Kletti and Stefan, 1997). Regression analysis aims to find a function to develop a model of the data in an expected error range and deals with numerical (continuous) values. Correlation analysis examines the degree and direction of relationships between two numerical variables. Association analysis is different than other methods in that it examines the relationships between various categorical variables, it can analyse large datasets successfully and can produce rules that show the cause-effect relationships between various combinations of variables at different spatial and temporal scales. It is therefore advantages over other methods when large datasets including several variables are used in the analysis (Changpetch and Lin, 2012).

Association analysis has been used extensively in financial services, banking, advertising, manufacturing, and e-commerce. The use of this method in environmental research, however, is very limited. Kumar *et al*. (2001), Tan *et al*. (2001), and Steinbach *et al*. (2002) used association analysis to find interesting spatio-temporal patterns in earth science data. Tadesse *et al*. (2004) determined association rules between climatic and oceanic variables to analyze drought in Nebraska. Dhanya and Kumar (2009) analyzed rainfall data to discover association rules for droughts and floods in India.

In this study, we aim to determine the associations between stream flow and climatic variables in the Kızılırmak River Basin in Turkey. In section 2, we introduce the study area. In section 3, we explain the methods used in this study. In section 4, we provide the analysis results, and finally in section 5, we provide the conclusions.

2. STUDY AREA

The Kızılırmak River Basin is located in Central Turkey (Figure 1) and covers an area of 78,180 km^2. The Kızılırmak River (in English, *Red River*) originates from Karadağ Spring near the city of Sivas in Cental Anatolia, moves toward north, and discharges into Black Sea near the city of Samsun. The climate in the basin changes from continental in the southern basin to Mediterranean in the northern part. Annual average temperature in the basin change from 7 to 15 ^{o}C from south to north. Annual precipitation is 300 to 900 mm and annual evaporation is 435 to 1500 mm in the south-north direction (Tubitak, 2010). Most of the basin is covered with agricultural areas and forests or semi-natural areas. Agricultural areas cover about 55% of the basin and forests and semi-natural areas cover another 43% of the basin (Tubitak, 2010).

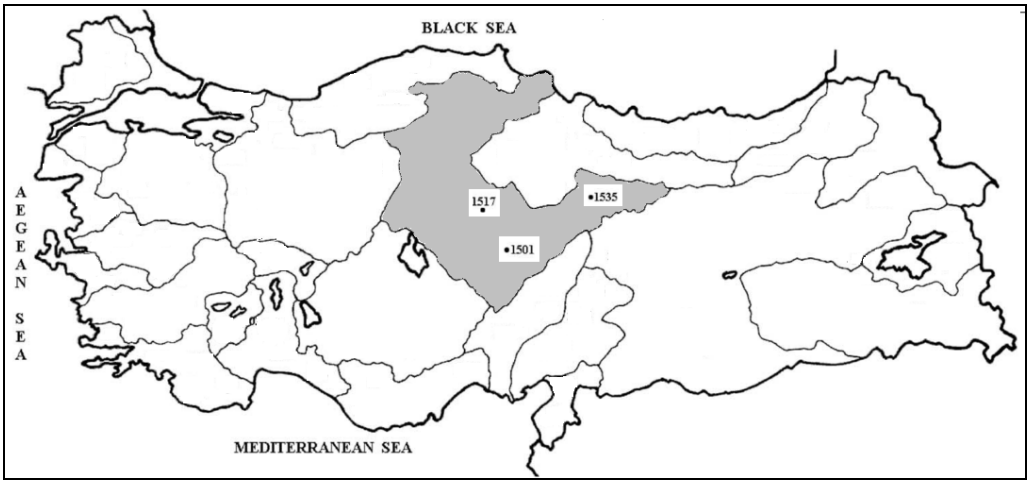

Figure 1. Location of Kızılırmak River Basin in Turkey and the names (given as station number) and locations of gauging stations used in this study

Monthly average stream flow obtained from daily stream-flow data collected at three gauging stations (1501, 1517 and 1535) in the Kızılırmak River Basin were used in this study. The stations 1501 and 1535 are located on the Kızılırmak River and station 1517 is located on a tributary to the Kızılırmak River. All stations are located in regions where continental climate is prevalent. The characteristics of the three gauging stations are given in Table 1. As can be seen, stations 1501 and 1535 show much larger variation in flows compared to the variation observed at station 1517. In addition, monthly average flows at these two stations are higher than the average flow at station 1517.

The climate data used in this study consist of total monthly precipitation, average monthly temperature, average monthly relative humidity, and average monthly wind speed obtained from three weather stations located in the same basin. All data were available from 1975 to 2000.

Table 1. Characteristics of stream flow at gauging stations used in this study. Minimum, maximum, and mean flows were calculated for the 1975-2000 period

Station Number	Drainage Area (km^2)	Station Elevation (m)	Minimum Flow ($m^3 s^{-1}$)	Maximum Flow ($m^3 s^{-1}$)	Average Flow ($m^3 s^{-1}$)	Standart Deviation ($m^3 s^{-1}$)
1501	15,582	995	7	500	73	83
1517	8,592	895	0.05	62	13	12
1535	6,607	1,243	3	344	41	54

3. DATA ANALYSIS AND METHODS

3.1. Data discretization and preparation

In association analysis, the data should be in discrete format. In this study, we discretized the data by using their statistical properties (i.e. mean (μ) and standard deviation (σ)). First, the data were arbitrary divided into three groups: “low (L)” if they were “smaller than μ-0.5σ”, “medium (M)” if they were between “μ-0.5σ” and “μ+0.5σ”, and “high (H)” if they were “higher than “μ+0.5σ”.

Low and high ranges of data were further divided into three subgroups each to emphasize more on extreme events (Figure 2). If the data were between “μ+0.5σ” and “μ+0.75σ”, they were named as “moderately high (MH)”; if the data were between “μ+0.75σ” and “μ+1σ” the data were named as “severely high (SH)”; and if the data were higher than “μ+1σ”, the data were named “extremely high (EH)”. In a similar way, the low ranges of data were divided into three subgroups as “moderately low (ML)”, “severely low (SL)”, and “extremely low (EL)”.

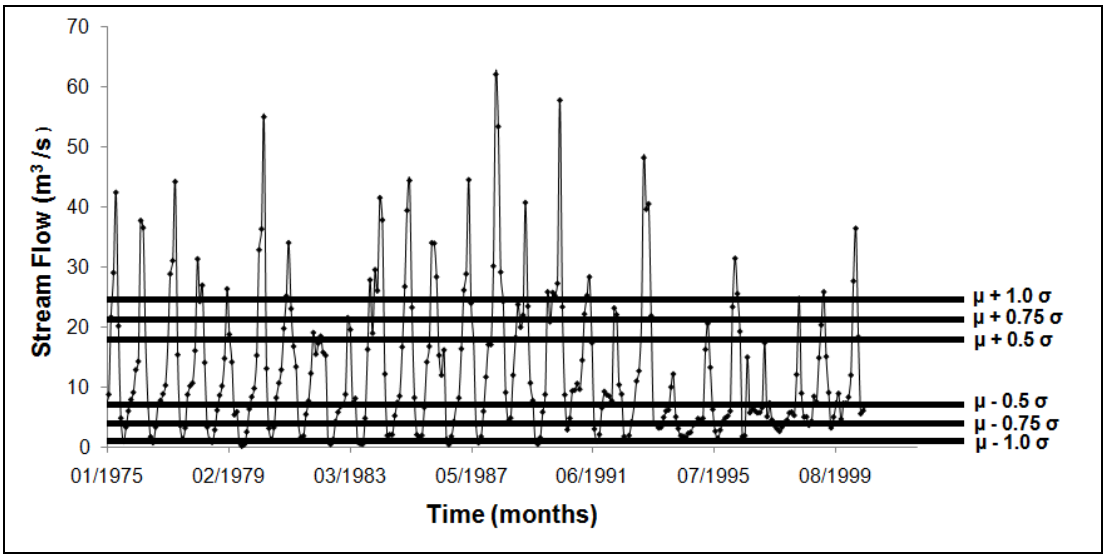

Figure 2. Average monthly stream-flow data for station 1517 and the threshold values used for data discretization

The analysis was conducted at the annual and seasonal timescales. Annual analysis included data from all months in a year (from January to December), while in seasonal analysis, the data were

divided into the summer (from April to September) and winter (from October to March) periods. In general, temperatures are low and precipitation is in the form of snow from October to March. In contrast temperatures are high and precipitation is in the form of rain from April to September. In addition, the analysis was conducted separately on the datasets with low (ML, SL, EL) and high (MH, SH, EH) stream-flow values.

3.2. Association analysis

In the association rule mining, the aim is to discover any rules of the form A→B that seem to occur in the data with frequency above a given threshold. Here A and B are events of a certain type, connected by the rule 'if A occurs then B occurs. The rules can be extended into the form A_1, A_2, …, A_H→B, which can be interpreted as 'if A_1, A_2, …, A_H all occur, then B will occur'.

Interest measures are used to identify and characterize interesting associations. Support and confidence are two measures that are widely used (Han and Kamber, 2001). Support defines the probability of occurrence of the set of {A, B} in an itemset (Eq.1). Confidence shows how often itemsets containing A also contain B (Eq.2).

Support ({A, B}) = (records containing A and B) / (all records) (1)

Confidence (A→B)= support (A∪B) / support (A) (2)

A basic example for association analysis can be formed using precipitation and stream flow. Let A be "low precipitation" and B be "high stream flow". If in a hydroclimatic database includes 10 records and {A, B} pattern is present in 6 of them, then, the support of the pattern {A, B} will be 0.6. If the user-defined threshold for interesting patterns is 0.5, then {A, B} will be an interesting pattern as its support value is higher than the threshold. Confidence measures whether A follows B, or vice versa. In this case, if the support of pattern {A} is 0.8, then confidence for A→B pattern will be 0.75. If the user-defined confidence threshold is 0.5, then the rule A→B will be meaningful. An association between A and B will be read as 'if low precipitation occurs, then high stream flow occurs at 75% confidence'.

A commonly used association discovery algorithm is the Apriori algorithm (Agrawal and Srikant, 1994) which aims to discover frequent itemsets (patterns) that satisfy the user-defined support threshold. The basic principle of the Apriori algorithm can be summarized as "if an itemset is frequent, then all of its subsets must also be frequent". Based on this principle, Apriori algorithm discovers size k+1 candidate itemsets using size k frequent itemsets. Then, the algorithm checks if the candidate size k+1 itemsets satisfy the support threshold. In Apriori algorithm, size k subsets are extended by one item in each iteration (a step known as candidate generation), and candidate size k+1 itemsets are tested against the data. If it is true, the size k+1 candidate itemset is called frequent itemset and the algorithm continues to find size k+2 candidate (superset) itemsets. This iteration (process) continues until finding all possible superset itemsets. The algorithm outputs all size frequent patterns, and finally the directions of the rules are determined by evaluating the confidence measures.

4. RESULTS

4.1. Association rules between stream flow and climatic data at the annual timescale

Apriori algorithm was applied to find association rules between stream flow and climatic data for the 1975-2000 period. The analysis was conducted by using stream flow and climate data from all months. A support threshold of 0.2 and a confidence threshold of 0.5 were used for extraction of frequent rules. The targets were low (L), medium (M), and high (H) stream flow. Even though all rules were generated, the longest rules, which are the rules with the largest number of variables, were reported. It should be noted that that all sub-rules of the longest rule also satisfy the support and confidence thresholds.

The generated rules (see Table 2) clearly show that there is an association between stream flow and precipitation, and stream flow and temperature, particularly for low flow values. Stream flow appears low, when precipitation is low and temperature is high in all three stations. Confidence is equal to or higher than 0.8 (80%) in all stations for this rule. The associations of stream flow with relative humidity and wind speed are also identified. Stream flow is low at all stations when relative humidity is low. Medium stream flow is associated with low temperatures and low stream flow is associated with medium wind speed at station 1501.

Table 2. Selected association rules at the annual timescale (from January to December)

Determinant (if ---- occur)	Result (then ---- occur)	Support	Confidence
Station No: 1501			
Precipitation-L; Rhumidity-L;	Stream flow-L	0.20	0.86
Precipitation-L; Temperature-H	Stream flow-L	0.20	0.84
Temperature-H; Rhumidity-L	Stream flow-L	0.21	0.73
Temperature-L	Stream flow-M	0.21	0.58
WSpeed-M	Stream flow-L	0.26	0.52
Station No: 1517			
Precipitation-L; Temperature-H	Stream flow-L	0.20	0.84
Rhumidity-L	Stream flow-L	0.23	0.73
Station No:1535			
Precipitation-L; Temperature-H	Stream flow-L	0.20	0.86
Rhumidity-L	Stream flow-L	0.21	0.59

4.2. Association rules between hydrologic and climatic variables at the seasonal timescale

We extracted association rules between stream flow and climatic data for two seasons, winter (from October to March) and summer (from April to September). In this experiment, the aim was to determine if the rules change for different seasons. As in the previous case, a support threshold of 0.2 and a confidence threshold of 0.5 were used and only the longest rules were reported.

Table 3 shows that temperature is the major determinant of stream flow during the winter season. Stream flow is medium, when temperature is low in stations 1501 and 1535. In station 1501, low temperature and high relative humidity conditions are associated with medium stream flow. The support values of the rules generated using the dataset for winter season are higher than that of the rules generated using the entire dataset. No frequent rules were generated for station 1517. In other words, the rules generated for station 1517 did not satisfy the support and confidence thresholds. This can be due to the fact that no climatic variable strongly affects stream flow in this station during the winter season.

Table 3. Selected association rules for winter season (October to March)

Determinant (if ---- occur)	Result (then ---- occur)	Support	Confidence
Station No: 1501			
Temperature-L; Rhumidity-H	Stream flow-M	0.35	0.61
Temperature-L	Stream flow-M	0.43	0.59
Precipitation-M; Temperature-L	Stream flow-M	0.20	0.53
WSpeed-M	Stream flow-L	0.25	0.53
Station No:1535			
WSpeed-L; RHumidity-H	Stream flow-L	0.20	0.56
WSpeed-L	Stream flow-L	0.31	0.56
Temperature-L	Stream flow-M	0.36	0.51

Table 4. Selected association rules for summer season (April to September)

Determinant (if ---- occur)	Result (then ---- occur)	Support	Confidence
Station No: 1501			
Precipitation-L; Temperature-H; Wspeed-M; RHumidity-L	Stream flow-L	0.20	0.83
Station No: 1517			
Precipitation-L; Temperature-H; Wspeed-M; RHumidity-L	Stream flow-L	0.20	0.83
Station No:1535			
Precipitation-L; Temperature-H; Wspeed-L; RHumidity-L	Stream flow-L	0.30	0.56
Precipitation-H	Stream flow-H	0.22	0.92

The generated rules for the summer season show associations between all climatic variables and stream flow (see Table 4). Stream flow is low, when precipitation is low, temperature is high, wind speed is medium, and relative humidity is low in stations 1501 and 1517. The same rule is valid for station 1537, except for low stream flow which occurs when wind speed is high. At station 1535, high stream flow in summer season is associated with high precipitation.

4.3. Association rules between extreme hydrologic events and climatic variables

As a final experiment, we aimed to find the association rules for extreme events. In this case, the low and high ranges of data were divided into three subgroups (see Figure 2). For the low flow events, the targets were moderate (ML), severe (SL), and extreme (EL) low flow. For the high flow events, the targets were set as moderate (MH), severe (SH), and extreme (EH) high flow. A support threshold of 0.2 and a confidence threshold of 0.5 were used for extraction of frequent rules and only the longest rules were reported.

Table 5. Selected association rules for low flows

Determinant (if ---- occur)	Result (then ---- occur)	Support	Confidence
Station No: 1501			
Rhumidity-EL	Stream flow-ML	0.24	1.00
Rhumidity-M	Stream flow-ML	0.20	0.97
Precipitation-M	Stream flow-ML	0.32	0.98
Temperature-M	Stream flow-ML	0.22	0.97
WSpeed-M; Precipitation-EL	Stream flow-ML	0.20	0.93
Precipitation-EL; Temperature-EH	Stream flow-ML	0.23	0.89
Station No: 1517			
Precipitation-EL	Stream flow-SL	0.25	0.57
Temperature-EH	Stream flow-SL	0.24	0.57
Station No:1535			
Precipitation-M	Stream flow-ML	0.36	1.00
Precipitation-EL; Temperature-EH	Stream flow-ML	0.26	1.00
Rhumidity-EL; Precipitation-EL	Stream flow-ML	0.21	1.00
Temperature-EL	Stream flow-ML	0.20	1.00
Rhumidity-M	Stream flow-ML	0.23	1.00
Rhumidity-EH	Stream flow-ML	0.20	1.00
WSpeed-EL	Stream flow-ML	0.20	1.00
WSpeed-EH	Stream flow-ML	0.21	1.00

Table 6. Selected association rules for high flows

Determinant (if ---- occur)	Result (then ---- occur)	Support	Confidence
Station No: 1501			
Precipitation-EH; Temperature-M	Stream flow-EH	0.24	0.88
Rhumidity-M; Temperature-M	Stream flow-EH	0.41	0.84
Rhumidity-M; Precipitation-EH	Stream flow-EH	0.21	0.76
Rhumidity-M; Precipitation-M	Stream flow-EH	0.20	0.67
WSpeed-M	Stream flow-EH	0.25	0.62
Station No: 1517			
Rhumidity-M; Temperature-M	Stream flow-EH	0.26	0.80
Precipitation-EH	Stream flow-EH	0.22	0.74
WSpeed-M	Stream flow-EH	0.26	0.69
Precipitation-M	Stream flow-EH	0.28	0.60
Station No:1535			
Rhumidity-M; Precipitation-EH	Stream flow-EH	0.22	0.94
Precipitation-EH; Temperature-M	Stream flow-EH	0.26	0.90
Rhumidity-M; Temperature-M	Stream flow-EH	0.28	0.86
WSpeed-M	Stream flow-EH	0.21	0.56

Table 5 lists the selected association rules for low flows. The algorithm generated rules only for moderate low stream flow because moderate low flows were more frequent in the dataset. For all stations, stream flow is associated with precipitation and temperature. In stations 1501 and 1535, stream flow is moderately low, when precipitation is extremely low and temperature is extremely high. In station 1517, stream flow is severely low, when precipitation is extremely low and temperature is extremely high. The associations with wind speed and relative humidity were not consistent between the three stations.

Table 6 presents the association rules generated for high flows. Most of the rules generated for stations 1501 and 1535 are common. Stream flow is extremely high when precipitation is extremely high and temperature is medium. Another association for these two stations is that if relative humidity is medium and precipitation is extremely high, then extremely high stream flow occurs. For station 1517, extremely high levels of stream flow is associated strongly with medium temperature and medium relative humidity.

5. CONCLUSIONS

In this study, we used a data mining technique, called association analysis, for discovering the relationships between stream flow and climatic variables in the Kızılırmak River Basin in Turkey. We used a 25 year period of stream-flow data from three gauging stations and precipitation, temperature, wind speed, and relative humidity data from three weather stations.

The analysis showed that the relationships between stream flow and precipitation, and stream flow and temperature are the strongest. As can be expected, low stream flow was found to be associated with low precipitation and high temperature. In contrast, high stream flow was found to be associated with high precipitation and medium temperature. It is interesting to note that some associations were detected between relative humidity and wind speed and extreme stream-flow events. In general, the rules generated for the stations 1501 and 1535 were similar, which can be due to the fact that watershed characteristics (e.g., geology, slope, land use) of these two stations are similar.

This study showed that association analysis can successfully be applied for identifying the interesting and non-trivial relationships between hydrologic and climatic variables. Once defined, the rules generated between hydrologic and climatic variables using association analysis can be used for filling missing hydrologic records based on known climatic conditions, for predicting impacts of climatic changes on hydrologic systems, and estimating hydrologic responses in ungauged basins with certain confidence.

6. ACKNOWLEDGMENTS

This study was partially supported by the Scientific and Technological Research Council of Turkey (TUBITAK), Project Number: CAYDAG 110Y110 and the Research Fund of the Erciyes University, Project Number: FBA-09-866. We would like to thank Eda Cengiz for her contribution at the data preparation step. We would also like to thank anonymous reviewers for their constructive comments.

REFERENCES

1. Agrawal R. and Srikant R. (1994), Fast algorithms for mining association rules, Proceedings of the 20th Int'l Conf. on Very Large Data Bases (VLDB), 12-15 September 2004, Santiago, Chile.
2. Changpetch P. and Lin D.K.J. (2012), Model selection for logistic regression via association rules analysis, *Journal of Statistical Computation and Simulation*, **82**, 1-14
3. Dhanya C.T. and Kumar D.N. (2009), Data mining for evolution of association rules for droughts and floods in India using climate inputs, *Journal of Geopyhsical Research*, **114**, D02102.
4. Ganguly A.R. and Steinhaeuser K. (2008), Data mining for climate change and impacts. Proceedings of the International Workshop on Spatial and Spatiotemporal Data Mining, IEEE International Conference on Data Mining, 15 December 2008, Pisa, Italy.
5. Han J. and Kamber M. (2001), Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, San Francisco, CA, USA.
6. Kletti L.L. and Stefan H.G. (1997), Correlations of climate and streamflow in three Minnesota streams, *Climatic Change*, **37**, 575-600.
7. Kumar V., Steinbach M., Tan P., Klooster S., Potter C. and Torregrosa A. (2001), Mining scientific data: Discovery of patterns in the global climate system, 2001 Joint Statistical Meeting, 5-9 August 2001, Atlanta, Georgia, USA.

8. Shekhar S., Vatsavai R.R. and Celik M. (2009), Spatial and spatiotemporal data mining: Recent advances. In: Next Generation of Data Mining, Kargupta, H., Han, J., Yu, P.S., Motwani, R., and Kumar V. (Eds.), CRC Press, New York, USA:
9. Steinbach M., Tan P., Kumar V., Klooster S., Potter C. (2002), Data mining for the discovery of ocean climate indices, Proceedings of the Fifth Workshop on Scientific Data Mining, 2nd SIAM International Conference on Data Mining, 11-13 April 2002 Arlington, VA
10. Tadesse T., Wilhite D.A., Hayes M.J., Harms S.K. and Goddard S. (2004), Discovering associations between climatic and oceanic parameters to monitor drought in Nebraska using data-mining techniques, *Journal of Climate*, **18**, 1541-1549.
11. Tan P., Steinbach, M., Kumar, V. (2005). Introduction to Data Mining, First Edition, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
12. Tan P., Steinbach M. Kumar V., Klooster S., Potter C. and Torregrosa A. (2001), Finding spatio-temporal patterns in earth science data. KDD 2001 Workshop on Temporal Data Mining, 26 August 2001, San Francisco, CA, USA.
13. Tubitak (2010). Watershed Protection Action Plan Project for Kızılırmak River Basin - Final Report, The Scientific and Technological Research Council of Turkey – Marmara Research Center, Kocaeli, Turkey (in Turkish)