Multivariate Outlier Detection by Using Two-Dimensional Correlation

Fatih Dikbas (f_dikbas@pau.edu.tr)
Pamukkale University https://orcid.org/0000-0001-5779-2801

Research Article

Keywords: Outlier identification, Two-dimensional correlation, Multivariate analysis, Averages of parts

Posted Date: April 6th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-399196/v1

Multivariate Outlier Detection by Using Two-Dimensional Correlation

Fatih DIKBAS*

*Civil Engineering Department, Pamukkale University, Denizli, Turkey

e-mail: f_dikbas@pau.edu.tr

ORCID: 0000-0001-5779-2801

ABSTRACT

The accuracy of descriptive statistics might be influenced by the existence of outliers in data sets. An observation which might not be considered as an outlier in the univariate case might be a multivariate outlier. Therefore, determination of outliers might make multivariate analysis more robust by providing an opportunity for making required corrections before modelling studies. This paper presents the implementation of the two-dimensional correlation method in the determination of multivariate outliers among the observations of six precipitation stations in Turkey. The two-dimensional correlation method considers the averages of the parts of the whole series instead of the average of the whole series and enables determination of the location of the outlier in the compared series. The obtained results point out that an outlier analysis for hydrologic variables should consider the two-directional behavior and the presented two-dimensional correlation method proves to be a strong alternative to be used in outlier and irregularity detection studies even with a limited number of available data. The 2DCorr software used in the study is freely provided as a supplementary material.

Keywords: Outlier identification, Two-dimensional correlation, Multivariate analysis, Averages of parts.

Declarations:

Funding: Not applicable

Conflicts of interest/Competing interests: Not applicable

Availability of data and material: Data used in the study is commercially available.

Code availability: The 2DCorr software used in the study is freely provided as a supplement.

1. INTRODUCTION

Researchers generally face the problem of outliers when analyzing data. There are several causes of outliers, therefore different definitions were suggested by various authors (Hodge and Austin 2004):

- An observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism (Hawkins 1980);
- An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs (Grubbs 1969);
- An observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data (Barnett et al. 1994).

Outliers arise due to mechanical faults, changes in system behavior, fraudulent behavior, human error, instrument error or simply through natural deviations in populations (Hodge and Austin 2004). Determination of an outlier or a group of outliers in a data set with an appropriate method provides important information about the mechanisms generating the outliers. The detection of outliers might allow determination of system faults and fraud before causing potentially catastrophic results. Outliers can also have negative impacts on the selection of the appropriate model as well as on the estimation of the

associated parameters (Chebana et al. 2012) and might influence the accuracy and reliability of computed statistics. Some favor censoring outliers while others oppose censoring measured values (Kirk and McCuen 2008). For example, if outliers are determined when working with hydrologic data, it would be a good practice to compare the investigated dataset with datasets of neighboring stations to decide whether the outliers are real observations or caused by faulty readings. Checks for spatial outliers use nearby stations to estimate a monthly value for a specific time series in a specific month (Eischeid et al. 1995). If the outliers are determined to be real observations, then, instead of ignoring, they should be included in further evaluations for a more realistic modeling by using a method not adversely influenced by outliers.

There is always a possibility of outlier existence for any dataset showing random fluctuations around the mean, which is also the case for the natural processes. The most important step in outlier detection is the selection of the most suitable method (or methods) for the evaluated data. The outliers are determined from the 'closeness' of vectors using some suitable distance metric. Different approaches work better for different types of data, for different numbers of vectors, for different numbers of attributes, according to the speed required and according to the accuracy required (Hodge and Austin 2004). Extensive overviews of outlier determination methods were provided by Tsay (1988), Hu (1987), Davies & Gather (Davies and Gather 1993), Lanzante (1996), Pegram (1997), Aggarwal (2016), Hodge et al. (2004), Wu et al. (2010) and Zhang et al. (2010). Outlier detection has been used in many studies investigating high-dimensional data, uncertain data, streaming data, network data and time series data (Gupta et al. 2014; Pan et al. 2018). Many proposals for outlier identification (Fried, 2004; Hill et al., 2009; Hill and Minsker, 2010) are available, but as Gupta et al. (2014) point out in their overview of outlier detection in temporal data, outlier detection is very challenging because methods for different data types do not generalize easily (Anderson et al. 2016).

The evaluation of outliers might be categorized into two groups of graphical and statistical tests. However, in graphical tests, the assessment of outliers is subjective because they don't allow hypothesis testing and in cases with multiple outliers the graphical methods become harder to apply (Jeong et al. 2017). Outliers, especially the hydrologic ones, generally have a random behavior. Accordingly, climatic outliers are extreme anomalies of any given climatic variable (Hunt 2007). Therefore, the task of determining outliers and their causes is becoming more and more important with the influence of climate change causing more frequent extreme hydrologic events worldwide.

The Grubbs-Beck test is recommended by the federal guidelines for detection of low outliers in flood flow frequency computation in the United States. Cohn et al. (2013) present a generalization of the Grubbs-Beck test for normal data (Rosner 1983; Spencer and McCuen 1996) that can provide a consistent standard for identifying multiple low outliers (potentially influential low flows). Lamontagne Jonathan et al. (2013) provide a Monte Carlo analysis of the performance of low-outlier tests, including the Bulletin 17B GB test and variations of a multiple Grubbs-Beck test. Spencer and McCuen (1996) point out that the Bulletin 17B outlier detection procedure is not designed to test for multiple outliers, and that many flood records contain multiple outliers. Their test considers the three smallest observations, but many flood records in arid regions of the United States can contain more than three outliers (Cohn et al. 2013; Lamontagne and Stedinger 2016).

The validity of the estimates of basic traditional parametric statistical techniques is based on underlying assumptions that sometimes are not met by real climate data. Two of these assumptions are normality and homogeneity. In particular, contamination from a relatively few 'outlying values' may greatly distort the estimates. Sometimes these common techniques are used in order to identify outliers; ironically they may fail because of the presence of the outliers! (Lanzante 1996).

Statistical texts frequently illustrate that one extreme outlier can greatly influence the correlation between two variables (Legates and McCabe Jr 1999). For example, Legates and Davis (1997) illustrate that correlation based measures are more sensitive to outliers than to observations near the mean. Therefore, the correlation method itself becomes a good candidate for detection of outliers. Aggarwal (2016) discusses the different methods for using linear correlation analysis for outlier detection. In his book, Aggarwal also mentions some of the limitations of the currently used correlation-based outlier detection methods. He

states that, even when the data is weakly correlated on a pairwise basis between different dimensions, it is often the case that subspaces of much lower dimensionality contain most of the variance in the data, because of the cumulative effect of inter-attribute correlations. Another related issue is that the correlations in the data may not be global in nature. It is suggested by a number of recent analytical observations (Aggarwal and Yu 2000) that the subspace correlations are specific to particular localities of the data.

This paper presents the implementation of the two-dimensional correlation method for detecting outliers in time series data (Dikbaş 2017). The main advantage of the two-dimensional correlation method is that the correlations are calculated by using the averages of the parts of the whole series instead of considering the average of the whole series. This approach enables the determination of the location of the outlier in the investigated series and improves efficiency of outlier detection. The method is applied on precipitation data from southwest Turkey and the software written for the implementation of the method is freely provided as a supplementary material.

2. MATERIAL AND METHODS

2.1. Two-dimensional correlation

The developed procedure for the detection of outliers in time series data uses the approach of two-dimensional correlation. First, the details of the two-dimensional correlation method which assesses associations between matrices in two directions by making use of the horizontal and the vertical covariances will be given (Dikbas 2017; Dikbas 2018). The approach is based on the idea that the values of a variable generally have different variances in the horizontal and vertical directions when placed on a matrix.

2.1.1. Two-dimensional variance

Variance is the measure of the spread around the average. The following equations show the calculation of the horizontal and the vertical variance for a variable located on a two-dimensional matrix:

$$Var_h(A) = \frac{\sum_m \sum_n (A_{mn} - \overline{A_m})^2}{m \times n} \tag{8}$$

$$Var_v(A) = \frac{\sum_m \sum_n (A_{mn} - \overline{A_n})^2}{m \times n} \tag{9}$$

In the above equations the averages of the mth row and the nth column of the matrix A are indicated by $\overline{A_m}$ and $\overline{A_n}$ respectively.

When the column averages in the matrices are more scattered around the overall average, the horizontal variance takes high values and when the column averages are closer to the overall average, the horizontal variance decreases in value. The same situation is valid for the row averages when calculating the vertical variance.

2.1.2. Two-dimensional covariance

The linear dependence among variables is calculated by covariance which is a quantitative indicator of the co-variation of the variables. In the calculation of two-dimensional correlation, the horizontal covariance provides a measure of how changes in the column averages of one matrix are associated with changes in the column averages of a second matrix. Correspondingly, the vertical covariance provides a similar information among the rows of the compared matrices. The association between the compared variables is

generally expected to be high when the covariance takes higher values. The horizontal and vertical covariances between scalar matrices A and B are defined by the following equations:

$$Cov_h(A,B) = \frac{\sum_m \sum_n (A_{mn} - \overline{A_m})(B_{mn} - \overline{B_m})}{m \times n} \quad (10)$$

$$Cov_v(A,B) = \frac{\sum_m \sum_n (A_{mn} - \overline{A_n})(B_{mn} - \overline{B_n})}{m \times n} \quad (11)$$

In the above equations, $\overline{B_m}$ and $\overline{B_n}$ are the averages of the mth row and the nth column of the matrix B respectively.

2.1.3. Two-dimensional correlation

Covariance is a scale dependent measure of the joint variability among variables. The covariance is positive when the higher values of a variable correspond with the higher values of the compared variable and the same behavior is observed between the lower values of the variables. Covariance takes negative values when the values of the variables tend to show inverse relationship (i.e., higher values of a variable correspond to the lower values of the compared variable). The value of covariance is influenced with the magnitudes of the compared variables; therefore, its value has to be normalized when there is need for comparison of associations among multiple variable pairs. For example, it would not be reasonable to make comparisons of precipitation series by only using covariance because generally there are significant differences between scales of the observed precipitation series in different basins of varied climate zones. Correlation coefficient, which should be preferred in comparisons of variables, is the scaled and normalized version of covariance and it is dimensionless. Based on the horizontal and the vertical covariance values, the horizontal and vertical correlation coefficients are obtained by using the following equations:

$$r_h = \frac{Cov_h(A,B)}{\sqrt{Var_h(A)\,Var_h(B)}} \quad (8)$$

$$r_v = \frac{Cov_v(A,B)}{\sqrt{Var_v(A) Var_v(B)}} \quad (9)$$

The horizontal and the vertical correlations might also be calculated directly as follows:

$$r_h = \frac{\sum_m \sum_n (A_{mn} - \overline{A_m})(B_{mn} - \overline{B_m})}{\sqrt{[\sum_m \sum_n (A_{mn} - \overline{A_m})^2][\sum_m \sum_n (B_{mn} - \overline{B_m})^2]}} \quad (10)$$

$$r_v = \frac{\sum_m \sum_n (A_{mn} - \overline{A_n})(B_{mn} - \overline{B_n})}{\sqrt{[\sum_m \sum_n (A_{mn} - \overline{A_n})^2][\sum_m \sum_n (B_{mn} - \overline{B_n})^2]}} \quad (11)$$

In the above equations, r_h and r_v are the horizontal and the vertical correlations between the matrices A and B.

2.2. The Study Area and Data.

The two-dimensional correlation coefficient is used for determining the outliers of six precipitation observation stations in the regions of Mugla and Denizli cities located in the southwest of Turkey (Figure 1). The distance of the station 07-016 to the remaining stations is relatively higher. The investigated period of monthly precipitation observations covers the 11 years from 1993 to 2003. A longer or a shorter period

might be selected but the results presented below have shown that the selected period was sufficient for illustrating the success of the developed method and software in the determination of existing outliers. The method is applied for each year separately, therefore the length (or the shortness) of the dataset does not influence the results but each station has to have the same observation period. The heatmaps of the precipitation observations and their averages are presented in Figure 2.

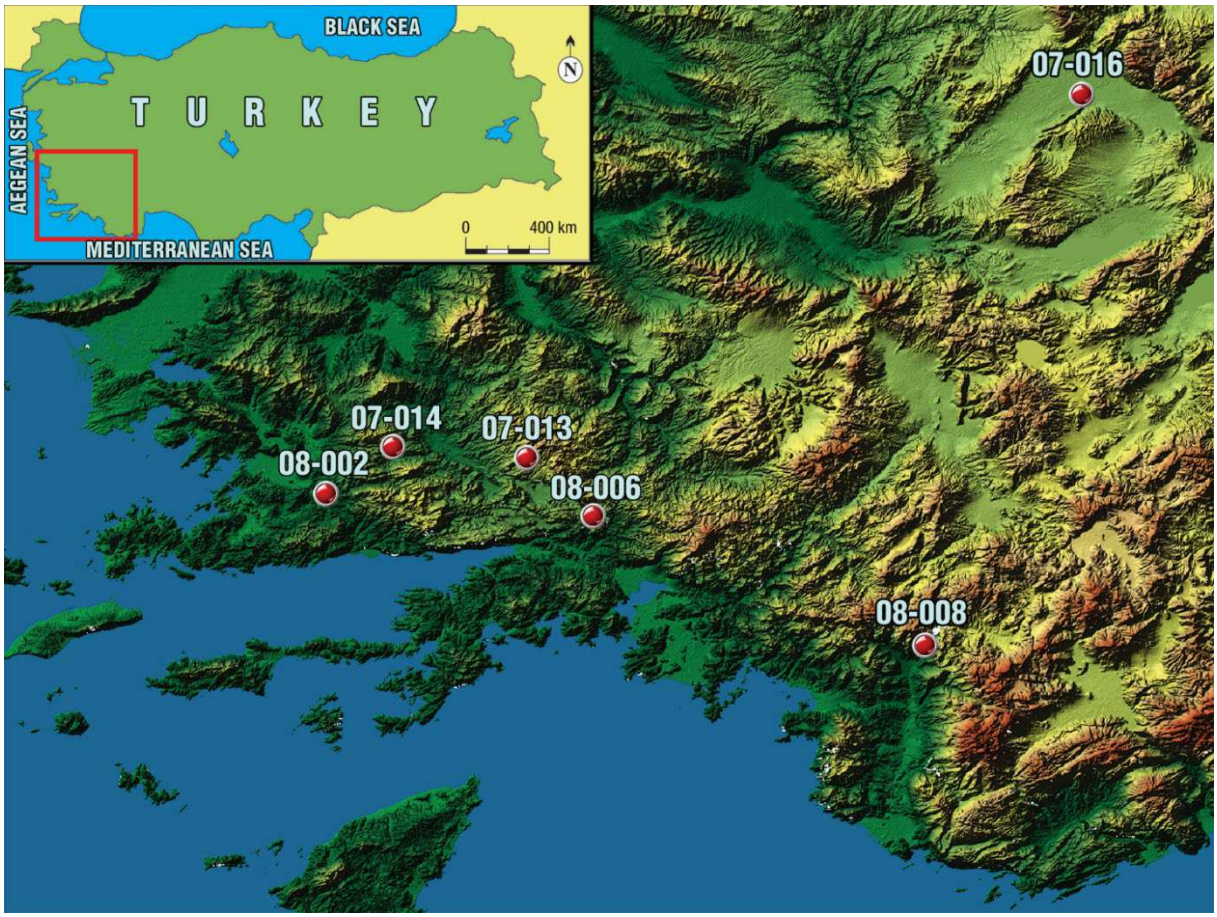

Fig. 1 The precipitation observation stations selected for implementation of the outlier detection method

The correlations between the monthly total precipitation observations of the selected stations covering the 11 years from 1993 to 2003 are calculated by using both the traditional and the two-dimensional correlation approaches. Figure 2 shows the heatmaps of the monthly precipitation observations of the investigated stations together with the row and column averages. The average of the whole data matrix is shown in the bottom right cell. In the investigated area, precipitation is lower in summer months and higher in winter months. The figures show that while the interannual variation is high (as seen in rows), year-wise variation is more limited (as seen in columns).

Table 1 shows the descriptive statistics and the percentiles for all stations. The maximum values are shown in bold. The observations of station 07-016 fit to the Wakeby distribution and have a lower average and variance than the remaining stations which fit to the Johnson SB (J. SB) distribution. The highest precipitation was received by the station 08-006. No precipitation was observed in at least 10% of the investigated period in all stations except the station 07-016.

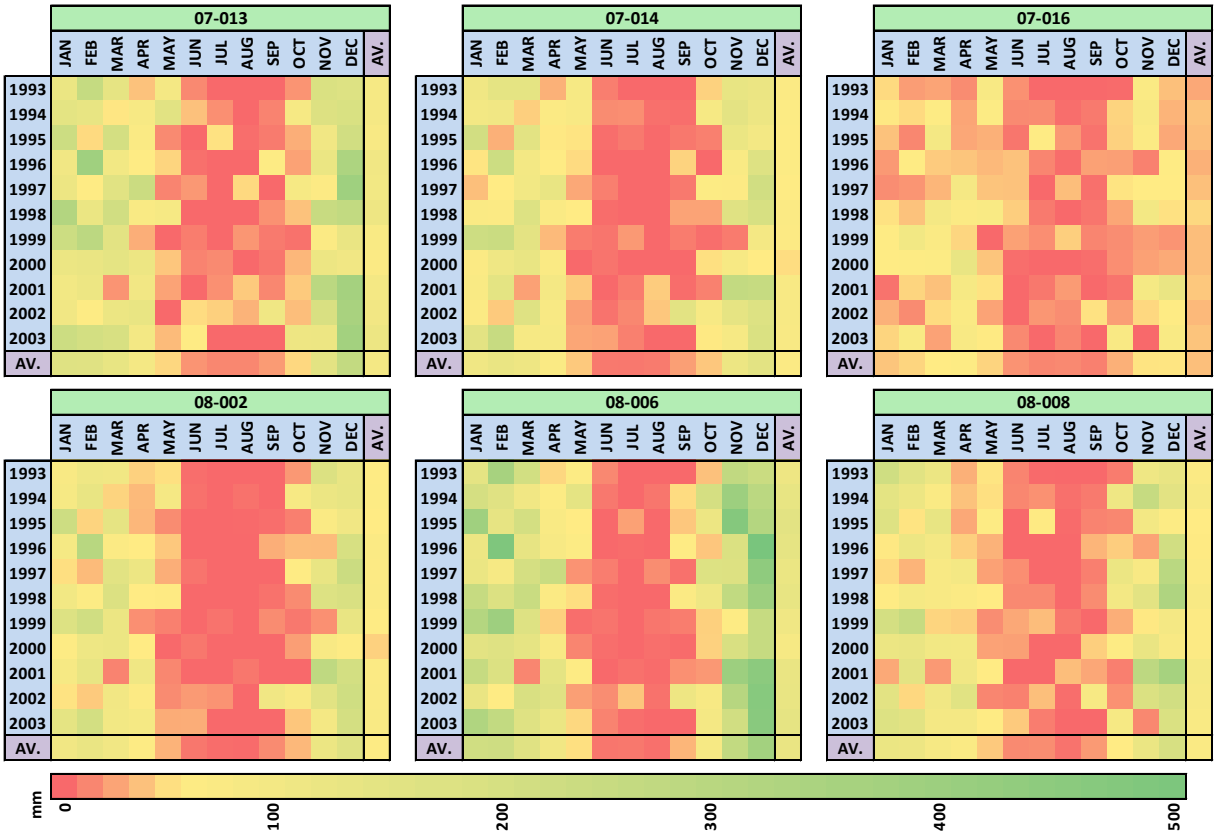

Fig. 2 The heatmaps of the monthly observations of the precipitation stations

Table 1. The descriptive statistics and the percentiles of the observations of the selected stations

	Station	07-013	07-014	07-016	08-002	08-006	08-008
	Elevation (m)	885	630	825	240	730	240
Statistic	Sample Size	132	132	132	132	132	132
	Range	392.1	266.7	132.9	303.0	**525.9**	366.5
	Mean	95.9	69.0	36.4	65.1	**129.3**	73.3
	Variance	9877	4789	890	5067	**19016**	5829
	Std. Deviation	99.4	69.2	29.8	71.2	**137.9**	76.3
	Coef. of Variation	1.04	1.00	0.82	**1.09**	1.07	1.04
	Std. Error	8.7	6.0	2.6	6.2	**12.0**	6.6
	Skewness	1.2	1.0	0.8	1.1	1.1	**1.5**
	Excess Kurtosis	1.0	0.2	0.2	0.6	0.4	**2.3**
Percentiles	Min	0.0	0.0	0.0	0.0	0.0	0.0
	5%	0.0	0.0	0.0	0.0	0.0	0.0
	10%	0.0	0.0	0.4	0.0	0.0	0.0
	25% (Q1)	**14.3**	8.0	13.0	2.8	9.4	13.3
	50% (Median)	62.5	52.4	30.4	40.6	**78.7**	49.2
	75% (Q3)	136.3	103.8	56.3	107.1	**204.0**	111.4
	90%	238.2	171.7	79.2	177.4	**355.5**	174.6
	95%	330.6	223.0	93.2	214.4	**452.7**	235.0
	Max	392.1	266.7	132.9	303.0	**525.9**	366.5
	Best-Fit Distribution	J. SB	J. SB	WAK	J. SB	J. SB	J. SB

2.3. Determination of interannual variability.

Table 2 shows the Pearson's (r), the horizontal (rh) and the vertical (rv) cross correlations among the stations. All horizontal correlation values calculated between the stations are higher than the Pearson's

correlations and all vertical correlation values are lower than the Pearson's correlation. The highest Pearson's (0.951), horizontal (0.954) and vertical (0.888) correlations were obtained between the stations 07-014 and 08-002 while the lowest correlations were obtained between the stations 07-016 and 08-006 ($r = 0.569$, $r_u = 0.578$ and $r_v = 0.507$). This result is also validated by the heatmaps in Figure 2 where the precipitations of the station 07-016 significantly varies from the others. The highest difference between the Pearson's correlation and vertical correlation (0.230) was obtained for the stations 07-013 and 08-008 showing that the relationship between the long year monthly averages of these stations is not as high as the relationship pointed out by the Pearson's correlation.

Table 2. The Pearson's (r), the horizontal (r_h) and the vertical (r_v) correlations between all stations

	07-014	07-016	08-002	08-006	08-008
07-013	$r = 0.897$ $r_h = 0.902$ $r_v = 0.784$	$r = 0.638$ $r_h = 0.645$ $r_v = 0.594$	$r = 0.923$ $r_h = 0.924$ $r_v = 0.819$	$r = 0.923$ $r_h = 0.929$ $r_v = 0.809$	$r = 0.830$ $r_h = 0.842$ $r_v = 0.600$
07-014		$r = 0.683$ $r_h = 0.696$ $r_v = 0.609$	$r = 0.951$ $r_h = 0.954$ $r_v = 0.888$	$r = 0.888$ $r_h = 0.895$ $r_v = 0.783$	$r = 0.860$ $r_h = 0.865$ $r_v = 0.732$
07-016			$r = 0.638$ $r_h = 0.648$ $r_v = 0.555$	$r = 0.569$ $r_h = 0.578$ $r_v = 0.507$	$r = 0.660$ $r_h = 0.664$ $r_v = 0.626$
08-002				$r = 0.906$ $r_h = 0.911$ $r_v = 0.771$	$r = 0.826$ $r_h = 0.835$ $r_v = 0.615$
08-006					$r = 0.837$ $r_h = 0.853$ $r_v = 0.592$

Even though there are some significant differences between the correlation values, the correlations obtained by taking the whole series into account are not sufficient for assessing the interannual variability of the investigated precipitation series. The interannual associations and irregularities between the observations of the stations are determined by dividing each year into 2, 3, 4 and 6 subgroups. In each division, a matrix called subgroup matrix is generated by using the subgroups as shown in Figure 7. When a row is divided into two subgroups, a subgroup matrix of size 2 x 6 (consisting of 2 rows and 6 columns) is formed as shown in the figure. For each year, four different subgroup matrices of sizes 2 x 6, 3 x 4, 4 x 3 and 6 x 2 are generated by using the 12 monthly observations in the row. Then the horizontal and vertical correlations between the 4 subgroup matrices are calculated for each year between the compared stations with the aim of finding the outliers that cause lower correlations. This approach enables determination of many temporal and numerical associations between the compared matrices as shown with the examples below. For example, when the row containing 12 monthly values for a year is divided into 3 subgroups, then the horizontal correlation between the 3 x 4 subgroup matrices measures the association according to the averages of each subgroup and reflects the influence of interannual variability. Similarly, the vertical correlation between the 3 x 4 subgroup matrices reflects the relationships according to the averages of the 1st, 2nd, 3rd and 4th observations of each subgroup and detects associations or irregularities between each month of the compared subgroups. The details of the implementation of the two-dimensional correlation method for finding the outliers is presented below for assessing the relationships between the stations 07-013 and 07-016. The presented procedure is repeated for all station pairs.

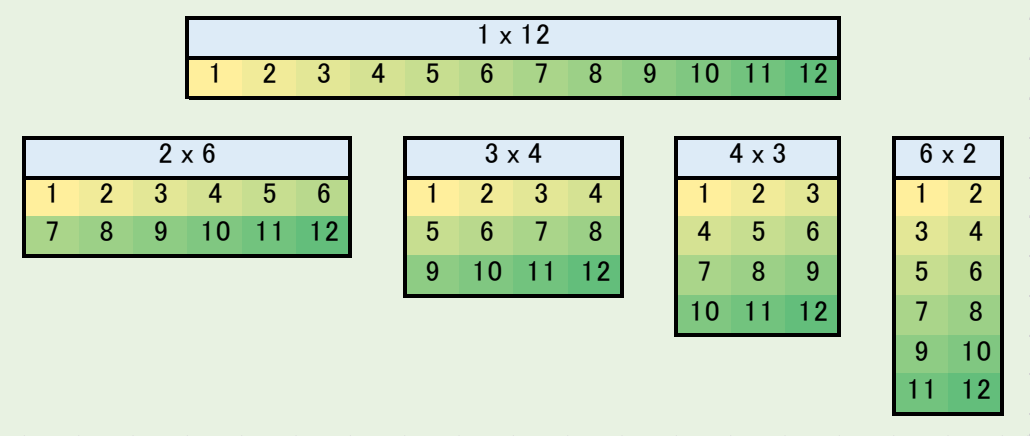

Fig. 3 The 2 x 6, 3 x 4, 4 x 3 and 6 x 2 subgroup matrices generated by dividing the 1 x 12 matrix into 2, 3, 4 and 6 subgroups

3.4. Calculation of Correlations between the Stations 07-013 and 07-016

The Pearson's correlation between the monthly total precipitation records of the stations 07-013 and 07-016 calculated by using Equation 1 is 0.638. The horizontal correlation (0.645) between the stations is 1.1% higher and the vertical correlation (0.594) is 6.9% lower than the Pearson's correlation showing that there is a higher inconsistency between the lines (years in this case) than that pointed out by Pearson's correlation. Table 3 shows the Pearson's correlations between each year and the horizontal and vertical correlations between each subgroup matrices of each year (m and n are the number of rows and columns of the subgroup matrices respectively). The lowest direction-based correlations are shown in bold font for each year. All of the Pearson's correlations between the annual series (shown in the bottom line of the table) are lower than the Pearson's correlation calculated between the whole series and vary between 0.507 (for 1995) and 0.837 (for 2001). The horizontal and vertical correlations between the subgroup matrices vary between -0.131 (the horizontal correlation for the 6x2 matrix for 2000) and 0.939 (the horizontal correlation for the 6x2 matrix for 1996).

Table 3. Correlations between rows and subgroup matrices of stations 07-013 and 07-016

		1993		1994		1995		1996		1997		1998	
m	**n**	r_u	r_v	r_u	r_v	r_u	r_v	r_u	r_v	r_u	r_v	r_u	r_v
2	6	0.533	0.766	0.717	0.769	0.522	0.553	0.797	0.802	0.592	0.745	0.623	0.728
3	4	0.658	0.659	0.687	0.860	0.476	0.564	**0.705**	0.738	0.562	**0.342**	0.552	0.606
4	3	0.560	0.481	**0.387**	0.698	**0.318**	0.524	0.821	0.782	0.579	0.662	0.553	0.632
6	2	**0.217**	0.611	0.557	0.783	0.447	0.457	0.939	0.726	0.343	0.508	**0.495**	0.595
r		0.574		0.727		0.507		0.746		0.591		0.626	

		1999		2000		2001		2002		2003	
m	**n**	r_u	r_v	r_u	r_v	r_u	r_v	r_u	r_v	r_u	r_v
2	6	0.785	0.868	0.666	0.812	0.864	0.832	0.698	0.899	0.556	0.761
3	4	0.646	0.850	0.271	0.794	0.835	**0.801**	0.536	0.609	0.430	0.613
4	3	**0.420**	0.826	0.808	0.771	0.852	0.831	0.704	0.695	**0.395**	0.657
6	2	0.670	0.838	**-0.131**	0.760	0.841	0.833	**0.440**	0.717	0.546	0.557
r		0.826		0.743		0.837		0.696		0.572	

Table 3 provides all the clues about the relationships and irregularities between the observations of the stations. For example, the horizontal correlation between the 3x4 subgroup matrix for the year 2000 is 0.271 while the Pearson's correlation for the year is 0.743. The horizontal correlation for the same year between the 6x2 subgroup matrices is -0.131. These contrasting correlation values are used for determining the inversely related parts and outliers of the compared data series as explained below.

Figure 4 shows the time series graph and the scatterplot of the stations 07-013 and 07-016 for the year 2000 and Figure 5 shows the 3x4 and 6x2 sized subgroup matrices for both stations for the same year together with the averages of each subgroup and correlations between the subgroups. The correlation between the first subgroup (first row) of the 3x4 sized matrices is -0.45. This low correlation value is caused by the inversely related 3rd and 4th (March and April) values. The 3rd value (148.5 mm) of station 07.013 is higher than the subgroup average (129.6 mm) while the 3rd value (66.2 mm) of station 07-016 is lower than the subgroup average (80.9 mm) and the 4th value (119.6 mm) of station 07-013 is lower than the subgroup average while the 4th value (132.9) of station 07-016 is higher than the subgroup average. This inverse relationship causes negative correlation (-0.45) between the first subgroups consequently the horizontal correlation between 3x4 matrices also reflects this inverse relationship as the row averages are used in the determination of the horizontal correlation. Similarly, 3 out of 6 correlations between the 6x2 sized subgroup matrices for the same year are equal to -1 showing the inverse relationships between the 2nd, 4th and 6th subgroups. The negative correlations between the observations in March-April, July-August and November-December reveal the inverse relationships which also can be observed by making a careful visual inspection of the time series data (a scatterplot which is very helpful in determining outliers is not useful for this purpose).

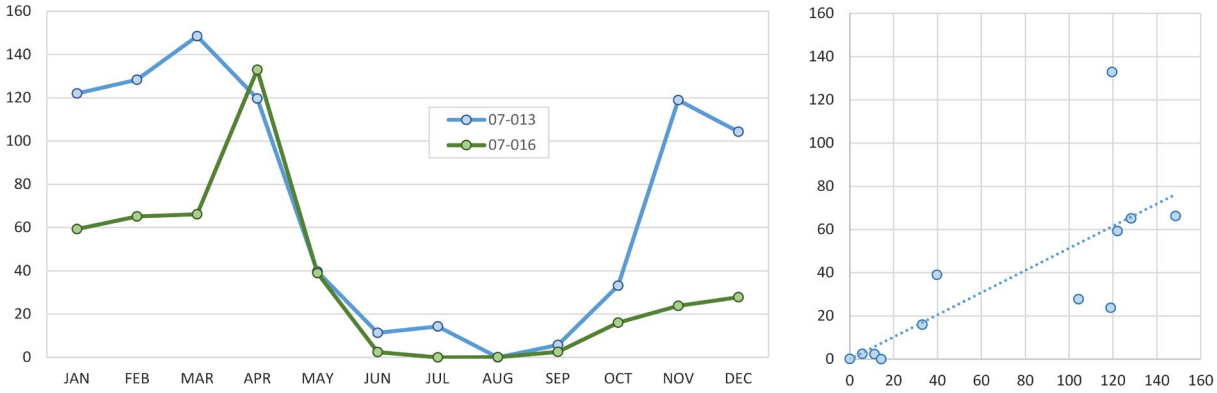

Fig. 4 The time series graph and the scatterplot of the observations of the stations 07-013 and 07-016 for the year 2000

07-013				AV.
122.0	128.3	148.5	119.6	129.6
39.7	11.3	14.3	0.0	16.3
5.7	33.1	118.9	104.3	65.5

r
-0.45
0.93
0.92

07-016				AV.
59.3	65.1	66.2	132.9	80.9
39.0	2.4	0.0	0.1	10.4
2.5	16.0	23.8	27.8	17.5

07-013		AV.
122.0	128.3	125.2
148.5	119.6	134.1
39.7	11.3	25.5
14.3	0.0	7.2
5.7	33.1	19.4
118.9	104.3	111.6

r
1.00
-1.00
1.00
-1.00
1.00
-1.00

07-016		AV.
59.3	65.1	62.2
66.2	132.9	99.6
39.0	2.4	20.7
0.0	0.1	0.1
2.5	16.0	9.3
23.8	27.8	25.8

Fig. 5 The 3x4 and 6x2 subgroup matrices of the stations 07-013 and 07-016 for the year 2000 and the correlations between each subgroup

3.5. Detection and influence of outliers

The uppermost point in the scatterplot in Figure 4 shows the observations of the stations in April 2000 (119.6 mm in station 07-013 and 132.9 mm in station 07-016). The subgroups containing these values in the matrices in Figure 5 are inversely correlated causing low horizontal correlation values with the influence of the inverse relationships. The data pair in April 2000 also produces the highest Mahalanobis distance equal to 6.0 in multivariate outlier detection analysis while all the remaining pairs have distances below 0.84. It is evident from the time series graph and the scatterplot that the high value in April 2000 in station 07-016 causes the high Mahalanobis distance and it is outside the interquartile range as shown by the descriptive statistics in Table 1. IBM SPSS software was used to check if this value is an outlier and the descriptive statistics analysis in the software reported the April 2000 observation as an outlier when the observations of the years 1999 and 2000 were concerned (The 12 values in a year was not sufficient for the conventional outlier detection method used by the software to determine the outlier when only the year 2000 was evaluated). The boxplot in Figure 6 shows the only univariate outlier of the station 07-016 in the years 1999 and 2000 as determined by IBM SPSS. The observation of station 07-013 in April 2000 is not an outlier because it is within the interquartile range of the investigated period (1993-2003). These results show that the presented direction-based correlation might be used for detecting outliers even with a low number of observations.

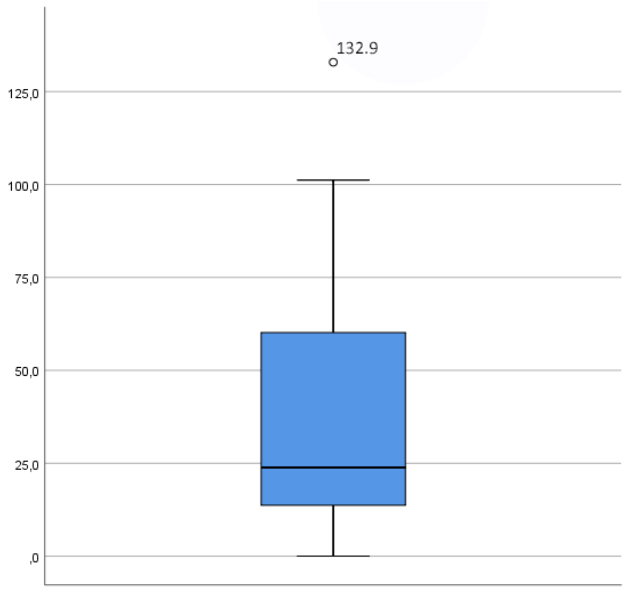

Fig. 6 The boxplot showing the outlier of the observations of the station 07-016 for the years 1999 and 2000

To test the influence of outliers on the value of Pearson's correlation and the direction- based correlations, the outlier of station 07-016 in April 2000 (132.9 mm) is replaced with the long year average of the station in April (62.85 mm) and the correlations are recalculated. After the intervention on this single observation, the Pearson's correlation for the year 2000 increased from 0.743 to 0.879 and the horizontal correlation between the 3x4 subgroup matrices in 2000 increased from 0.271 to 0.678 and from -0.131 to 0.698 for the 6x2 sized subgroup matrices. These results show that the low horizontal correlations were caused by the single outlier in the observations of the station 07-016 and that the direction-based correlation can be used reliably for detecting outliers and inversely related sections in observed data series.

The low values of direction-based correlations in Table 3 point out that there are more outliers or inversely related periods in the compared data series. For example, the horizontal correlation for the 4x3 sized matrices in 1999 is 0.420 while the Pearson's correlation is 0.826. By looking at the averages and correlations between the subgroups, it is determined that the low horizontal correlation value is caused by the inverse relationship in the months November and December where the December value in station 07-

016 reached a historically low level (18.3 mm). This value might easily be ignored when all the observations of the station is included in an outlier analysis but it must never be forgotten that, in the northern hemisphere, a low value might be regarded as normal in summer months but it might be an extreme value (in the lower range) when it is experienced in winter months. Consequently, an outlier analysis for hydrologic variables should consider the two-directional behaviour and the presented two-dimensional correlation method proves to be a strong alternative to be used in outlier and irregularity detection studies.

4. The 2DCorr Software

All results presented in this paper were obtained by using the 2DCorr software developed for the implementation of the presented two-dimensional correlation method. 2DCorr is a Visual Basic code making use of the interoperability feature of MS Visual Studio. The software accepts two MS Excel files containing the observed matrices as input data and generates two output files. One file contains all of the calculated correlations: the horizontal, vertical and Pearson's correlations between the matrices, between the rows and between the subgroup matrices of each row. The other output file contains the subgroup matrices for each row for easier post investigation of associations, outliers and irregularities. All computation process is fully automatic and the software only requires the input matrices to calculate all correlations. The software calculates direction-based correlations between subgroup matrices when the number of columns is 12 and the number of rows is not limited. The software is provided under the terms of the GNU Free Documentation License, Version 1.3.

5. Conclusion

The two-dimensional correlation method as used in this study, successfully determines outliers existing in precipitation data with only 12 observed values (months in a year). This feature of the method is enabled by the pairwise approach used in comparisons of the observations instead of considering the distributions as a whole. This advantage might allow researchers to detect outliers which are not found by conventional methods. The presented approach is not only applicable in the area of hydrology, in fact, as it has a general approach, it might be useful in finding outliers in many other areas of research. Future work might aim at applying the method on areas in which outliers have significant importance where the outliers carry a lot of weight where more "normal" data don't. The method might also be easily used for determining univariate outliers by comparing seasonal series, for example, comparison of yearly comparison of observations of a hydrologic station. The 2DCorrel software developed for the implementation of the methodology is freely provided as a supplement for ease of re-implementation of the approach by researchers.

6. References

Aggarwal CC (2016) Outlier Analysis. Springer Publishing Company, Incorporated,

Aggarwal CC, Yu PS (2000) Finding generalized projected clusters in high dimensional spaces. Paper presented at the Proceedings of the 2000 ACM SIGMOD international conference on Management of data, Dallas, Texas, USA,

Anderson AN, Browning JM, Comeaux J, Hering AS, Nychka D (2016) A comparison of automated statistical quality control methods for error detection in historical radiosonde temperatures. Int J Climatol 36:28-42. https://doi.org/10.1002/joc.4327

Barnett V, Barnett PSV, Lewis T (1994) Outliers in Statistical Data. Wiley,

Chebana F, Dabo-Niang S, Ouarda TBMJ (2012) Exploratory functional flood frequency analysis and outlier detection. Water Resour Res 48. https://doi.org/10.1029/2011WR011040

Cohn TA, England JF, Berenbrock CE, Mason RR, Stedinger JR, Lamontagne JR (2013) A generalized Grubbs-Beck test statistic for detecting multiple potentially influential low outliers in flood series. Water Resour Res 49:5047-5058. https://doi.org/10.1002/wrcr.20392

Davies L, Gather U (1993) The identification of multiple outliers. J Am Stat Assoc 88:782-792. https://doi.org/10.1080/01621459.1993.10476339

Dikbas F (2017) A novel two-dimensional correlation coefficient for assessing associations in time series data. Int J Climatol 37:4065-4076. https://doi.org/10.1002/joc.4998

Dikbas F (2018) A new two-dimensional rank correlation coefficient. Water Resour Manage 32:1539-1553. https://doi.org/10.1007/s11269-017-1886-0

Eischeid JK, Baker CB, Karl TR, Diaz HF (1995) The quality control of long-term climatological data using objective data analysis. J Appl Meteorol 34:2787-2795. https://doi.org/10.1175/1520-0450(1995)034<2787:TQCOLT>2.0.CO;2

Grubbs FE (1969) Procedures for detecting outlying observations in samples. Technometrics 11:1-21. https://doi.org/10.1080/00401706.1969.10490657

Gupta M, Gao J, Aggarwal CC, Han J (2014) Outlier detection for temporal data: A survey. IEEE Trans Knowl Data Eng 26:2250-2267. https://doi.org/10.1109/TKDE.2013.184

Hawkins DM (1980) Identification of outliers / D.M. Hawkins. Monographs on applied probability and statistics., vol Accessed from https://nla.gov.au/nla.cat-vn18307. Chapman and Hall, London ; New York

Hodge VJ, Austin J (2004) A survey of outlier detection methodologies. Artif Intell Rev 22:85-126. https://doi.org/10.1023/B:AIRE.0000045502.10941.a9

Hu S (1987) Problems with outlier test methods in flood frequency analysis. J Hydrol 96:375-383. https://doi.org/10.1016/0022-1694(87)90167-3

Hunt BG (2007) Climatic outliers. Int J Climatol 27:139-156. https://doi.org/10.1002/joc.1379

Jeong J, Park E, Han WS, Kim K, Choung S, Chung IM (2017) Identifying outliers of non-Gaussian groundwater state data based on ensemble estimation for long-term trends. J Hydrol 548:135-144. https://doi.org/10.1016/j.jhydrol.2017.02.058

Kirk AJ, McCuen RH (2008) Outlier detection in multivariate hydrologic data. J Hydrol Eng 13:641-646. https://doi.org/10.1061/(ASCE)1084-0699(2008)13:7(641)

Lamontagne Jonathan R, Stedinger Jery R, Cohn Timothy A, Barth Nancy A (2013) Robust national flood frequency guidelines: What is an outlier? World Environmental and Water Resources Congress 2013

Lamontagne JR, Stedinger JR (2016) Examination of the Spencer-McCuen outlier-detection test for Log-Pearson type 3 distributed data. J Hydrol Eng 21. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001321

Lanzante JR (1996) Resistant, robust and non-parametric techniques for the analysis of climate data: Theory and examples, including applications to historical radiosonde station data. Int J Climatol 16:1197-1226. https://doi.org/10.1002/(SICI)1097-0088(199611)16:11<1197::AID-JOC89>3.0.CO;2-L

Legates DR, Davis RE (1997) The continuing search for an anthropogenic climate change signal: Limitations of correlation-based approaches. Geophys Res Lett 24:2319-2322

Legates DR, McCabe Jr GJ (1999) Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation. Water Resour Res 35:233-241. https://doi.org/10.1029/1998WR900018

Pegram G (1997) Patching rainfall data using regression methods. 3. Grouping, patching and outlier detection. J Hydrol 198:319-334. https://doi.org/10.1016/S0022-1694(96)03284-2

Rosner B (1983) Percentage points for a generalized esd many-outlier procedure. Technometrics 25:165-172. https://doi.org/10.2307/1268549

Spencer CS, McCuen RH (1996) Detection of outliers in Pearson type III data. J Hydrol Eng 1:2-10. https://doi.org/10.1061/(ASCE)1084-0699(1996)1:1(2)

Tsay RS (1988) Outliers, level shifts, and variance changes in time series. J Forecast 7:1-20. https://doi.org/doi:10.1002/for.3980070102

Wu E, Liu W, Chawla S Spatio-temporal outlier detection in precipitation data. In, Berlin, Heidelberg, 2010. Knowledge Discovery from Sensor Data. Springer Berlin Heidelberg, pp 115-133

Zhang Y, Meratnia N, Havinga P (2010) Outlier detection techniques for wireless sensor networks: A survey. IEEE Communications Surveys & Tutorials 12:159-170. https://doi.org/10.1109/SURV.2010.021510.00088

Figures

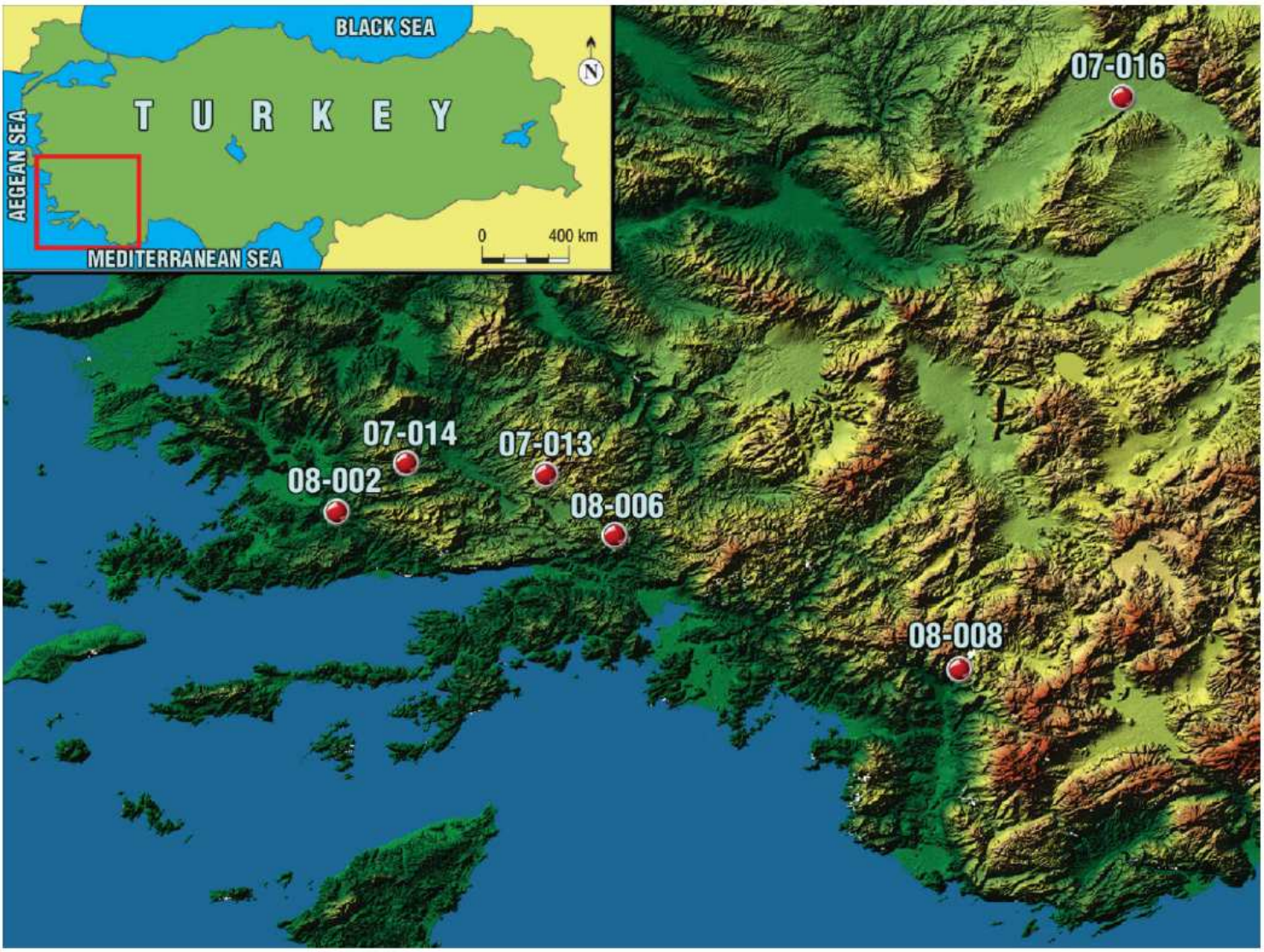

Figure 1

The precipitation observation stations selected for implementation of the outlier detection method

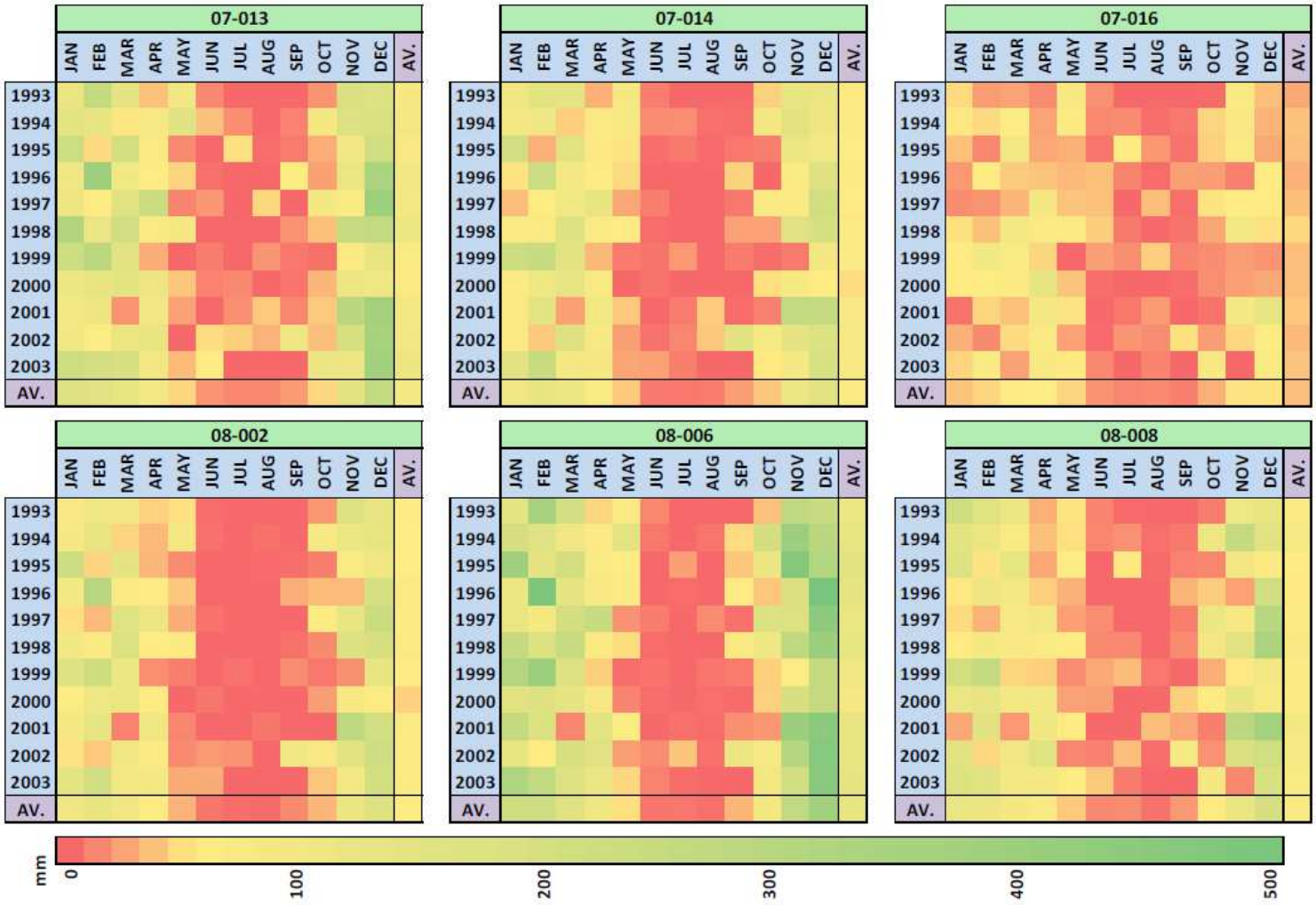

Figure 2

The heatmaps of the monthly observations of the precipitation stations

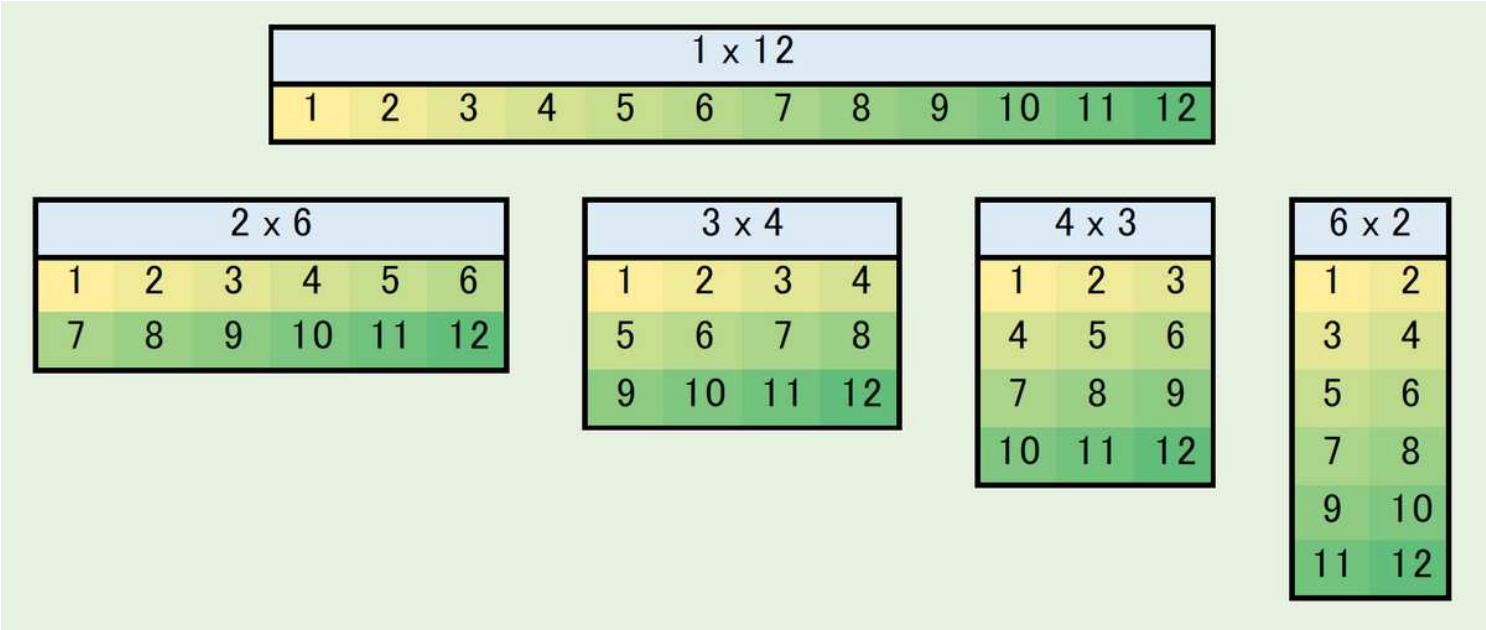

Figure 3

The 2 x 6, 3 x 4, 4 x 3 and 6 x 2 subgroup matrices generated by dividing the 1 x 12 matrix into 2, 3, 4 and 6 subgroups

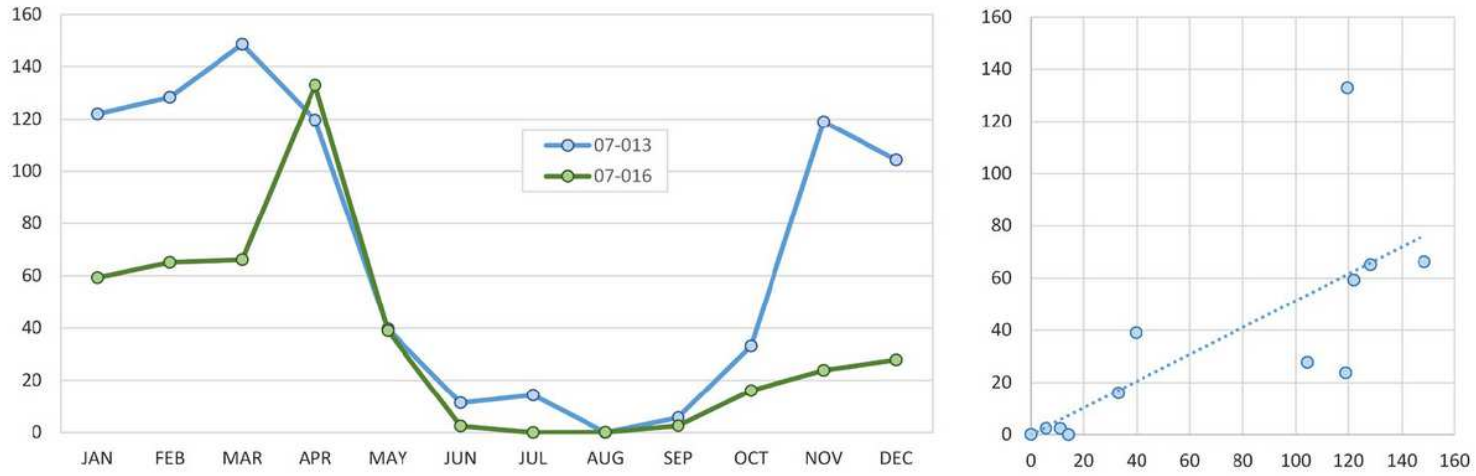

Figure 4

The time series graph and the scatterplot of the observations of the stations 07-013 and 07-016 for the year 2000

07-013				AV.
122.0	128.3	148.5	119.6	129.6
39.7	11.3	14.3	0.0	16.3
5.7	33.1	118.9	104.3	65.5

r
-0.45
0.93
0.92

07-016				AV.
59.3	65.1	66.2	132.9	80.9
39.0	2.4	0.0	0.1	10.4
2.5	16.0	23.8	27.8	17.5

07-013		AV.
122.0	128.3	125.2
148.5	119.6	134.1
39.7	11.3	25.5
14.3	0.0	7.2
5.7	33.1	19.4
118.9	104.3	111.6

r
1.00
-1.00
1.00
-1.00
1.00
-1.00

07-016		AV.
59.3	65.1	62.2
66.2	132.9	99.6
39.0	2.4	20.7
0.0	0.1	0.1
2.5	16.0	9.3
23.8	27.8	25.8

Figure 5

The 3x4 and 6x2 subgroup matrices of the stations 07-013 and 07-016 for the year 2000 and the correlations between each subgroup

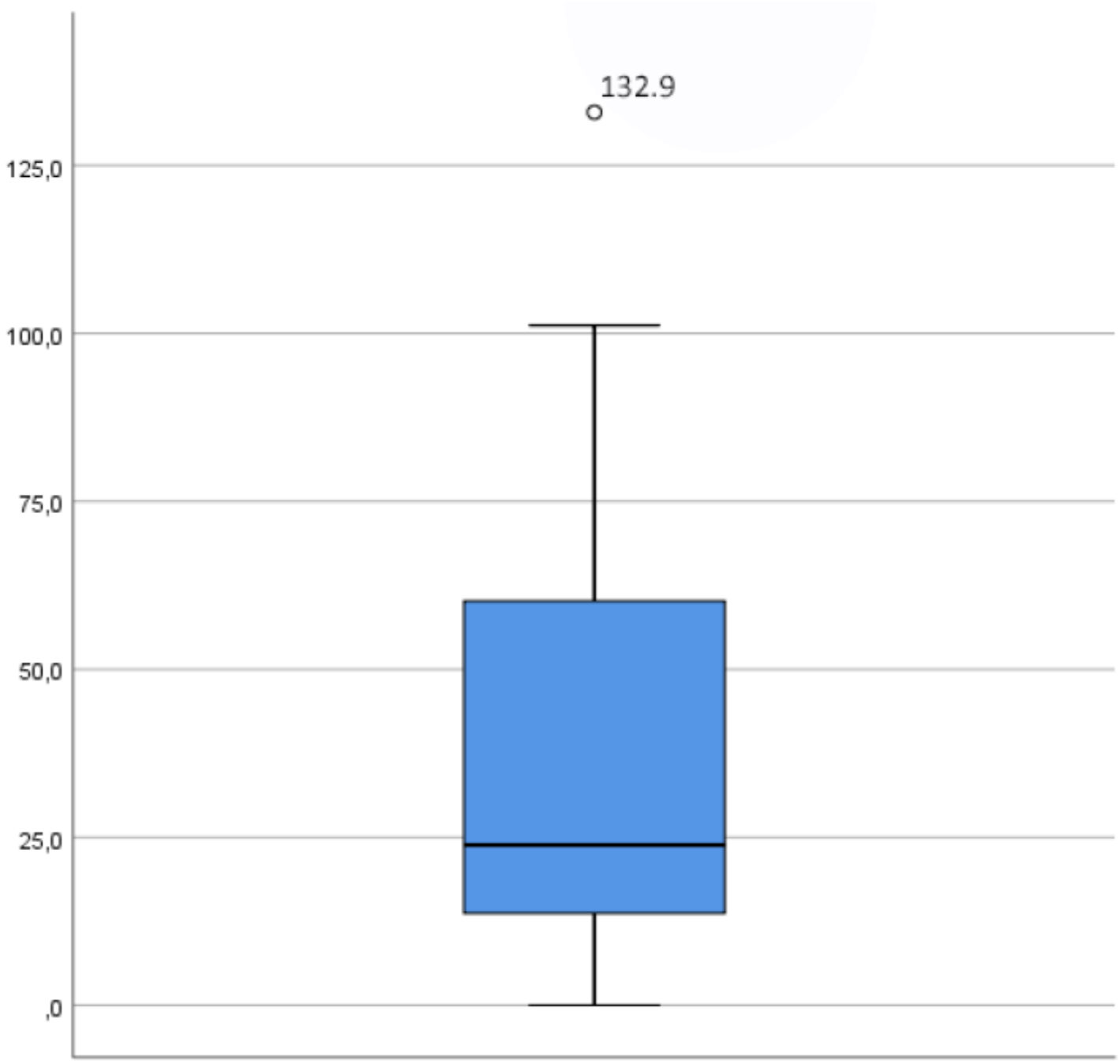

Figure 6

The boxplot showing the outlier of the observations of the station 07-016 for the years 1999 and 2000

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementary.Material.pdf