Proceedings of the Mediterranean Meeting on "Monitoring, modelling and early warning of extreme events triggered by heavy rainfalls". PON 01_01503 - MED-FRIEND project
University of Calabria, Cosenza (Italy), June 26th-28th, 2014

ASSESSMENT OF METHODS FOR OUTLIER DETECTION AND TREATMENT IN FLOOD FREQUENCY ANALYSIS

J. Plavšić[1]*, V. Mihailović*[2] *and B. Blagojević*[3]

[1] University of Belgrade, Faculty of Civil Engineering, Belgrade, Serbia
[2] University of Belgrade, Faculty of Forestry, Belgrade, Serbia
[3] University of Niš, Faculty of Civil Engineering and Architecture, Niš, Serbia

ABSTRACT

In the framework of the development of the guidelines for flood flow assessment in Serbia, the outlier detection procedures and their consequent treatment are investigated. The outcomes of different methods are analysed in terms of both detected outliers and estimated flood quantiles using data from 68 hydrologic stations in Serbia. Three different tests are applied which are commonly proposed in the outlier detection procedures and which are all based on the assumption of normally distributed samples (original or after transformation). The tests identified the same series with outliers, but the number of outliers differs from test to test. Removal of low outliers from the sample has a significant impact on the resulting flood frequency and quantiles. The log-Pearson type III and the log-normal distributions are highly sensitive to presence of low outliers, while the general extreme value distribution is not. It is also concluded that these tests are not appropriate for flood data that cannot be assumed to be log-normally distributed (or normal after some other transformation).

1. INTRODUCTION

The major uncertainty associated with the flood frequency analysis stems from the limited length of the observed hydrologic records and uncertainty inherent in extreme flood flow measurement and estimation from the extrapolated stage-discharge relationships. In practical applications of flood frequency analysis, these sources of uncertainty lead to a difficulty in recognizing the underlying distribution of the flood flows. One of the typical problems is that the annual maximum flows in different years are caused by different runoff generation mechanisms, like snowmelt or heavy rainfall. This is the case when the flood data sample comes from a mixed population, and it requires separating the floods into two or more groups according to the flood generation mechanism and later combining the separately estimated distributions into a composite one. However, the number of the flood events due to one of the generating mechanisms is usually small for defining a reliable probability distribution.

The problem of outliers in the flood data is also closely related to the problem of a unknown parent distribution of floods and the uncertainty in flood flow measurements and estimation. The presence of outliers in a data sample can lead to problems in formulating a probabilistic model and fitting an appropriate theoretical distribution from the observations. Given that finding a correct model that would allow extrapolation of the flood flows outside the range of observed values is very important for flood estimation, the effect of high outliers on the choice of the theoretical distribution is usually considered to be crucial. However, it has been recognized that low outliers can significantly affect not only the choice of the best distribution, but also the distribution parameter estimates and consequently the flood quantile estimates.

The most popular method proposed in the literature for detection and treatment of outliers is based on the approach presented in the well-known Bulletin 17B (IACWD, 1982). However, the problem of outliers is still an open problem; calls for revision of these methods have been made since the available series are now 30 years longer and new perspective of the outlier problem might be possible (Stedinger and Griffis, 2008; 2011).

In Serbia, development of the "Guidelines for flood flow assessment at hydrologic stations (at gauged catchments)" is currently in progress (Blagojevic et al, 2013). In this framework, a study has been undertaken with an aim to determine the differences that arise from different approaches to outlier detection and treatment and to identify the problems that may arise in application of the proposed procedures. A preliminary analysis of the available data (Blagojevic et al, 2014) has shown that the main doubts in the outlier detection emerge from the application of the traditional Grubbs-Beck test for outlier detection, which is valid for normal (or transformed to normal) data while the flood data is assumed to follow other distributions. Moreover, the annual maximum flood series that exhibit outliers usually come from mixed populations, and a question remains whether the outliers are the exceptional values of small probability of occurrence or just "regular" values from another population.

In this paper, we investigate the outlier detection procedures and consequent treatments of outliers. We analyse the outcomes of the different methods for outlier detection using data from 68 hydrologic stations in Serbia in terms of both detected outliers and estimated flood quantiles. We also focus on these results in respect to an assumed underlying distribution of the annual maximum floods and try to make recommendations for practitioners.

2. DETECTION AND TREATMENT OF OUTLIERS IN HYDROLOGIC DATA

2.1 Outlier detection

The prevailing method for the outlier detection and accommodation in hydrologic series is the procedure given in the Bulletin 17B (IACWD, 1982). These guidelines recommend the Grubbs-Beck test for detecting outliers, which is applied as the one-sided test by comparing the logarithmic flows in the sample to a threshold corresponding to the critical value of the test statistic at 10% significance level.

The Grubbs-Beck test, originally proposed by Grubbs (1969) and Grubbs and Beck (1972), examines the presence of either high or low outliers in the series. It is used under the assumption that the data in the original form or after some transformation is normally distributed. If x_i (i = 1, 2, ..., n) are the data values from the sample of size n sorted from the smallest to the largest value, the test statistic of the Grubbs-Beck test is the studentized deviate of the lower or upper most extreme value in the sample x_1 or x_n, respectively:

$$K = \frac{\overline{x} - x_1}{s} \text{ or } K = \frac{x_n - \overline{x}}{s} \tag{1}$$

where $\bar{x}$ and s are the sample mean and standard deviation, and K is the test statistic that depends on the significance level α and the sample size n. For the 10% significance level and sample size n, an approximate expression for K is given with (Stedinger et al, 1993):

$$K_{10\%} = -0.9043 + 3.345\sqrt{\log n} - 0.4046 \log n \tag{2}$$

Although the Grubbs-Beck test is aimed at testing a single outlier, the Bulletin 17B (IACWD, 1982) recommends that all values below the low threshold defined by $K_{10\%}$ are considered low outliers and all values above the upper threshold are considered high outliers. The order of testing depends on sample skew. For samples with negative skew (smaller than –0.4), low outliers are tested first and high outliers are tested in the sample censored for detected low outliers. For samples with skew greater than –0.4, high outliers are tested first and low outliers are tested second. In the latter case, if the high outliers are identified in the record containing historical floods, the sample statistics are adjusted for historical information before testing for low outliers for skews greater than 0.4, and after testing for low outliers for skews smaller than 0.4.

The major criticism to application of the Grubbs-Beck test to flood data is twofold. The first is that this test cannot deal with multiple outliers and the second is related to the assumption of normality (Spencer and McCuen, 1996). The first problem is related to sequential application of this test by testing e.g. the smallest value, then removing it from the sample if it is identified as a low outlier, and repeating the test on the reduced (censored) sample, and so on. However, the test power is reduced in its repeated applications (Tietjen and Moore, 1972). The sequential procedure can also lead to the masking effect in case when a group of values is separated from the rest of the values. The sequential testing can then be unable to detect a single value as an outlier within the group of suspected outliers. On the other hand, if the tests for multiple outliers are made on a group of values simultaneously, then the swamping effect may occur. In this case all values may be declared outliers while in fact there is only a single outlier.

Grubbs (1950) also developed a test for testing multiple outliers as a group of k values at one of the distribution tails, with null hypothesis that all the values come from the same normal distribution. This test was later also proposed by Tietjen and Moore (1972), with the test statistic for examining k greatest values:

$$L_k = \frac{\sum_{i=1}^{n-k} (x_i - \bar{x}_k)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{3}$$

where

$$\bar{x}_k = \frac{1}{n-k} \sum_{i=1}^{n-k} x_i \tag{4}$$

In case when k smallest values are examined, the summation in the nominator in equations (3) and (4) ranges from $i = k + 1$ to $i = n$. The tables of critical values for L_k for sample sizes up to $n = 50$ and up to $k = 10$ outliers can be found in Tietjen and Moore (1972), and for sample sizes up to $n = 100$ with k up to 4 in Verma and Quiroz-Ruiz (2006). Tietjen and Moore (1972) recommend the backward procedure with this test in order to avoid the masking effect. This means that a large k is assumed at the beginning. If the null hypothesis is not rejected, then k is reduced for one and the test is repeated.

However, this procedure does not prevent the swamping effect. Tietjen and Moore (1972) also examined several alternatives on how to assume the number k of outliers if no outliers are anticipated, and they showed that this should be the number of observations below or above the largest gap in data values.

McCuen (2002) proposed the test of Rosner (1975) which can identify more than one outlier, but the number k of potential outliers must also be specified in advance. The test looks for both low and high outliers by looking at k values that are the farthest from the mean. All k values are considered outliers if the null hypothesis that no outliers are present is rejected, but none of the individual values can be considered an outlier if the null hypothesis is not rejected. However, by identifying the outliers on both tails, this test is not helpful for the flood frequency analysis since the nature of high and low outliers in flood flow series is completely different and they cannot be treated in the same manner. If both high and low outliers are eliminated from the sample, it would get closer to the normality assumption; however, in the flood statistics we do not want to exclude the greatest values and loose valuable information about the upper distribution tail.

More recently, Cohn et al. (2013) developed the generalized test for detecting multiple potentially influential low outliers in flood series, based on the Grubbs-Beck test statistic (1) and similar to the Rosner's (1975) method but without an *a priori* defined number of outliers. This is a great step forward for the flood frequency analysis and the hydrologic statistics in general. However, the method is rather complicated for use by practitioners unless incorporated in a statistical software package.

All the previously mentioned tests and the critical values of the test statistics are based on the assumption of normality. This means that rejection of the null hypothesis may be a result of a non-normal parent population rather than the presence of outliers in the sample. The usual recommendation is to transform the data in order to make it closer to normality, including logarithmic, Wilson-Hilferty, Box-Cox and other transformations. Logarithmic transformation is the most common to hydrologic practice. While it is generally thought that it provides a sufficient degree of normality in the sample, at the same time this transformation can provoke occurrence of low outliers because the small values get more weight in the log-transformed sample.

For flood flow series and their logarithms for which distributions other than normal are assumed, NRCS (2012) implicitly proposes in the practical examples that the critical values x_U and x_L for declaring a lower or an upper outlier respectively in the Grubbs-Beck test can be estimated from the assumed distribution for the normal probability of test statistic K:

$$x_L = F_X^{-1}(1-p) \quad \text{or} \quad x_U = F_X^{-1}(p), \quad p = \Phi(K) \tag{5}$$

where $\Phi(.)$ is the normal cumulative distribution function, and $F_X(.)$ is the assumed parent distribution for the sample data. This approach is referred to as the probability mapping.

The sample data can be tested for normality by different tests, including the goodness-of-fit tests (e.g. the probability plot correlation coefficient test or the Anderson-Darling test), or by looking into the confidence intervals for sample skewness and kurtosis (see e.g. Kottegoda and Rosso, 2008).

2.2 Treatment of outliers

Ghosh and Vogt (2012) state: "Outlier treatment is an art." In the flood frequency analysis, high and low outliers have different nature and have different effects on the probability distribution; their treatment should therefore be different. High outliers may indeed be exceptional and rare events, but they can also be a result of measurement errors. Even if there is no proof of a measurement error, uncertainty in the upper end of the stage-discharge relationship is usually high due to its extrapolation and our confidence in extreme values is generally lower. Furthermore, the exceptional values might not have occurred naturally, but under some anthropogenic effect. This calls for a careful review of the extreme values in the record before proceeding to frequency analysis. On the other hand, low outliers are not expected to result from a measurement error, nor they are burdened by high uncertainty from the stage-discharge relationship; however, they can significantly affect the sample statistics and the results of frequency analysis.

In general, the outliers can be excluded from the sample from either lower or upper side (censoring or trimming). This approach is commonly used for low outliers in flood series. Another approach is to adjust the outliers so that their effect on the sample statistics is reduced ("winsorising"). According to Vukmirovic and Pavlovic (2000), possible adjustments of high outliers can be to replace them with: (1) the second greatest value from the same year, (2) the second greatest value in the sample, or (3) an estimate based on the 50- or 100-year flood from the neighbouring stations. High outlier adjustment is generally considered a better procedure than excluding it from the sample, except when a doubt exists that it is caused by measurement error (BLFUW, 2011). Keeping high outliers in the sample and treating them as an ordinary value seem to be a preferred approach because it is believed that they carry important information. However, keeping the high outliers may lead to overestimation of flood flows, while the censoring and winsorising approaches introduce statistical bias and may lead to underestimated flood flows. The Austrian guidelines (BLFUW, 2011) propose, among other approaches, that an apparent return period $T^* > n$ of a high outlier can roughly be estimated as the return period of the same flood event at neighbouring stations or by using information on historical floods.

The Bulletin 17B (IAWCD, 1982) recommends that high outliers should be retained as a part of the systematic record. If they represent historical floods from the non-systematic record, the sample statistics and the plotting positions are adjusted for this historical information. The low outliers are excluded from the record and the distribution function $F_1(x)$ resulting from the censored sample is adjusted for removing k values from the sample of size n by using conditional probability:

$$\begin{aligned} F(x) &= P\{X \le x\} = P\{X \le x \mid X > x_L\}P\{X > x_L\} + \\ &+ P\{X \le x \mid X < x_L\}P\{X < x_L\} = F_1(x) \cdot \frac{n-k}{n} + \frac{k}{n} \end{aligned} \tag{6}$$

The statistical literature (Kottegoda and Rosso, 2008; Barnett and Lewis, 1994) also advocates the use of robust statistical measures (e.g. trimmed mean, interquartile range etc.) instead of those sensitive to outliers. However, we still need more research on the use of such measures in flood frequency analysis.

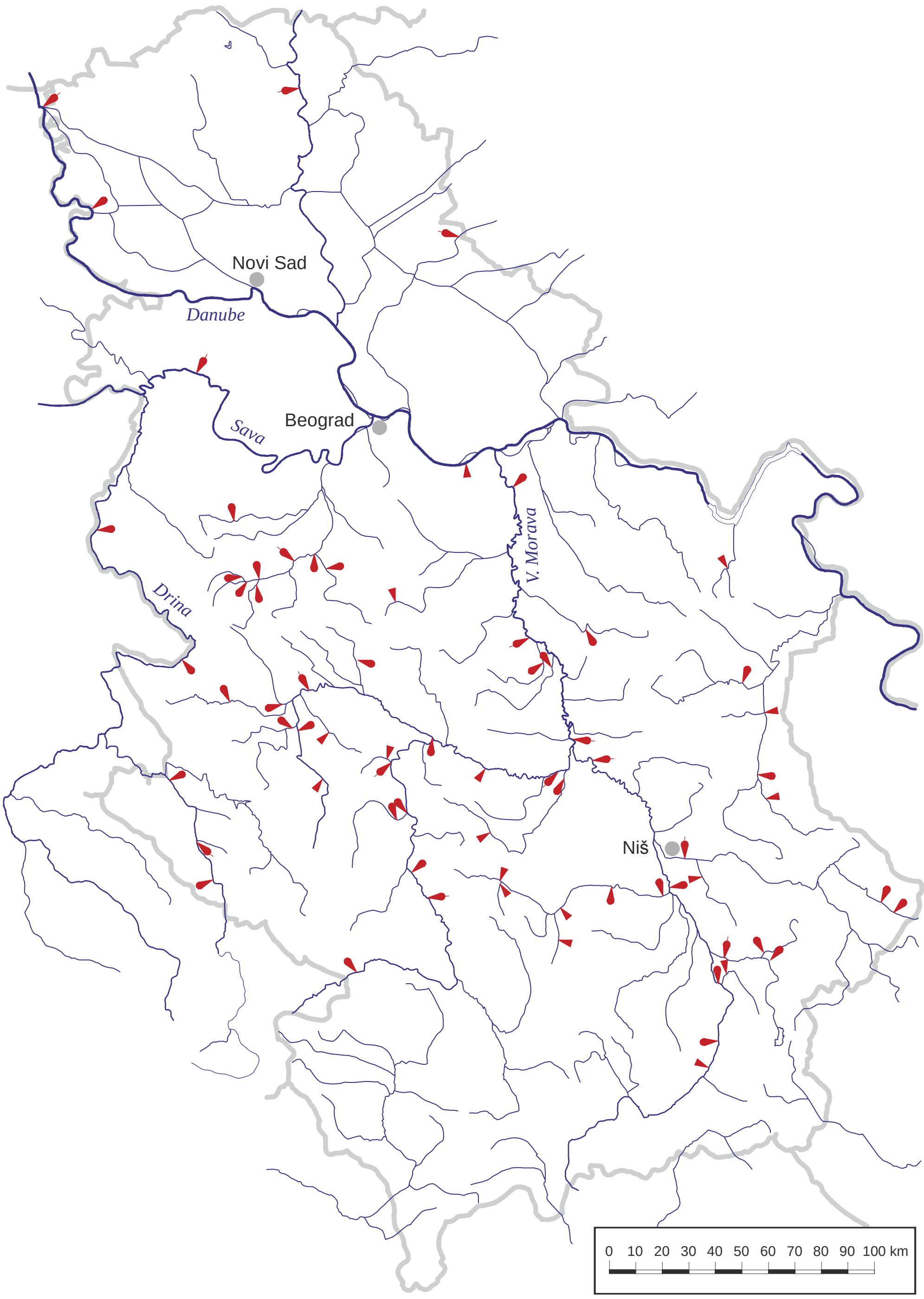

Figure 1. Hydrologic stations in Serbia used in this study.

3. DATA AND METHODS

In this paper, we investigate the outcomes of different approaches for outlier detection and treatment in the annual maximum series of flood flows in Serbia. A total of 68 hydrologic stations at which hydrologic regime is not under significant anthropogenic influence were used in the analysis (Figure 1). The drainage areas range from 84 to 37200 km^2 for the inner Serbia, but stations on the transboundary rivers Sava and Danube are also included. Record lengths span from 37 to 85 years, until 2010. The following approaches were applied for outlier detection:

1. the Bulletin 17B procedure, denoted here with as B17B;

2. the original one-sided Grubbs-Beck test for single outlier detection, denoted GB1, applied sequentially to the sample after removing detected low outliers;
3. the multiple outlier test by Grubbs (1950) according to Tietjen and Moore (1972), denoted TM.

The above listed approaches were applied to the log-transformed data and therefore an assumption of the log-normal underlying distribution was made. All log-transformed series were tested for normality using two tests: the probability plot correlation coefficient (PPCC) test and the Anderson-Darling (AD) test. The two tests did not produce consistent results. The null hypothesis that the log-transformed data are normally distributed could not be rejected by PPCC test in only five cases out of 68, all in the series without detected outliers. The same hypothesis was rejected by the AD test in 12 cases only, including five series that were later shown to contain outliers. In further considerations, the distribution providing the best fit for a particular series was identified by ranking the results from the Anderson-Darling and the Kramer-von Mises goodness-of-fit tests. The general extreme value (GEV) distribution was generally found to be the best fit to the observed data, and the log-Pearson type III (LP3) was equally good for stations with greater positive skew.

A preliminary testing of homogeneity was also undertaken by applying the non-parametric Mann-Whitney or the rank-sum test (Helsel and Hirsch, 2002) to all series by looking into differences between two halves of each sample. The null hypothesis of homogeneity was rejected for 28 stations out of 68. This number includes six stations in which outliers were later detected. This shows that additional investigation would be needed to identify the sources of non-homogeneity and eventually non-stationarity.

4. RESULTS AND DISCUSSION

The results of the outlier detection procedures according to the three tests are presented in Tables 1, 2 and 3 for low outliers, high outliers and both low and high outliers respectively. All tests were applied at the 10% significance level.

Eight series with low outliers detected with the three methods are listed in Table 1. All the methods identified the same series as those with low outliers, so it could be said that the results are rather consistent. The differences are in the number of detected outliers at three stations out of eight (stations 30, 52 and 53), where the sequential application of the Grubbs-Beck test and the Tietjen-Moore test detected two outliers instead of one. After removing the detected number k of low outliers, neither of the series exhibited presence of high outliers with each method.

The effect of removing a different number of outliers is described in Figure 2 for station 53. It can be seen that removal of outliers mostly affects the log-normal (LN) and the log-Pearson type III (LP3) distribution. The performance of the LN distribution becomes much better without low outliers according to the goodness-of-fit tests since the log-sample skew approaches zero. However, the upper tail of this distribution is lowered (Table 4). The LP3 distribution also adjusts to higher skew and results in a higher upper tail, while its goodness-of-fit improves. The GEV distribution is the most resistant to the outlier removal; its upper tail without outliers is just slightly higher and the goodness-of-fit statistics are almost unchanged. Similar results are obtained for other stations with low outliers.

Table 1. Low outliers detected in the series of annual maximum flows (*n* denotes sample size and 'best fit' indicates the result according to the goodness-of-fit tests).

No.	River/Station	Area (km^2)	*n*	Min. (m^3/s)	Max. (m^3/s)	Log skew	Best fit	No. of low outliers *k*		
								B17B	GB1	TM
20	J. Morava/Mojsinje	15390	60	115	1830	-0.40	GEV	1	1	1
29	Kolubara/Valjevo	340	55	7.2	287	-0.31	LN, LP3*	1	1	1
30	Kolubara/Slovac	995	57	28.7	322	-1.06	GEV	1	2	1
31	Kolubara/Beli Brod	1896	52	27.3	767	-0.57	GEV	1	1	1
51	V. Morava/Varvarin	31548	63	336	2550	-0.58	GEV	2	2	2
52	V. Morava/Bagrdan	33446	61	354	2930	-0.53	GEV	1	2	2
53	V. Morava/Lj. Most	37320	63	353	2354	-0.70	GEV	1	2	2
69	Z. Morava/Trstenik	13902	48	160	1750	-0.31	LN, GEV*	1	1	1

*Best fit atfer removing low outliers.

Table 2. High outliers detected in the series of annual maximum flows (*n* denotes sample size and 'best fit' indicates the best result according to the goodness-of-fit tests).

No.	River/Station	Area (km^2)	*n*	Min. (m^3/s)	Max. (m^3/s)	Log skew	Best fit	No. of high outliers		
								B17B	GB1	TM
13	Ibar/Batrage	703	56	27.0	519	0.92	GEV	1	1	2
17	Ibar/Lopatnica Lakat	7818	63	115	1520	0.62	GEV	1	1	3
24	Lužnica/Svođe	319	49	6.6	488	0.05	LN	1	1	1
25	Vlasina/Svođe	350	56	4.86	578	0.48	GEV	1	1	1
48	Lukovska/Merćez	113	42	2.12	106	0.74	GEV	1	1	4
58	Z. Morava/K. Stena	3077	82	99.4	1250	0.18	GEV	1	1	2
60	Z. Morava/Jasika	14721	60	327	1870	0.50	LP3	1	1	1
62	Moravica/Ivanjica	475	64	16.1	429	0.56	GEV	1	1	1

Table 3. Series with both low (LO) and high (HO) outliers (*n* denotes sample size; HO* indicate number of high outliers after removing low outliers).

No.	River/Station	Area (km^2)	*n*	Min. (m^3/s)	Max. (m^3/s)	Log skew	Type	No. of outliers		
								B17B	GB1	TM
18	Studenica/Ušće	540	57	9.84	276	0.19	LO	1	1	1
							HO	1	1	3
							HO*	1	1	3
45	Toplica/Magovo	180	37	2.33	192	0.48	LO	1	1	1
							HO	1	1	2
							HO*	1	1	2

However, the outcomes are different with one or two outliers deleted from the record and the question of the number of low outliers in the sample remains. The values of the goodness-of fit statistics indicate that only the LN distribution benefits significantly from removing two low outliers instead of one, while performance of the GEV and LP3 distributions is almost the same. This is expected since all tests assume that the log-transformed floods are normally distributed and the removal of outliers would actually contribute to getting closer to normality. With this in mind, we performed the tests using the probability mapping given by eq. (5) under the hypothesis that the data is GEV distributed. This procedure gave the same results for low outliers as the tests with the LN assumption for stations 31 and 69, while the other stations from Table 1 did not exhibit low outliers. The GEV mapping also identifies low outlier at station 8, which was not detected under the LN assumption. Although the probability mapping testing procedure may not be theoretically sound, it is indicative that different assumptions about the parent distribution can lead to different results.

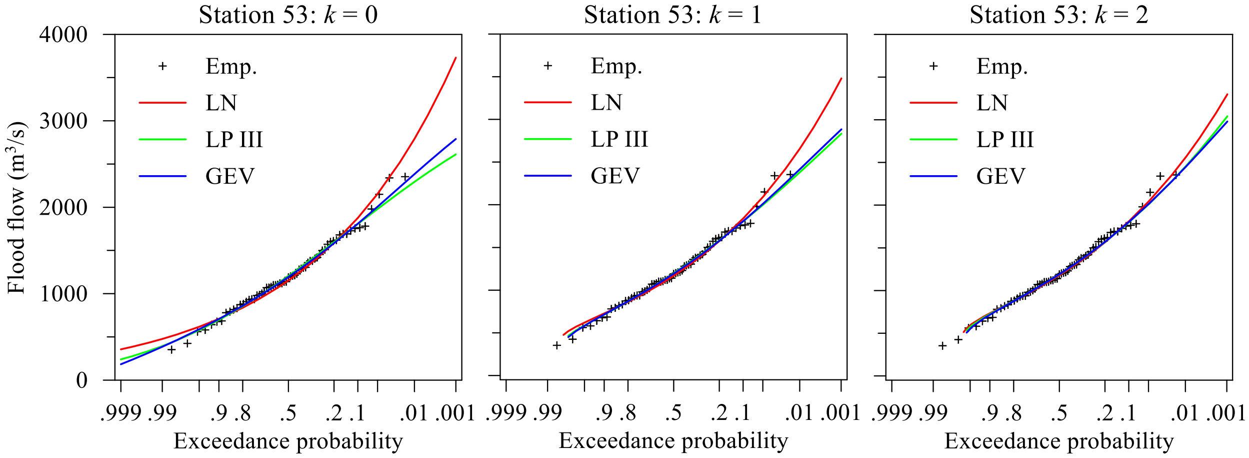

Figure 2. Effect of low outliers on distribution fitting: complete record (left), one outlier removed (middle) and two outliers removed (right).

Table 4. The 100-year and 500-year flood quantiles X_T after removing low outliers for station 53 (percentage in parentheses indicates change relative to the complete record results).

No. of outliers	X_{100} (m^3/s)			X_{500} (m^3/s)		
	LN	LP3	GEV	LN	LP3	GEV
$k = 0$	2792	2294	2383	3442	2529	2682
$k = 1$	2656 (-4.9%)	2380 (3.7%)	2415 (1.3%)	3225 (-6.3%)	2704 (6.9%)	2753 (2.6%)
$k = 2$	2556 (-8.5%)	2451 (6.8%)	2445 (2.6%)	3066 (-10.9%)	2862 (13.2%)	2824 (5.3%)

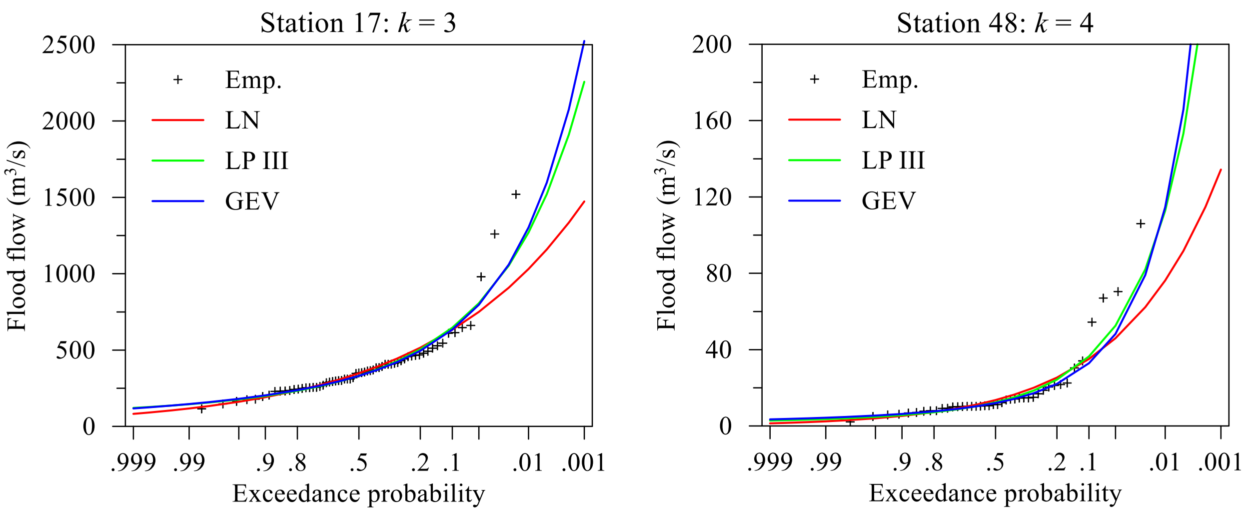

Figure 3. Examples of data with high outliers where the multiple outliers were detected with the Tietjen-Moore test.

Table 2 lists the series with identified high outliers and it indicates again that all the applied methods identify the same series with outliers. All the stations in this group have positive log-sample skew. The Bulletin 17B procedure and the sequential application of the Grubbs-Beck test yield the same results of one high outlier per series, while the Tietjen-Moore test detects a larger group of outliers in half of the series. While this may be the consequence of the swamping effect, at the same time a greater number of outliers may indicate presence of mixed distributions in the sample, for which an additional analysis should be made. Two examples are shown in Figure 3.

Only two series exhibited both high and low outliers, as shown in Table 3. The results obtained for these series are similar to those with individual low and high outliers. All tests detected the same number of low outliers while the Tietjen-Moore test again indicates that more than one high outlier is present. After removing the low outlier from each series, the same number of high outliers remains in the series. The effects of removing the low outliers from these two series are the same as already described for series exhibiting only low outliers (see Figure 4 for an example). The LP3 distribution has the greatest sensitivity to low outliers. For example, the 100-year and 500-year LP3 quantiles without low outlier are greater for 16% and 36% respectively. For the GEV distribution, these percentages are 2% and 6%, thus indicating that the GEV distribution is almost insensitive to low outliers.

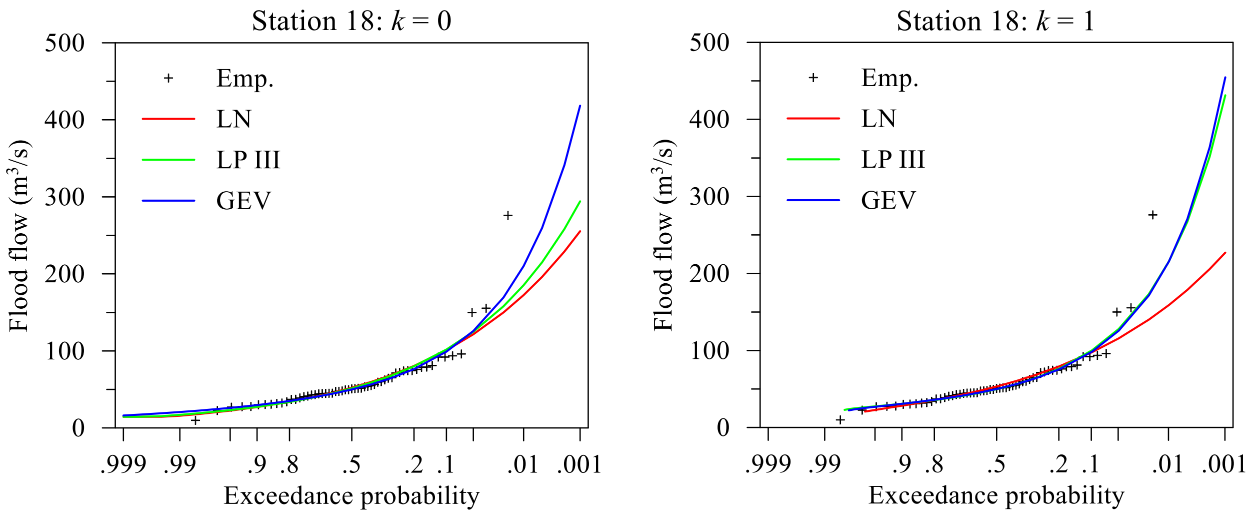

Figure 4. An example of data with both low and high outliers (station 18): complete record (left) and one outlier removed (right).

5. Conclusions

The results of three different tests for outlier detection were compared under the hypothesis that the annual maximum flood data are log-normal. The tests identified the same series with outliers, but the number of outliers differs from test to test. The tests for multiple outliers such as the Tietjen and Moore (1972) test or the new test proposed by Cohn et al. (2013) are useful not only for multiple low outliers, but can also be useful for high outliers in order to indicate mixed distributions.

Removal of low outliers from the sample has a significant impact on the resulting flood frequency and quantiles. The log-Pearson type III and the log-normal distributions are very sensitive to presence of low outliers. On the contrary, the general extreme value distribution has a very low sensitivity to low outliers and can therefore be recommended for a robust flood frequency analysis. On the other hand, impact of low outliers on the upper tail of the distribution can be avoided with the peaks over threshold (POT) approach. This approach has an additional benefit of including more values above the threshold so that the high outliers from the annual maximum series are not exceptional values in the POT series. The outliers should also be investigated in the regional context flood frequency analysis, which is considered to be more robust than at-site analysis.

The analysis of the series used in this study has shown that the tests are obviously designed to identify and remove data that violate normality. While this can be beneficial when applied to data at the lower tail, it cannot be reasonably applied to data at the upper tail. Therefore, these methods are not appropriate for flood data that cannot be assumed to be log-normally distributed (or normal after some other transformation). Such cases call for application of tests with an assumption of other underlying distributions. Further research is needed in this direction, since a range of tests for non-normal samples are available in the literature (Barnett and Lewis, 1994).

However, it has been again shown that the problem of outliers is related to many other problems such as the lack of knowledge on parent distributions, non-homogeneity of the data or presence of mixed distributions, and relatively short records from which these distributions are difficult to identify. Therefore, it is impossible to recommend a single and a straightforward procedure for flood frequency analysis. The practising hydrologists still have to deal carefully with each particular series and each outstanding value in these series, and to strive to apply the more robust approaches like the POT method or regional analysis.

ACKNOWLEDGEMENTS

The study in this paper was undertaken in the framework of the “The study on development and improvement of the protection from floods in Serbia: Development of the methodology for standardization of flood assessment procedures in Serbia – Phase One: Guidelines for frequency analysis of floods at hydrologic stations (gauged basins)” funded by the Public Water Management Company “Srbijavode”. The authors are grateful to the Republic Hydrometeorological Service of Serbia for the making the data available and for participating in this study.

REFERENCES

Barnett, V., and T. Lewis (1994). *Outliers in Statistical Data*, John Wiley, New York.

Blagojević, B., Mihailović, V., Plavšić, J. (2013). New Guidelines for Flood Flow Assessment at Hydrologic Stations in Serbia. Electronic Proceedings of the International Conference on Flood Resilience: Experiences in Asia and Europe, 5-7 September 2013, Exeter, United Kingdom. Djordjević, S., Butler, D., Chen, A. (Eds.).

Blagojević, B., Mihailović, V., Plavšić, J. (2014). Outlier treatment in the flood flow statistical analysis. Proc. Int. Conf. on Contemporary Achievements in Civil Eng., Facilty of Civil Engineeing Subotica, University of Novi Sad, Subotica, Serbia, 24-25 April 2014, pp. 603-609, DOI: 10.14415/konferencijaGFS2014.081.

BLFUW (2011). Leitfaden Verfahren zur Abschätzung von Hochwasserkennwerten (Guidelines for Flood Flow Assessment), P. Lorenz (ed.), Bundesministerium für Land- und Forstwirtschaft, Umwelt und Wasserwirtschaft, Wien, Austria. Available at http://www.wasseraktiv.at/

resources/files/2012/3/22/2192/leitfaden-verfahren-zur-abschaetzung-von-hwkennwerten.pdf (accessed 23 June 2014).

Cohn, T. A., J. F. England, C. E. Berenbrock, R. R. Mason, J. R. Stedinger, and J. R. Lamontagne (2013). A generalized Grubbs-Beck test statistic for detecting multiple potentially influential low outliers in flood series, *Water Resour. Res.*, 49, 5047–5058, doi:10.1002/wrcr.20392.

Ghosh, D. and Vogt, A. (2012). Outliers: An Evaluation of Methodologies. *Proceedings of the Survey Research Methods Section*, American Statistical Association. Available at http://www.amstat.org/sections/srms/Proceedings/y2012/Files/304068_72402.pdf (accessed 15 June 2014).

Grubbs, F.E. (1950). Sample criteria for testing outlying observations. *Ann. Math. Stat.*, 21(1), 27-58.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples, *Technometrics*, 11(1), 1–21.

Grubbs, F. E., and G. Beck (1972). Extension of sample sizes and percentage points for significance tests of outlying observations, *Technometrics*, 14(4), 847–854.

Helsel, D.R. and Hirsch, R.M. (2002). *Statistical Methods in Water Resources*, US Geological Survey, available at http://water.usgs.gov/pubs/twri/twri4a3/ (accessed 23 June 2014).

Hodge, V.J., and Austin, J. (2004). A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, 22(2), 85-126.

IACWD (1982). Guidelines for determining flood flow frequency: Bulletin 17B (revised and corrected), Interagency Advisory Committee on Water Data, Hydrol. Subcomm., Washington, D.C.

Kottegoda, N., Rosso R. (2008). *Applied Statistics for Civil and Environmental Engineers*, Blackwell, 2008.

McCuen, R.H. (2002). Outlier Detection. Chapter 4 in *Modelling Hydrologic Change: Statistical Methods*, Lewis Publishers, CRC Press.

NRCS (2012). National Engineering Handbook, Part 630: Hydrology, Chapter 18: Selected Statistical Methods, National Resources Conservation Service, USDA.

Rosner, B. (1975). On the detection of many outliers, *Technometrics*, 17(2), 221–227.

Spencer, C.S., and McCuen, R.H. (1996) Detection of Outliers in Pearson type III data. *J. Hydrol. Eng.*, 1(1), 2-10.

Stedinger J. R. i Griffis V W. (2008). Flood Frequency Analysis in the United States: Time to Update. *Journal of Hydrologic Engineering*, 13(4), 199–204.

Stedinger J. R. i Griffis V W. (2011). Getting from here to where? Flood frequency analysis and climate. *Journal of the American Water Resources Association* (JAWRA), 47(3), 506–513.

Stedinger, J.R., Vogel, R.M., Foufoula-Georgiou, E. (1993). Frequency Analysis of Extreme Events. In: *Handbook of Applied Hydrology*, Mc-Graw Hill, D. Maidment (ed.).

Tietjen, G., and R. Moore (1972). Some Grubbs-type statistics for the detection of several outliers, *Technometrics*, 14(3), 583–597.

Verma, S.P., and Quiroz-Ruiz, A. (2006). Critical values for 22 discordancy test variants for outliers in normal samples up to sizes 100, and applications in science and engineering. *Revista Mexicana de Ciencias Geológicas*, 23(3), 302-319.

Vukmirović, V., Pavlović, D. (2000). Utvrđivanje kriterijuma za izbor merodavnih velikih voda (Establishing criteria for defining design flood flows). Report of the national research project TSI 114 „Contemporary methods in water resources", Faculty of Civil Engineering, University of Belgrade.