27| Оригинални научни рад
Research paper
doi 10.7251/STP1813386K ISSN 2566-4484

NEIZVJESNOST U ODREĐIVANJU VELIKIH VODA U PRAKSI NA PRIMJERU SREMSKE MITROVICE, RIJEKA SAVA

Nina Kondić, *nina92.bl@hotmail.com*,
University of Banja Luka, Faculty of Architecure, Civil Engineering and Geodesy
Zana Topalović, *zana.topalovic@aggf.unibl.org*,
University of Banja Luka, Faculty of Architecure, Civil Engineering and Geodesy

Apstrakt

Određivanje velikih voda različitog ranga pojave je jedan od najvažnijih zadataka koji se postavlja pred inženjere hidrotehnike. Projektovanje hidrotehničkih objekata i sistema zahtijeva pouzdane ocjene kvantila što nije uvijek jednostavno isporučiti. U ovom radu prikazuje se neizvjesnost određivanja velikih voda uobičajena u praktičnoj primjeni a na primjeru rijeke Save, na stanici Sremska Mitrovica. Na nizu od 42 godine osmotrenih dnevnih proticaja izvršena je statistička analiza gdje je uzorkovanje sprovedeno prema dvije metode: metodi godišnjih maksimuma i metodi pikova. Pri ocjenjivanju neizvjesnosti, uzorci su analizirani prema nekoliko scenarija, tj. mijenjana je dužina uzorka kao i bazna vrijednost oticaja kod metode pikova. Dobijeni rezultati ukazuju na veliku neizvjesnost sračunatih kvantila, posebno u domenu malih vjerovatnoća pojave, a time i na problem usvajanja kvantila u praksi koji obezbjeđuje sigurnost i optimalnu cijenu objekta.

Ključne riječi: neizvjesnost, statistička analiza, velike vode, funkcija raspodjele, kvantili

FLOOD FREQUENCY ESTIMATION UNCERTAINTY IN DESIGN PRACTICE: CASE STUDY OF SREMSKA MITROVICA, SAVA RIVER

Abstract:

Flood frequency estimation is one of the most important tasks for hydraulic engineers. Design of hydraulic structures and systems require reliable estimates of high waters, which is not always easy to deliver. In this paper, uncertainty of flood frequency estimation common in practical use is presented in the case study of the Sava River at the Sremska Mitrovica hydrological station. Time series of 42 years daily flow records are statistically analysed on two samples, comprising annual maxima (AM), and peaks over threshold (POT). For uncertainty assessment, samples are analysed for several different scenarios, i.e. varying AM sample length as well as threshold flow for POT. Results indicate large uncertainty of flood frequency estimates, especially in the domen of low probabilities, as well as problem of adopting final value for practical use that will provide safety and optimum cost of the structures.

Keywords: uncertainty, statiatical analysis, flood frequency, distribution function, quantiles

1. INTRODUCTION

Flood frequency analysis imparts flood frequency estimates (FFE) that play an important role in the design of almost all hydraulic structures and systems such as dykes, bypass channels, bridges, floodwalls, spillways, culverts, etc. Safety of these structures, as well as human lives in the cases of large flood protection systems depend upon the reliability of FFE. On the other hand, estimated design flood must be economically justified, wherefore each country has set the standard design flood return periods for different structure types (e.g. 100 years for dykes, 5-10 years for storm drainage system, 1000 years for a concrete dam spillway, 10000 for earth dam spillway, etc.).

There are three possible methods for design flood estimation, depending on the available data: (a) statistical analysis of the observed flows, (b) statistical analysis of the modelled flow, obtained with hydrological rainfall-runoff model and observed precipitation and (c) by transformation of design storms (obtained from statistical analysis of observed precipitations) into design flows based on a rainfall-runoff model. Here, statistical analysis is assumed to be establishing relationship between the flows and return period or probability of (non-)exceedance with one of the defined theoretical probability distribution functions. In this text from now on, the flow defined in this way is called quantile.

FFE is usually obtained from two sampling methods [1]: annual maxima (AM) and peaks over threshold (POT). Research shows that results of these two methods are quite similar above return period of 10 years [2] or POT is found to be advantageous over AM due to possibility to include more information about floods, i.e. more floods per year instead of only one as in AM method [3], [4].

During FFE in design practice, the uncertainty of the estimated quantiles is very rarely included in the calculation. Uncertainties in hydrological procedures can be classified into three categories [5]: natural or inherent, model and parameter uncertainty. Natural or inherited uncertainty arises from the random variability of hydrological variables (i.e. uncertainty of measured data) while model uncertainties arise from the model structure and approximations made when representing hydrological phenomena. Parameter uncertainty is due to unknown nature, and therefore errors compiled in the methods of parameter estimation. In design practice, usually just one uncertainty source is addressed through confidence intervals estimation. These intervals only deal with data sample uncertainty from which quantile is estimated [6].

Investors, designers and managers usually think that quantile estimated by hydrologist is an exact value while those values, depending on the available data and methodology applied can be found in a very wide range [7]. For this reason, the aim of this paper is to show the possible quantiles range depending on data sample method. This is demonstrated on the Sava River case study, the Sremska Mitrovica hydrological station that records runoff from almost all of the Sava River basin. Similar study [8] shows a wide range of runoff from relatively small catchment, while here, the idea is to see how this range in changed when dealing with large rivers data.

2. METHODOLOGY

2.1. Case study information and data

The Sava River is right and by the discharge largest tributary of Danube, the second longest river in Europe (after Volga River). The river basin area is over 97000 km2 with the watercourse length of cca 990km. The river is formed in Slovenia from the Sava

Dolinka and Sava Bohinjka from where it flows through Croatia and Bosnia and Herzegovina, discharging into the Danube in Belgrade, Serbia. A small share of the river basin is located in Montenegro (7.09%) and in Albania (0.18%).

The Sremska Mitrovica station is located 139.3 km from the Sava river mouth into the Danube. Station controls 87.996 km2 of the basin. The station is founded in 1878 but flow observation commenced not before 1926. Observed flow data available for this paper are for the period 1969-2010.

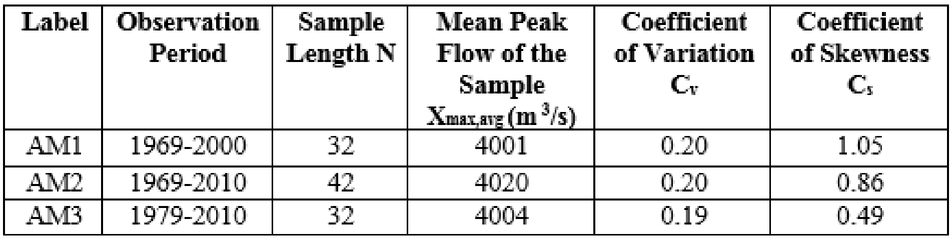
2.2. Sampling data for statistical analysis

In this paper time series of 42 years daily flow observations are used, which are statistically processed with two sampling methods: annual maxima (AM) and peak over the threshold (POT).

For AM method, samples were formed taking one maximum flow per each year. Because of the relatively short available observation period (1969-2010), here are used three samples with different observation period and different sample length: 1969-2010, 1969-2000 (this way excluding new data after 2000) and 1979-2010.

For each sample, the following theoretical probability distributions are analyzed: log-normal (logarithmic Gaussian distribution), Pearson type III, log-Pearson type III, Gumbel and Generalized Extreme Value (GEV). Review of the fitness of theoretical and empirical distributions is conducted with following tests [9], [10]: Kolmogorov–Smirnov, Anderson-Darling and Cramer-von-Mises. Based on the results of the named tests, log-Pearson type III is accepted as the most adjustable theoretical distribution to the observed data. The statistics of the samples, as well as the parameters of the theoretical distributions, are estimated by the method of moments.

Table 1. Overview of samples analyzed with annual maximum flows method



Label	Observation Period	Sample Length N	Mean Peak Flow of the Sample $X_{max,avg}$ (m ^{3}/s)	Coefficient of Variation C_v	Coefficient of Skewness C_s
AM1	1969-2000	32	4001	0.20	1.05
AM2	1969-2010	42	4020	0.20	0.86
AM3	1979-2010	32	4004	0.19	0.49

Another method used to determine the maximum flow quantiles is POT method, in which samples were formed by taking maximum flows above a certain limit - the threshold X_b. Eight different thresholds are defined from which eight samples within fixed observation period (1969-2010) are extracted. The data samples of maximum flows, shown in the Table 2 are adjusted to the two-parameter Weibull distribution function.

Peak selection for POT method is done by considering independence criteria. General criterion is to use only one peak in one event, while peaks filtration when two (or more) consecutive peaks occur is processed following Water resources Council (USWRC 1982) criteria [16]:

$$\theta < 5\ \text{days} + \log(\text{A})$$
$$\text{Or } \text{x}_{\text{MIN}} < 0.75\text{min}[\text{x}_{\text{s1}}, \text{x}_{\text{s2}}] \qquad (1)$$

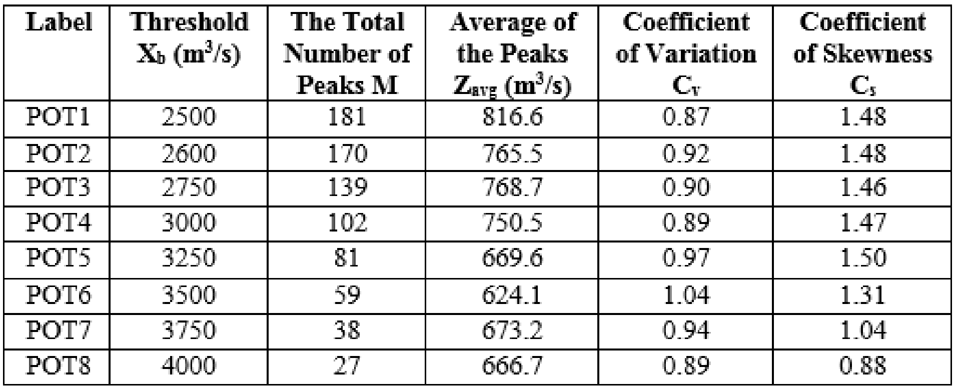
where

A is basin area in miles,

xs1 and xs2 are two consecutive peak values.

The second peak should be rejected if one of the conditions in above equation is met.

Table 2. Overview of samples analyzed with a peak over threshold method



Label	Threshold X_b (m^3/s)	The Total Number of Peaks M	Average of the Peaks Z_{avg} (m^3/s)	Coefficient of Variation C_v	Coefficient of Skewness C_s
POT1	2500	181	816.6	0.87	1.48
POT2	2600	170	765.5	0.92	1.48
POT3	2750	139	768.7	0.90	1.46
POT4	3000	102	750.5	0.89	1.47
POT5	3250	81	669.6	0.97	1.50
POT6	3500	59	624.1	1.04	1.31
POT7	3750	38	673.2	0.94	1.04
POT8	4000	27	666.7	0.89	0.88

By using these methods for flood frequency estimation, eleven theoretical distribution functions have been obtained. For each of them, confidence intervals are calculated and then their envelopes (aggregate minimum and maximum limit values for each return period) are defined. These envelopes represent uncertainty intervals for quantiles of specific return period, which is the main aim of this paper. Analysed return periods (T) are: 2, 5, 10, 20, 50, 100, 200, 500, 1000 and 10.000 years.

Basic theories of applied methodology is given in the next section.

3. STATISTICAL MODELS FOR FLOOD FREQUENCY ESTIMATION

3.1. Annual maxima (AM) method

Annual maxima (AM) is a statistical model where sample data formed with one maximum flow per year is fitted to a probability function. Therefore, the sample consists of $Q_1, Q_2, ..., Q_N$ data where N is a number of years of observations and Q_i is a maximum flow recorded for i -th calendar year. The basic assumption is that all data in the sample are stochastic value, mutually independent, uncorrelated and homogenous; therefore adequate sample of the population which must be previously tested [1], [11].

There are many probability functions that can be used for fitting AM. In practice, several distributions are fitted and upon test of fitness (i.e. Kolmogorov-Smirnov, Cramer von Mises, etc. [12]) is decided which distribution fits the best to the observed data.

The three parameter log-Pearson type III distribution from the gamma group best describes the flow distribution Q_i for annual maxima series of Sremska Mitrovica station.

The function of this distribution is defined by the expression:

$$F(y) = \int_{\Upsilon}^{y} \frac{1}{\beta\Gamma(\alpha)} \left(\frac{y-\Upsilon}{\beta}\right)^{\alpha-1} exp\left(-\frac{y-\Upsilon}{\beta}\right) dy \tag{2}$$

where:

α (α >0), β (β≠0), Υ are distribution parameters,

$y = logx$,
$x = Q$,
$\Gamma(\alpha)$ is gamma function:

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-u} du. \tag{3}$$

Since the inverse distribution (the quantile y calculation) cannot be defined explicitly by the expression, numerical approximations (in combination with built-in excel functions) are used. The inverse distribution could be determined via tabulated factor of frequency K_T (which depends on the probability and coefficient of skewness of the series) as well [13]. In that case, the following expression is used:

$$\log(x) = y_{avg} + K_T S_y \tag{4}$$

where:
y_{avg} is mean value,
S_y is standard deviation of the log series.

3.2. Peaks over threshold (POT) method

POT represents good alternative to the AM method since it includes more flood information, i.e. more than one peak in the year [3]. Potential problem here may arise due to uncertainty of threshold determination. Threshold X_b defines the sample data that consists of peaks $Z = x - X_b$, where x is observed flow. There are several recommendations for threshold determination, such as to use minimum annual maxima as a threshold or to find the value on the graphic representation of the relationship X_b and X_b / Z_{avg}. In the later, Z_{avg} is an average of the peaks over threshold and the flow where the distinctive linearity is lost, represents X_b [14].

The theoretical distribution of the number of peaks per year is calculated according to one of the distribution functions for discrete variables. Selection criteria for theoretical distribution is the value of the dispersion index, obtained from:

$$I = \frac{N \cdot S_x^2}{M} = \frac{S_x^2}{n_{avg}} \tag{5}$$

where:
N is the total number of the observation years,
M is the total number of peaks above the threshold X_b,
S_x^2 is the variance of the number of peaks,
$n_{avg} = \frac{M}{N}$ is the mean of the number of peaks.

As the dispersion index for all series of peaks above the threshold X_b formed from observed data for Sremska Mitrovica station is greater than one, negative binomial distribution with two parameters is used. The function of this distribution is defined by the expression:

$$p_i = P\{X = i\} = \binom{b-1+i}{b-1} p^b (1-p)^i \tag{6}$$

where:

b is distribution parameter:

$$b = \frac{n_{avg}}{I-1} \tag{7}$$

p is distribution parameter:

$$p = \frac{b}{n_{avg} + b} \tag{8}$$

The distribution of the height $H(Z)$of the peaks Z is defined with a two-parameter Weibull distribution. The function of this distribution is defined by the expression:

$$H(Z) = 1 - exp\left\{-\left(\frac{Z}{\alpha}\right)^{\beta}\right\} \tag{9}$$

where:

β is distribution parameter that should be numerically calculated from the expression:

$$f(\beta) = \frac{\Gamma(1 + 2/\beta)}{\Gamma^2(1 + 1/\beta)} = 1 + C_{vz}^2 \tag{10}$$

α is distribution parameter:

$$\alpha = \frac{Z_{avg}}{\Gamma(1 + 1/\beta)} \tag{11}$$

The distribution of the probability of the annual maxima $F(x)$ is obtained by a combination of the previous distributions of the number and height of the peaks, in a way that it gives an inverse distribution function for quantile calculation according to Weibull's distribution:

$$Z = X_b + \alpha[-ln(1 - H)]^{1/\beta} \tag{12}$$

where:

$1 - H$ is obtained from the negative binomial distribution:

$$1 - H = \frac{F^{-1/b} - 1}{1/p - 1} \tag{13}$$

3.3. Standard quantile errors and confidence intervals

The confidence intervals of the distribution function are determined for quantiles X(T), where T denotes the return period, by defining the upper (u) and lower (l) interval boundaries according to the following:

$$X_{u,l}(T) = X(T) \pm |z_\alpha| S_{x(T)} \tag{14}$$

where:

z_α is a standardized variable of normal distribution for the significance threshold α:

$$z_\alpha = -z(1-\alpha) \tag{15}$$

The significance threshold α corresponds to the confidence interval β:

$$\beta = 1 - 2\alpha \tag{16}$$

$S_{x(T)}$ is the standard quantile error.

The standard quantile error $S_{x(T)}$ represents the square root of the quantile variance and is determined differently depending on the selected theoretical distribution.

For log-Pearson type III distribution, the standard quantile error is defined by the expression:

$$S_{y(T)} = \frac{S_y}{\sqrt{N}} \sqrt{1 + \frac{1}{2} K^2(T)} \tag{17}$$

where:

$K(T)$ is frequency factor:

$$K(T) = \frac{Y(T) - y_{avg}}{S_y} \tag{18}$$

while $Y(T)$ represents logarithmic quantile:

$$Y(T) = logX(T) \tag{19}$$

The determination of the standard quantile errors and confidence intervals for Weibull distribution is somewhat complicated and due to limited space is not given here. However, interested readers may find complete derived expressions in [15].

4. RESULTS AND DISCUSSION

Resulting quantiles for AM method are not in big discrepancy along the different return periods. The lowest quantiles are obtained for the period 1979-2010. The reason for this probably lies in the fact that largest floods on the Sava River recorded at this station occurred in 1970 and 1974, which are excluded from this period. Largest quantiles are obtained for the period 1969-2000, but not with significant differences from the whole period (1969-2010), which is shown in the Table 3.

Table 3. Overview of quantiles obtained from annual maximum flows method

Return period	Quantiles		
	1969-2000	1969-2010	1979-2010
2	3850	3885	3919
5	4580	4616	4591
10	5074	5097	5001
20	5558	5559	5376
50	6202	6161	5840
100	6700	6620	6178
200	7212	7085	6509
500	7917	7715	6940
1000	8474	8206	7263
10000	10496	9948	8337

These leads to conclusion that flood information content within the observation period used for statistical analysis is crucial for quantile estimation. Analysing data where large historical floods are excluded may lead to underestimated quantiles, and excluding periods with less and smaller floods may lead to quantile overestimation. Generally, sufficient length of the sample is maybe the most important in the proper statistical analysis. In this paper, there was not enough data to manipulate with, i.e. to form series of various length, but similar analysis concluded that, for example, parameters of the Pearson III distribution function are getting stabilized as length of the sample gets longer [8]. The problem of short sample data for FFE is not novelty in hydrology. Estimation of quantile of 100 years return period with the data sample of e,g, 40 years is never a good idea. However, in design practice this problem is usual and needs to be addressed with proper inclusion of uncertainty into the FFE.

In the POT method, largest quantiles are obtained with the threshold of 3500m3/s. Fitness of the theoretical and empirical distributions for both methods and for the same observation period is given in the Figure 1.

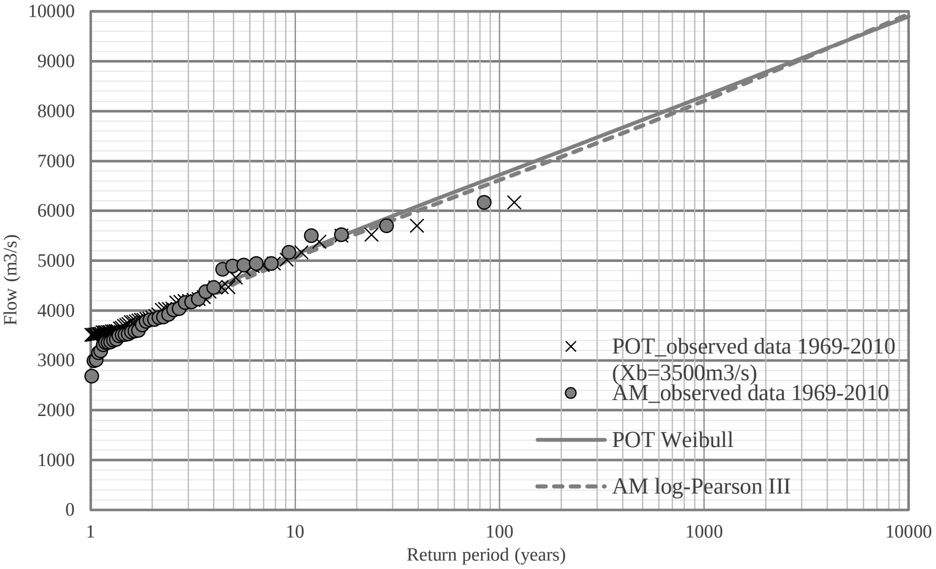


Figure 1. Distribution function plot for AM and POT methods for period 1969-2010

Summary results of all samples by both methods are depicted in the Figure 2. Quantiles obtained by POT are symbolised with dots while the ones obtained with AM are symbolised as squares. Confidence interval envelopes (i.e. boundary confidence intervals for both sampling methods) form upper and lower limits and are shown in the Figure 2 as continuous lines. The value of the significance threshold α used to determine confidence intervals is 0.05.

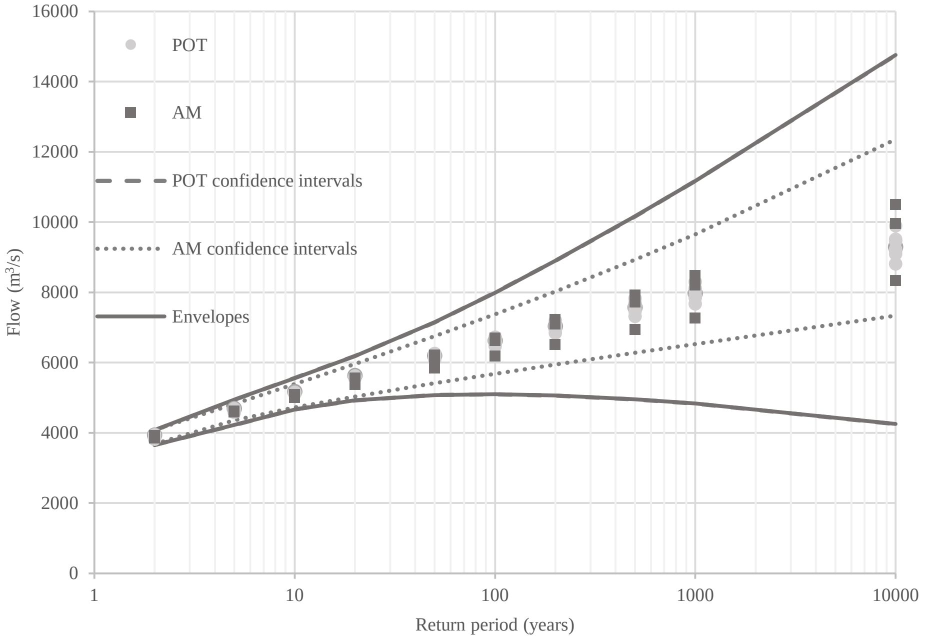


Figure 2. Distribution functions with boundary confidence intervals

For both methods, it is evident that range of quantiles increases as the return periods increases. For instance, defined quantile for a return period of 100 years, for POT method could take any value between 6470 and 6724 m^3/s, while for AM method between 6178 and 6700 m^3/s, depending on the available sample. These envelopes give wide range of possible quantile values, and by increasing return periods, uncertainties in flood frequency estimation increases as well.

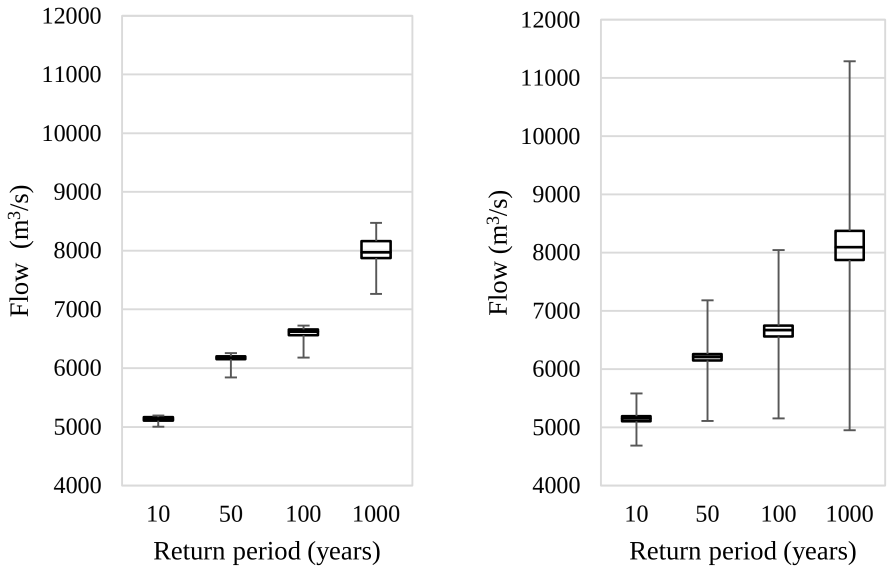


Figure 3. Box and whiskers plot for quantiles determined by AM and POT method (left) and with included confidence intervals (right)

Box and whiskers diagrams show the central value (median) of the series formed from obtained quantiles for both methods and its upper (75%) and lower (25%) quartiles, as well as the lowest and highest values of the quantiles. In this graph, bottom and top of the box are the upper and lower quartiles, so the box spans the interquartile range. A horizontal line inside the box marks the median. The ends of the vertical lines are the minimum and maximum values of all the obtained quantiles.

Introducing only one uncertainty due to the mistake of distribution functions, which in this case are confidence intervals, the range is significantly increased (Figure 3 right). It could be concluded that this range would be even greater, if uncertainty is calculated for output data from the probability analysis of the maximum flows due to other errors (i.e. measurements, distribution function and parameter estimation, etc.).

5. CONCLUSIONS AND RECOMMENDATIONS

Uncertainties in flow estimation occur due to many sources of error: poor data quality, determination of parameters and selection of the statistical model-distribution function, the data sampling method, assumptions about stochastic nature of hydrological variable, problems with water level measurement, equipment malfunction or incorrect cease-to-flow datum, etc. [6]. The best way of improving the data quality of flood flow behavior is to measure rainfall and streamflow – preceding, during and after a flood event. The longer is the period of record, the better the confidence in the flow estimate would be.

The uncertainty of the sample from which information about the maximum flow of a certain probability of occurrence is finally obtained, is mainly considered through determination of the confidence intervals. In this paper, this is extended to uncertainty that depends on sample data used for estimation, statistical model used and method (AM and POT). Still, problem in design practice exists. For example, design flood of 100 years return period is estimated with mean value of 6577 m3/s. Possible quantiles that can be used in design is ±20% with reference to mean value, including confidence intervals and estimation methodology, which is large span due to large flow values. Comparing results to similar study [8], it could be concluded that great catchments are less sensitive when it comes to the range of quantiles. This range also vary, but in comparison with relatively small catchments, these changes are smaller.

The analysis could be improved: by collecting more observed data to extend analysis, by varying the length of the sample for the peak method, analysing some other sources of error, such as various statistical models and other methods for model parameters estimation (i.e. L-moments). However, it is expected that the future extended analysis will show even more uncertainty intervals, especially under the conditions of changing climate. Still, in practice there is no proper mechanism for solving the problem, i.e. decreasing the uncertainty or methods for dealing with it. This is the topic that should be seriously addressed in order to prevent future designs to be under- or over dimensioned with respect to design floods.

LITERATURE

[1] V. Vukmirović, Analiza verovatnoće pojave hidroloških veličina. Beograd: Građevinski fakultet, Naučna Knjiga, 1990.

[2] S. Mkhandi, A. Opere, and P. Willems, “Comparison between annual maximum and peaks over threshold models for flood frequency prediction,” Int. Conf. UNESCO Flanders FIT FRIEND/Nile Proj. - ’Towards a better Coop., vol. 1, pp. 1–15, 2005.

[3] S. Fischer and A. Schumann, “Comparison between classical annual maxima and Peak over Threshold approach concerning robustness,” Sonderforschungsbereich 823, 2014.

[4] H. Madsen, P. F. Rasmussen, and D. Rosbjerg, “Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events: 1. At site modeling,” Water Resour. Res., vol. 33, no. 4, pp. 747–757, 1997.

[5] V. Te Chow, D. R. Maidment, and L. W. Mays, “Applied hydrology.” McGraw Hill, New York, p. 565, 1988.

[6] B. Merz and A. H. Thieken, “Separating natural and epistemic uncertainty in flood frequency analysis,” J. Hydrol., vol. 309, no. 309, pp. 114–132, 2005.

[7] C. Wright and D. Kemp, “Flow Estimation for Flood Management Understanding the Uncertainty .,” in 5th Flood Management Conference Warrnambool, 2007, pp. 1–8.

[8] Ž. Topalović, “Praktični problemi određivanja mjerodavnih velikih voda za potrebe projektovanja sistema odbrane od poplava,” in Zbornik radova sa 17. Savetovanja SDHI i SDH održanog 5.-6- oktobra u Vršcu, 2015, pp. 893–903.

[9] N. T. Kottegoda and R. Rosso, Applied Statistics for Civil and Environmental Engineers, 2nd ed. West Sussex: Blackwell Publishing, 2008.

[10] D. A. Darling, “The Kolmogorov-Smirnov, Cramer-von Mises Tests,” Ann. Math. Stat., vol. 28, no. 4, pp. 823–838, 1957.

[11] S. Jovanović, Primena metoda matematičke statistike u hidrologiji. Beograd: Građevinski fakultet, 1977.

[12] T. B. Arnold and J. W. Emerson, “Nonparametric Goodness-of-Fit Tests for Discrete Null Distributions,” R J., pp. 34–39, 2011.

[13] J. D. Salas et al., “Introduction to hydrology,” in Handbook of Environmental Engineering 15: Modern water resources engineering, C. T. Yang and L. K. Wang, Eds. New York: Humana Press, Springer Science+Business, 2014, pp. 61–94.

[14] J. Beirlant, Y. Goegebeur, J. Teugels, J. Segers, D. De Waal, and C. Ferro, Statistics of Extremes: Theory and Applications, vol. 47, no. 3. West Sussex: John Wiley & Sons Ltd., 2005.

[15] J. H. Heo, J. D. Salas, and K. D. Kim, “Estimation of confidence intervals of quantiles for the Weibull distribution,” Stoch. Environ. Res. Risk Assess., vol. 15, no. 4, pp. 284–309, 2001.

[16] USWRC (United States. Interagency Advisory Committee on Water Data. Hydrology Subcommittee), “Guidelines for deter- mining flood flow frequency”. Reston, VA: US Department of the Interior, Geological Survey, Office of Water Data Coordination, 1982.