JAES | JOURNAL OF APPLIED ENGINEERING SCIENCE | www.engineeringscience.rs
ISTRAŽIVANJA I PROJEKTOVANJA ZA PRIVREDU

Indexed by

MISSING DATA REPRESENTATION BY PERCEPTIONTHRESHOLDS IN FLOOD FLOW FREQUENCY ASSESSMENT

Nikola Đokić
University of Niš,
Faculty of Civil Engineering and Architecture,
Niš, Serbia

Borislava Blagojević
University of Niš,
Faculty of Civil Engineering and Architecture,
Niš, Serbia

Vladislava Mihailović
University of Belgrade,
Faculty of Forestry,
Belgrade, Serbia

Key words: *flood flow frequency analysis, bulletin 17C, perception thresholds, missing data, HS Senta*

Cite article:
Nikola, Đ., Borislava, B., & Vladislava, M. [2021]. Missing data representation by perceptionthresholds in flood flow frequency assessment. *Journal of Applied Engineering Science*, 19(2), 432 - 438. DOI:10.5937/jaes0-28902

Online aceess *of full paper is available at: www.engineeringscience.rs/browse-issues*

iipp Original Scientific Paper

Journal of Applied Engineering Science
Istrazivanja i projektovanja za privredu

doi:10.5937/jaes0-28902 Paper number: 19(2021)2, 810, 432-438

MISSING DATA REPRESENTATION BY PERCEPTION THRESHOLDS IN FLOOD FLOW FREQUENCY ASSESSMENT

Nikola Đokić[1*], Borislava Blagojević[1], Vladislava Mihailović[2]

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

Flood flow frequency analysis (FFA) plays one of the key roles in many fields of hydraulic engineering and water resources management. The reliability of FFA results depends on many factors, an obvious one being the reliability of the input data - datasets of the annual peak flow. In practice, however, engineers often encounter the problem of incomplete datasets (missing data, data gaps and/or broken records) which increases the uncertainty of FFA results. In this paper, we perform at-site focused analysis, and we use a complete dataset of annual peak flows from 1931 to 2016 at the hydrologic station Senta of the Tisa (Tisza) river as the reference dataset. From this original dataset we remove some data and thus we obtain 15 new datasets with one continuous gap of different length and/or location. Each dataset we further subject to FFA by using the USACE HEC-SSP Bulletin 17C analysis, where we apply perception thresholds for missing data representation. We vary perception threshold lower bound for all missing flows in one dataset, so that we create 56 variants of the input HEC-SSP datasets. The flood flow quantiles assessed from the datasets with missing data and different perception thresholds we evaluate by two uncertainty measures. The results indicate acceptable flood quantile estimates are obtained, even for larger return periods, by setting a lower perception threshold bound at the value of the highest peak flow in the available - incomplete dataset.

Key words: flood flow frequency analysis, bulletin 17C, perception thresholds, missing data, HS Senta

INTRODUCTION

Flood frequency analysis (FFA) is an important part of the flood risk management. The FFA result is a set of flood quantiles representing, for example, a 1000-year, 100-year and 50-year flood flow. In the case of gauged catchments, the most common FFA is performed on the datasets comprising annual flow maxima. These datasets often come with missing data, data gaps or broken records. There is a variety of techniques for dealing with missing values, from the simple ones where linear regression is used with the direct upstream or downstream hydrologic station (HS) record, to the sophisticated ones including dynamic regression models [1]. Some analyses concerning gaps in hydrometeorological time series, reveal that a very efficient gap filling of sporadic, single-value gaps, is achieved by the value obtained using only three values of the dataset: the one before, one after the missing value, and the sample mean [2].

The recent revision of the Guidelines for FFA in the U.S.A. [3] called Bulletin 17C (B17C), introduces the concept of Perception threshold that can be used for missing data representation. A proper setting of the perception threshold requires some idea about the missing flow(s), usually obtained from the public records, newspaper, interviews or modelling flood marks (physical evidence). However, practicing engineers often do not have resources for such an investigation. This situation is pronounced in a single-site analysis. Therefore, deriving flow information about the perception threshold from the available dataset (or its characteristics) would have a practical application.

The reference flow dataset in our research is the one gauged at the hydrologic station (HS) Senta of the Tisa (Tisza) river in the period from 1931 to 2016. We created 15 base datasets and 56 their variants with missing data of varying gap size and time of occurrence, and applied different perception thresholds in the FFA. When considering gap size for the investigation we also took care of the recommendations for maximum gap allowance in trend detection in extreme streamflow time series [4]. The flood flow quantiles assessed from the datasets with missing data and different perception thresholds we evaluated through percentage error (PE), and confidence interval width as uncertainty measure, while we also observed change in number of detected outliers in the datasets.

METHODOLOGY

In this research, we use the FFA methodology of B17C, successor of the famous B17B [5]. The latter has been in power over 30 years, while the former is a recent product of various situations that have emerged in practice [3]. There is a continuity of underlying theoretical flood frequency distribution in the B17C, which is the log-Pearson Type III (LPTIII) distribution. The Expected Moments Algorithm (EMA) is an analysis introduced in the B17C methodology for estimating the moments of the LPTIII distribution [3]. The EMA deals with multiple threshold data, which is a data representation option for missing data, incorporated in the Hydraulic Engineering Centre Statistical Software Package (HEC-SSP) [6]. We use the HEC-SSP as a tool in our research.

Peak flow datasets in our research involve systematic flood data – the observed (gauged) peak flows repre-

*nikolavdjokic995@gmail.com

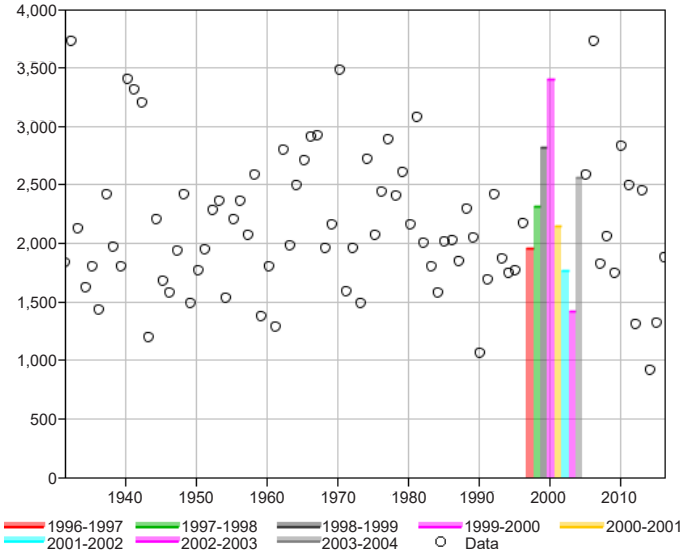

Figure 1: The data representation in HEC-SSP for B17C analysis of the dataset with perception thresholds applied in the created missing data period 1996-2004

sented by dots in Fig. 1, and missing data represented by perception thresholds. According to this new concept, a perception threshold is set to indicate that missing data is smaller than some flow value, i.e., the floods are not gauged above a certain flow. Perception thresholds are described by lower ($T_{Y,lower}$) and upper bound ($T_{Y,upper}$), where upper bound in the systematic records is assumed to be infinite. The lower bound should represent the smallest annual peak flow that could have been recorded. It is noted that '*the perception thresholds do not depend on the actual peak discharges that have occurred*' [3].

There are two more significant novelties in B17C for our research: improvement of the outlier detection procedure and different formula for the frequency distribution (quantile) confidence interval.

A new Multiple Grubbs-Beck outlier statistical test in B17C is a replacement for the simple Grubbs-Beck test of B17B, and as such, it is incorporated in the HEC-SSP. It is used in the outlier detection procedure, and in the outlier treatment, censoring for low outliers is applied [7]. Also, a multiple-threshold plotting positions according to Hirsch and Stedinger [8] is a default option in the HEC-SSP Bulletin 17C analysis. The novelty in confidence interval formulae by EMA reflects opportunity to include all available data, among which, potentially influential low floods and uncertainty in the at-site estimate of the skewness coefficients is significant for our research.

EXPERIMENTAL DATASETS

The Tisa (Tisza) River joins the Danube in Serbia. Along its 164 km course through Serbia, there is one flow gauge at Senta, in addition to 3 stage gages (Fig. 2). The river drainage basin area at HS Senta is 141715 km^2 [9]. The peak flow record comprises annual maxima starting from

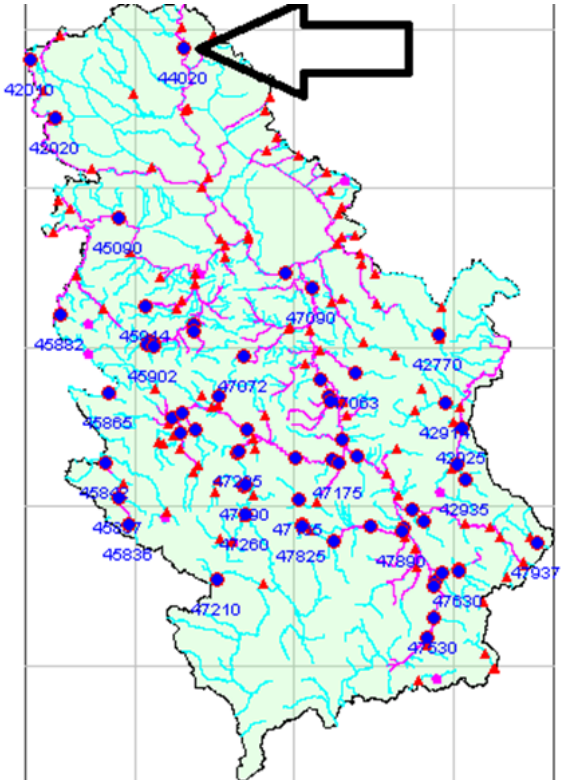

Figure 2: The map of rivers and some water gauges in Serbia generated in HEC-SSP. The arrow points to HS Senta

the year 1931 without missing data. The record length and data completeness are the main reasons for selecting this HS for studying. Our research covers the period from 1931 to 2016.

By the tests for homogeneity (*z*-test, *F*-test, Mann-Whitney test) and trend detection (Mann-Kendall trend test, Spearman's rank correlation coefficient, Regression slope test) at the 0.05 confidence level, we checked whether the reference set is suitable for statistical analysis. Pursuant to all of the tests, the 1931-2016 dataset of annual peak flows is suitable for statistical analysis. According to Multiple Grubbs-Beck test, neither low nor high outliers are present in the reference dataset. This reference dataset comprising gauged flows in the period of systematic record is labelled with number 1.

Labels of 15 base datasets with artificial gaps are shown in Table 1. Numbers 1-4 in the dataset labels denote length of the data gap in the reference data record in the following way: 1 – 1 year is missing, 2 – 5% of the dataset size is missing (4 years), 3 – 10% (8 years), 4 – 15% (12 years). Numbers 5-7 are labels for datasets with the missing data in a particular period as shown in Table 1, where dataset 5 misses 12% of the data, dataset 6, 17%, and dataset 7, 23%. Letters A-C show which data are missing: A – flows in the year of a minimum peak flow and around it – both before and after, B – in the year of a maximum peak flow and around it, C – flows in the year of a peak average flow and around it.

For the data representation in HEC-SSP we use infinity

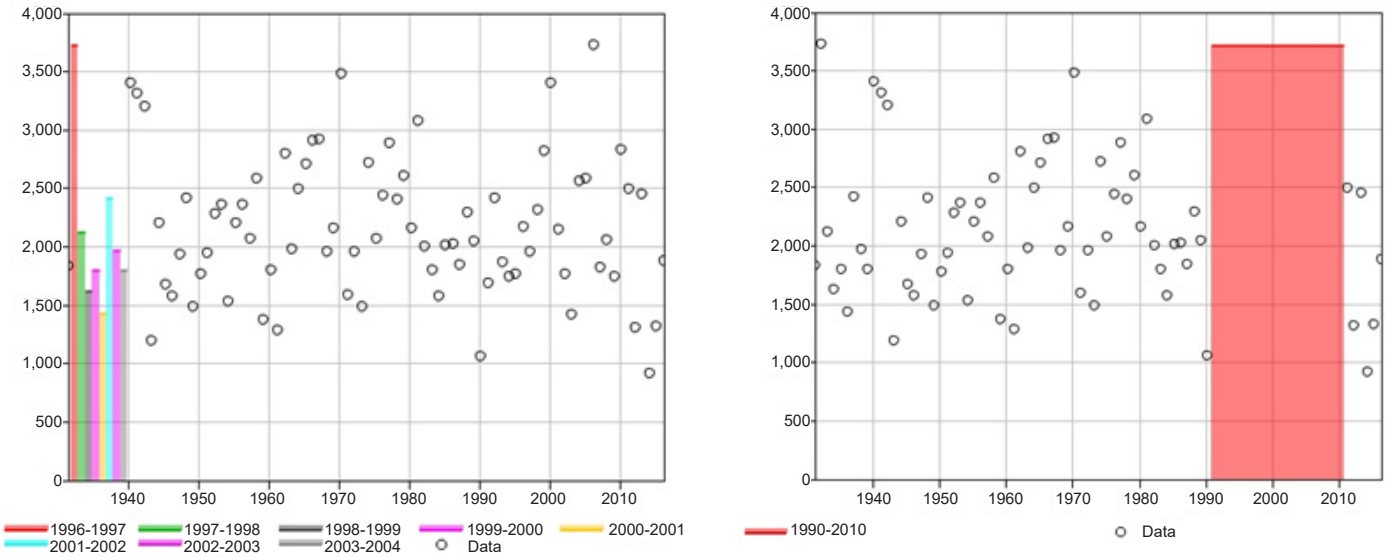

Figure 3: Data representation of the dataset 2B-iv (left) and the dataset.5-iii (right)

Table 1: Dataset labels depending on data gap size and reference dataset characteristics

Gap around Gap size / period	Min Q	Max Q	Avg Q
1 y.	1A	1B	1C
5%	2A	2B	2C
10%	3A	3B	3C
15%	4A	4B	4C
1991-2010	5		
1991-2005	6		
1991-2000	7		

for perception threshold upper bound ($T_{Y,upper}$) and different lower bounds for all missing flows in one dataset. Depending on the values set for $T_{Y,lower}$, datasets have an additional label (i-iv): i - $T_{Y,lower}$ is the average peak flow of the dataset remained after the gap is created, ii – $T_{Y,lower}$ is the maximum peak flow in the dataset remained after the gap is created, iii - $T_{Y,lower}$ is the maximum peak flow of missing data - where data gap is located, iv – $T_{Y,lower}$ are the single values of missing flows in each year within the gap. In Fig. 3 two examples of the dataset representation are shown.

There is a total of 56 datasets we subject to FFA, because some of the datasets have the same $T_{Y,lower}$ when its focus is on the flow we use for gap creation. The flood quantiles of our interest represent a 1000-, 500-, 200-, 100- and 50-year flood flow (*Q1000, Q500, Q200, Q100 and Q50*) i.e., 0.1%, 0.2%, 0.5%, 1.0% and 2.0% annual exceedance probability flow.

RESULTS

The results of FFA for the reference dataset are shown in Fig. 4 as screenshot of the HEC-SSP output table. Flood quantile estimates in this table (computed curve flows in cubic meters per second – cms) are considered reference values when calculating percentage error in comparison to flood quantile estimates from the other datasets. It can be seen that we use a width of 90% confidence interval to compare uncertainty in estimated quantiles (*QT*). Two probability plots, a graphic output from HEC-SSP, are shown in Fig. 5. Along with the percentage error of flood quantile estimates from the 15 base datasets (as listed in Table 1), the 90% confidence interval widths are shown in Fig. 6 for two quantiles (1000- and 100-year return period) and two $T_{Y,lower}$ options (i and ii).

The flood flow quantiles assessed from the datasets with missing data and different perception thresholds evaluated through percentage error (*PE*) are shown in Fig. 7.

DISCUSSION

The size of reference dataset allows for reliable probability curve extrapolation slightly beyond 400 - years return period (2 to 5 times the dataset size) [9]. Therefore, in the results section, a flood quantile of interest in many flood-related applications, *Q100* is selected to illustrate percentage errors and confidence interval widths. The quantile of the largest return period we studied, *Q1000*, a frequent focus of studies, projects and applications, is shown in parallel in Fig. 6. These results show the *PEs* in datasets A and B are larger than in datasets C, 5, 6 and 7. At the same time, uncertainty in 100- and 1000-year quantile estimates are generally higher in datasets A and for two options of lower threshold bound of 4B, while uncertainty is close to the reference one in the datasets C (Fig. 6 – right column). The only dataset where we detected low outliers is 4B-i. In the dataset 4B- and 4B-iii we detected 2 low outliers, and in the dataset 4B-iv, three of them. This might be source of higher uncertainty.

Frequency Curve for: Ceo niz-Senta-FLOW-PEAK

Percent Chance Exceedance	Computed Curve Flow in cms	Variance Log (EMA)	Confidence Limits Flow in cms	
			0.05	0.95
0.1	4593.0	0.00237	5827.0	3947.2
0.2	4374.4	0.00185	5401.1	3826.9
0.5	4075.3	0.00128	4859.3	3647.4
1.0	3839.5	0.00093	4462.4	3492.3
2.0	3593.0	0.00065	4073.2	3315.9
5.0	3244.8	0.00039	3563.9	3040.2
10.0	2956.6	0.00027	3179.4	2789.8
20.0	2633.5	0.00021	2792.6	2494.1
50.0	2090.2	0.00020	2207.4	1979.0
80.0	1637.8	0.00026	1737.1	1531.8

Distribution Parameters

Parameter	Value
Mean	3.316
Standard Dev	0.123
Station Skew	-0.193
Regional Skew	
Weighted Skew	
Adopted Skew	-0.193
EMA Estimate of MSE (G at-site)	0.071

Events

Event	Number
Historic Events	0
High Outliers	
Low Outliers and Zero Flows	0
Missing Flows	0
Systematic Events	86
Historic Period	86
Equivalent Record Length (years)	86.000

Figure 4: HEC-SSP B17C tabular output for the reference dataset – 1

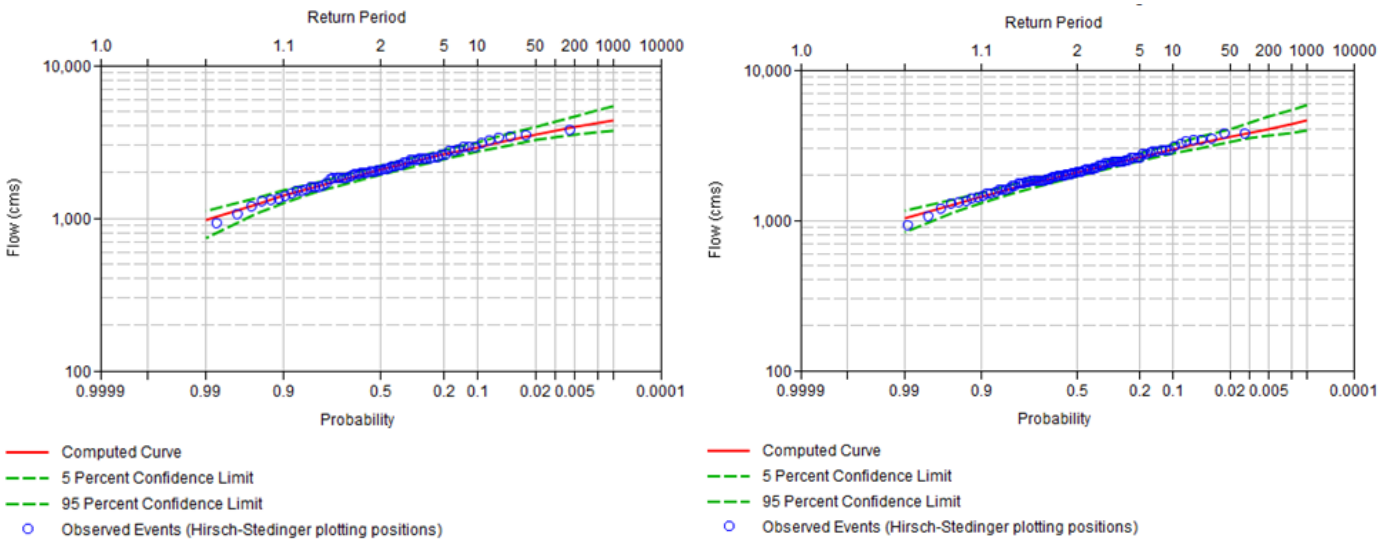

Figure 5: Probability plot for the series 5 (left) and 1 (right)

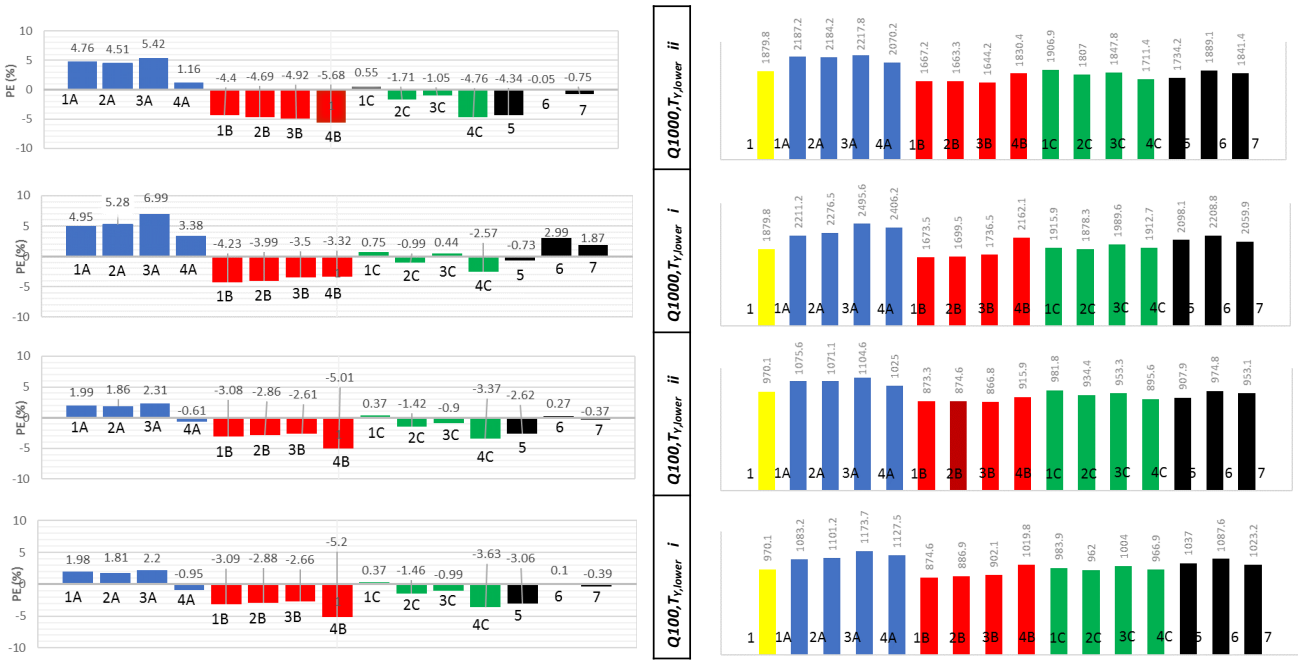

Figure 6: Percentage error, PE, (left column) and 90% confidence interval width in m³/s (right column) of Q1000 (rows 1, 2) and Q100 (rows 3, 4), $T_{y,lower}$ options: ii (rows 1, 3), i (rows 2, 4)

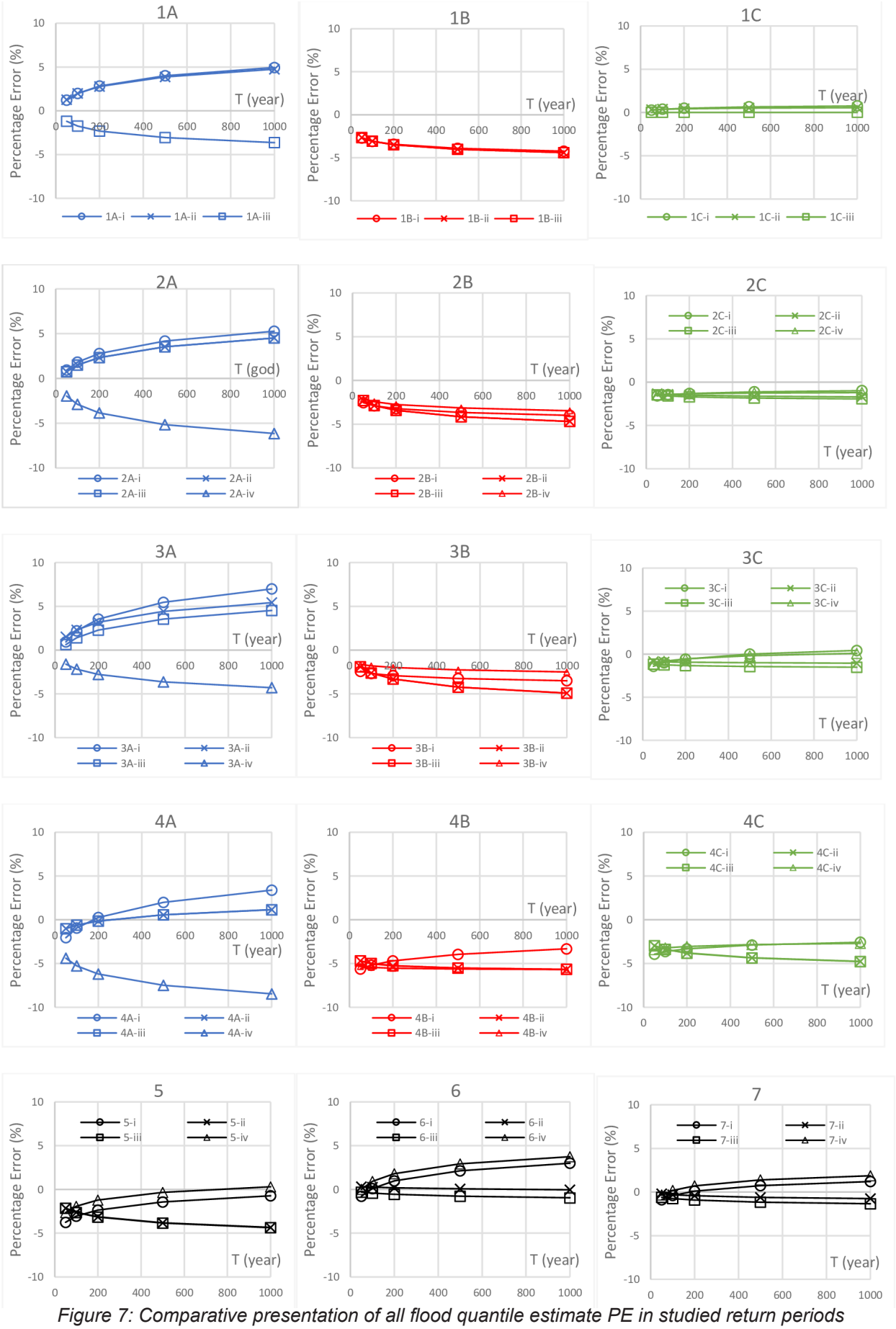

Figure 7: Comparative presentation of all flood quantile estimate PE in studied return periods T (50-, 100-, 200-, 500- and 1000-years) assessed from the base datasets by setting four $T_{Y,lower}$ options.

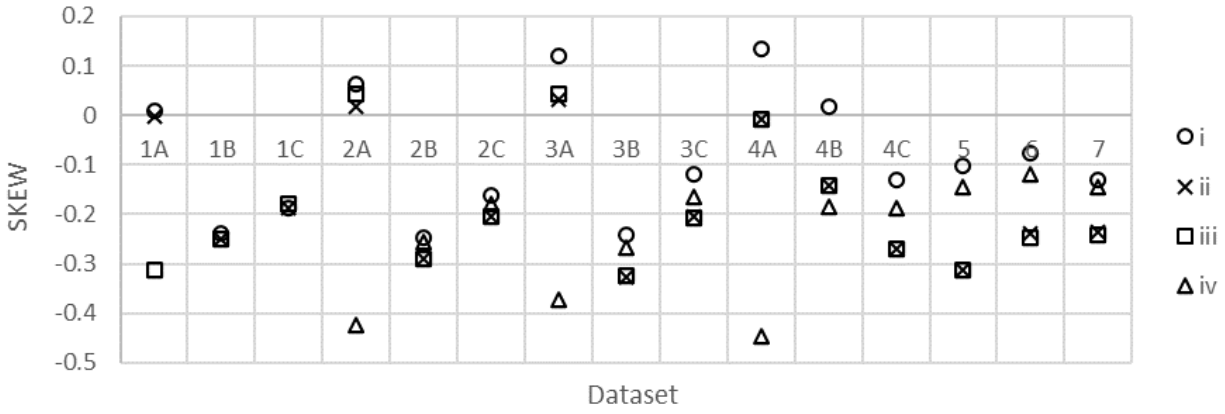

Figure 8: The base dataset skewness change depending on the lower threshold bound applied

For option i of lower threshold bound, uncertainty is higher in *QT* from datasets 5, 6 and 7. Quantiles from datasets 1B, 2B and 3B exhibit lower uncertainty in both quantile estimates shown in Fig. 6.

In general, majority of *QT* estimates of our interest exhibit *PE* in the range +/- 5%, except for datasets 2A-iv, and 4A-iv and 4B-iv, where *QT* are underestimated up to -8.5%.

Despite large data gaps after the year 1991 in the datasets 5, 6 and 7, *QT* estimates of our interest exhibit PE in the range +/- 5%, where options ii and iii yield underestimates, and threshold options i and iv overestimates of the *QTs*. The second largest flood in the reference period occurred in 2006, while the dataset with the largest gap that excluded this record entry is dataset 7. It is interesting the *QTs* do not show larger *PE* compared to datasets 5 and 6 where 2006 flood is present in the tested datasets.

Among the values we set for $T_{Y,lower}$, option iv (values of missing flows) in the dataset 1A (the same as option iii - Maximum peak flow of missing data) is the closest to the B17C recommendation. However, better results are obtained by setting other values of considered lower bound threshold options.

When setting lower threshold at the value of missing data (iv), the largest *PE* is obtained in majority of the datasets we created. At the same time, it may be seen in Fig. 8, the largest (absolute) value of the skewness coefficient (skew - *G*) is exhibited in the base datasets A for the $T_{Y,lower}$, option iv. In the reference dataset skew is low (-0.193), while in all considered datasets it ranges from -0.446 to 0.135. In general, all these skews are considered low.

Considering complexity of the skewness coefficient association with the quantile uncertainty response due to gaps in the dataset, further analysis is required. Not only that this issue is beyond research scope of this paper, but low skewness of both the reference dataset and base datasets, makes them unfit for further study in that respect.

CONCLUSION

In FFA according to B17C methodology, perception thresholds are primarily intended for unobserved historic flood representation, and period between its occurrence and beginning of systematic record. In some cases, they can be used for missing data during observation period. There is a variety of recommendations on setting both upper and lower threshold bound [3]. Most of them rely on previous knowledge and information about unrecorded floods. Ignoring some of the recommendations, we investigated the case of setting $T_{Y,lower}$ at values that could be derived from the existing record, not accounting for any other information and data source. In many parts of the world, this reflects reality in engineering practice when it comes to single site analysis and contents of flood data records. When flood data records offer information about a range of annual flood values, a concept of flow interval should be used for data representation.

The research results indicate that acceptable quantile estimates may be obtained using available peak flow record with data gaps. Encouraging results are obtained in the case $T_{Y,lower}$ is set to value of the largest flood within the remaining (incomplete) data record, regardless of the gap size and location. Percentage error of flood quantile estimates even in the case of 1000-year flood, indicate +/-5% deviation from the reference value when $T_{Y,lower}$ is set to value of the largest flood within a systematic data record with gap up to 23% of a period.

It should be noted our study case is HS Senta with large drainage area and low skewness coefficient. The results should be confirmed for HS with smaller drainage area, while observing the dataset skewness reflection on the quantile uncertainty due to gaps.

Improving the capability of dealing with missing data in FFA for engineering practice contributes to lessening uncertainty in flood studies, including flood hazard assessment. Such an improvement passes on to flood risk estimation results, and further, to flood risk zoning for insurance purposes [10].

ACKNOWLEDGEMENT

The authors are grateful to the Republic Hydrometeorological Service of Serbia for making data available for this research. The work for this study is financed from the research project TR37005 'Climate change impact on the rivers in Serbia' for the Ministry of Education, Science and Technological Development of the Republic of Serbia.

REFERENCES

1. Tencaliec, P., Favre, A. C., Prieur, C., Mathevet, T. (2015). Reconstruction of missing daily streamflow data using dynamic regression models. Water Resources Research, vol. 51, no. 12, 9447–9463, DOI:10.1002/2015WR017399
2. Pappas, C., Papalexiou, S. M., Koutsoyiannis, D. (2014). A quick gap filling of missing hydrometeorological data. Journal of Geophysical Research: Atmospheres, vol. 119, no. 15, 9290–9300, DOI:10.1002/2014JD021633
3. England, J.F., Jr., Cohn, T.A., Faber, B.A., Stedinger, J.R., Thomas, W.O., Jr., Veilleux, A.G., Kiang, J.E., and Mason, R.R., Jr. (2018). Guidelines for determining flood flow frequency—Bulletin 17C (ver. 1.1, May 2019). U.S. Geological Survey Techniques and Methods, Book 4, chap. B5, p. 1-168. https://doi.org/10.3133/tm4B5
4. Slater, L., Villarini, G. (2017). On the impact of gaps on trend detection in extreme streamflow time series. International Journal of Climatology, vol. 37, no. 10, 3976-3983, DOI: 10.1002/joc.4954
5. USWRC. (1982). Guidelines for Determining Flood Flow Frequency, Bulletin No. 17B. U.S. Water Resources Council, Subcommittee on Hydrology, Washington, D.C.
6. U.S. Army Corps of Engineers. (2019). Statistical Software Package HEC-SSP User's Manual, Version 2.2, US Army Corps of Engineers – Hydrologic Engineering Center
7. Blagojevic B., Mihailovic V., Plavsic J. (2014). Outlier treatment in the flood flow statistical analysis. Proc. Int. Conf. On Contemporary Achievements in Civil Eng., Facity of Civil Engineering Subotica, University of Novi Sad, Subotica, Serbia, p. 603-609, DOI: 10.14415/konferencijaGFS2014.081
8. Hircsh, R.M., Stedinger, J. (1987). Plotting positions for historical floods and their precision. Water Resources Research, Vol. 23, No.4, 715-727.
9. Vukmirovic, V. (1990). Analiza verovatnoce pojave hidroloskih velicina. Naucna knjiga. Beograd
10. Hanak, T., & Korytarova, J. [2014]. Zone rizika sa aspekta osiguranja - poredjenje poplava, sneznih lavina, olujnog vetra i oluja sa gradom. *Journal of Applied Engineering Science*, 12(2), 137-144.

Paper submitted: 18.10.2020.
Paper accepted: 15.12.2020.