Research Article

Estimation of the upper limit of confidence interval with the Monte Carlo method

Olsi Barko
Polytechnic University of Tirana, Albania

Idlir Lami
Polytechnic University of Tirana, Albania

DOI: https://doi.org/10.2478/bjir-2025-0002

Abstract

The assessment of peak flows with different return period is one of the important stages during the design of hydraulic works. The best method of their assessment is the statistic method, wich consists of using a measured series of maximum annual flows (preferably as long as possible) in order to be as reliable as possible and the error as small as possible during the probabilistic interpolation. Since the uncertainty of the assessment cannot be avoided due to different factors, then with the standard error method, a confidence interval is obtained which determines the interval where the flow value with a certain return period will be located on average in 95% of cases. Another method of dealing with the uncertainty in determining the maximum flow is by using the Monte Carlo method. The assessment of peak flows using the Monte Carlo method has been applied to the case study of the Shkumbin River at the Rrogozhine measurement site.
In the conclusions presented in this article, the results of the uncertainty assessment for the maximum flow found with the standard error method and with the Monte Carlo method will be compared.

Keywords: Peak flow, Confidence Interval, Stochastic method, Synthetic series, Monte Carlo method.

1. Introduction

Estimating peak flows with the statistic method is the best method for their estimation. This method is based on a observed series of annual peak flow in the river section where these peak flows are required. Uncertainty in the assessment of the flows with this method is inevitable regardless of the series measured in the field because there are errors in measurement or other errors for which we do not have information but which are located within the information contained in the measured series.

The simplest way of estimating this uncertainty is to treat it with the standard error method according to which the 95% confidence interval where the true value is located is ±1.96SE, where SE means the standard error.

Another method of uncertainty estimation is the Monte Carlo method. This method is applicable in cases where the measured series meets two conditions, that of independence of values and that of stationarity. If the measured series meets these conditions, then the values of this series are uncorrelated stochastic events that have any probability of occurrence (random) from zero to one, where the probability density distribution obeys a certain law. If we were to make further measurements and from the observation to form another series with the same length, we would have a series with other values but with the same probability distribution as the series measured at the beginning. So based on this logic we can generate endless synthetic series where the values of the series will be different from the measured series but the distribution will be the same. This is made possible by the Monte Carlo method. After generating a synthetic series where the inputs are the parameters of the measured series and the random probability of its values, based on the cumulative probability curve (cdf), we are able to determine the flow with a certain return period QT or with a certain non-occurrence probability. So for each generated series we have a value of QT. If we generate 10,000 such series, we get 10,000 QT values that would constitute a series with a certain probability distribution. This distribution can be determined by different software and then by the inverse function we determine the upper limit of the confidence interval with 95% certainty.

The main purpose of this paper is precisely to compare the standard error method when estimating the upper limit of the confidence interval with the Monte Carlo method.

In order to be able to compare, we must consider something important. The standard error method gives the confidence interval with 95% certainty, which means that 95% of the values will be in this interval and 5% will fall outside it. More precisely, 2.5% will be below the lower limit and 2.5% above the upper limit.

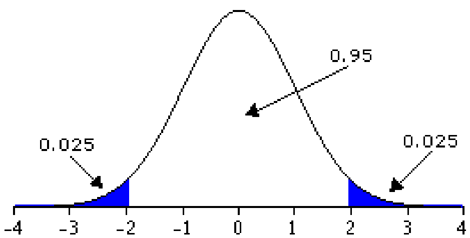


Figure 1: Confidence interval illustration

2. Study methods

2.1 *The upper limit of the confidence interval with the standard error method and the uncertainty in the flow estimation according to this method*

The statistical analysis of each hydrological variable, including flow, is based on the series of observed data. These data samples are somewhat limited in size from a statistical point of view. The parameters of the probability distribution estimated from the observed stream series data at a given location may not be the "true" parameters of the distribution of the population of streams at that location. Therefore, the question is always raised as to how much uncertainty exists in the estimation of flows that are determined by statistical analysis.
A commonly used method for estimating the uncertainty when estimating (flow with exceedance probability equal to p) is its standard error SE [], which is equal to the square root of the variance of the estimate, var [].
Confidence intervals for the quantile estimates are obtained using the standard error SE []. In general, when the series is very long, the quantile has a normal distribution. Lower and upper bounds of the confidence interval for are calculated from:

The upper limit:

$$x_{p,U} = x_p + z_{1-\alpha/2} \cdot SE[x_p] \qquad (1)$$

Lower limit:

$$x_{p,U} = x_p - z_{1-\alpha/2} \cdot SE[x_p] \qquad (2)$$

Where $z_{1-\alpha/2} = -z_{\alpha/2}$ is the standardized normally distributed variable for corresponding levels of a $1-\alpha$ confidence interval.

Where is the standardized normally distributed variable for corresponding levels of a confidence interval.

- Normal distribution.

The standard error of a quantitative estimate is:

$$SE[x_p] = \frac{S_x}{\sqrt{n}}\sqrt{1+\frac{1}{2}z_p^2} \qquad (3)$$

where is the standardized normally distributed variable corresponding to the probability of not exceeding p, is the standard deviation of the normal variable of the observed data, and is the length of the observed series.

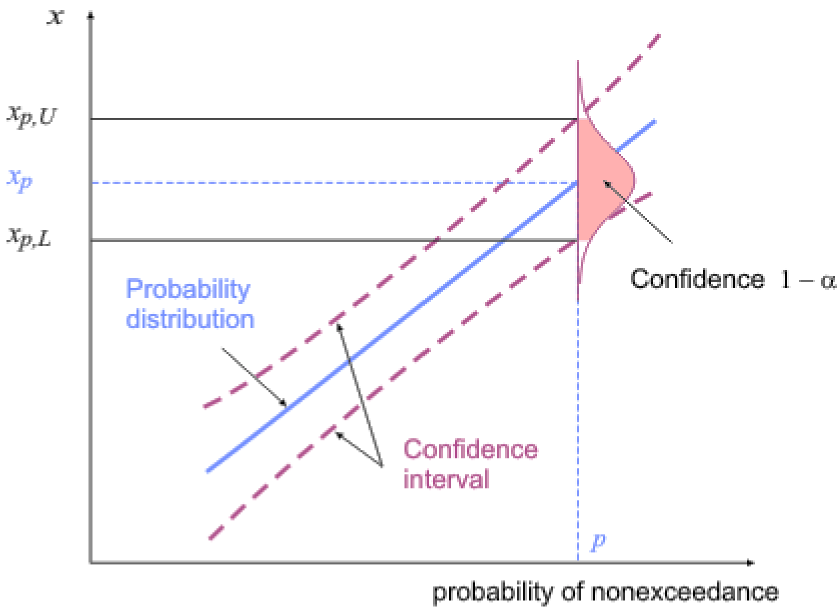


Figure 2: Confidence interval indicator chart. [2]

- Log-normal distribution.

For a log-normally distributed variable , the variable Y = ln X (or Y = log X) has a normal distribution. As a result, for a normal distribution, the standard error of a normal quantile estimate is:

$$SE[y_p] = \frac{s_y}{\sqrt{n}}\sqrt{1+\frac{1}{2}z_p^2} \quad (4)$$

where, s_y is the standard deviation of the logarithmic data and the limits of the confidence interval in this case are:

$$y_p \pm z_{1-\alpha/2} \cdot SE[y_p] \qquad (5)$$

From this we derive the approximate confidence interval limits for the original variable $X = e^Y$ which are:

where, is the standard deviation of the logarithmic data and the limits of the confidence interval in this case are:

$$e^{y_p \pm z_{1-\alpha/2} \cdot SE[y_p]} \qquad (6)$$

While in logarithmic base of 10, the limits of the confidence interval for the original variable $X = 10^Y$ are:

$$10^{y_p \pm z_{1-\alpha/2} \cdot SE[y_p]} \qquad (7)$$

There are also formulas for determining these limits for other distributions such as Pearson 3, Log-Pearson 3 and Gumbel.

2.2 *The upper limit of the confidence interval with the Monte Carlo method*

Once we have calculated a flow value corresponding to a given return period the next step is to consider the uncertainty. These uncertainties come from various sources, as it may.

Table 1: The cause of uncertainty according to Liu and Frigaard (2001)

• Variability of series values due to its limited length
• Errors related to measurement, visual observation.
• Selection of the distribution as a representative of the unknown true long-term distribution
• Variability of algorithms (choice of threshold, adaptation method, etc.)
• Climatological changes

The uncertainties that can be taken into consideration and reduced by the numerical simulation with the Monte Carlo method are those caused by the measurement and those due to the variability of the values of the series due to its limited length. The numerical simulation method that has been applied here is Monte Carlo Simulation with the following procedure, according to (Liu and Frigaard, 2001).

An array of random numbers in the range [0-1] is calculated or generated. The size of this group is equal to the length of the series in which the flow value with a certain return period Q_T was calculated.

With the previously calculated parameters of the distribution function (Lognormal: μ,σ), the inverse distribution function is calculated, the accumulated probability of which is the randomly generated values.

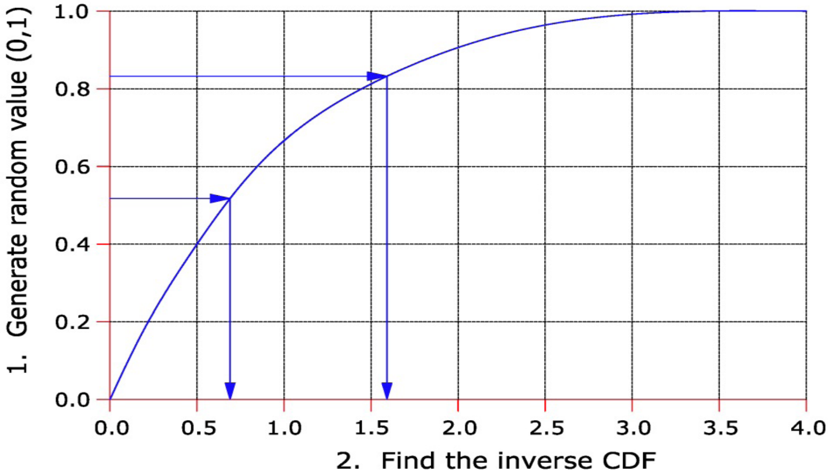


Figure 3: Graph indicating the use of the inverse function

From these calculations, a new series of flow values is created that are distributed according to the specified function. The best theoretical distribution is then fit to this set by calculating the new parameters again. Finally, with this new distribution function, the flow value with a certain return period is calculated. These steps are repeated until a suitable length of this group is obtained, e.g. 10,000 or 15,000 values (Liu and Frigard, 2001).

The value of the variable with a certain recurrence period usually follows a normal distribution but for the case of maximum flows this may change and may have a distribution Gumbel, Logpearson 3, Pearson 5 (3P), Johnson SU, Gen.Gamma (4P) etc. Below is an illustration of the case where the histogram fits the Johnson SU distribution better than the Normal distribution.

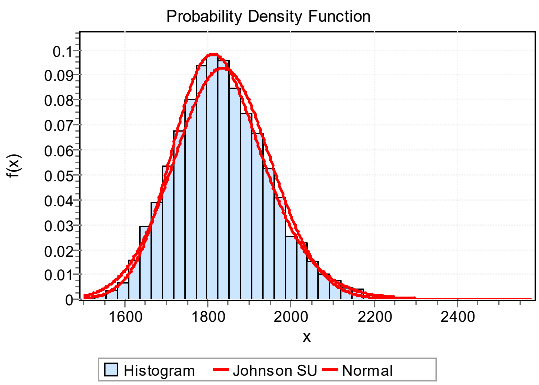


Figure 4: Best distribution graph

The value of the upper limit of the confidence interval (eg 90%), can be found by calculating the inverse of the distribution function of the series generated by the Monte Carlo method. This can be done only after we have found in advance which is the best distribution and we have found its parameters.

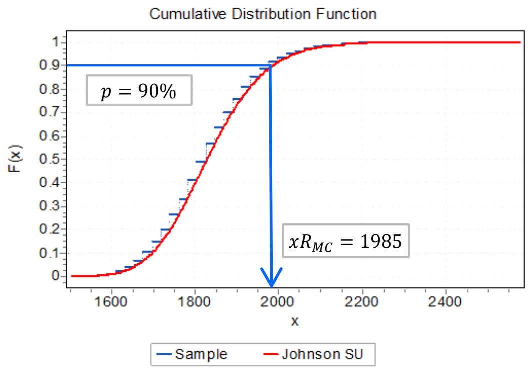


Figure 5: Finding of the upper limit by the inverse function

In this method, the uncertainty that arises due to the limited number of series values is reduced quite well. This will be seen better during an application we made for a specific case. The application of this method, i.e. the generation of synthetic series with the Monte Carlo method, requires that the observed series fulfill two conditions, that of the independence of the values against each other and that of stationarity.
These conditions are generally met for the series of maximum annual flows because the phenomena that cause them, i.e. the rains, are distant in time and have no correlation with each other. Even the condition of stationarity for these series is fulfilled, so we have the right to generate synthetic series where the variables have a distribution like the measured series and where the probability of occurrence is a random number.

3. Results for a case study

In the following, we are presenting in a clearer way, the difference between these two methods by applying it to a series measured in the Shkumbin River at the Rrogozhinë hydrological station.
The measured series has information for 43 consecutive years, from 1949 to 1991, including the years that have caused flooding in the lower part of this river.

Table 2: Measured Series (Instituti Hidrometereologjik, 1949-1991)

Nr.	Viti	Prurja [m3/s]	Nr.	Viti	Prurja [m3/s]
1	49	320	23	71	1820
2	1950	489	24	72	468
3	51	700	25	73	476
4	52	1200	26	74	1480
5	53	909	27	75	430
6	54	287	28	76	987
7	55	668	29	77	604
8	56	897	30	78	1090
9	57	362	31	79	536
10	58	700	32	1980	897
11	59	645	33	81	1030
12	1960	684	34	82	831
13	61	780	35	83	520
14	62	1630	36	84	624
15	63	1160	37	85	1080
16	64	850	38	86	970
17	65	731	39	87	942
18	66	836	40	88	720
19	67	756	41	89	570
20	68	582	42	1990	431
21	69	752	43	91	590
22	1970	561			

First, we determine the upper limit of the confidence interval according to the first method and then with the Monte Carlo method.

3.1 *Standard Error Method*

In the first method, to determine the probabilistic distribution that best fits our measured series, is the EasyFit software, which contains a large number of theoretical distributions in its archive.
This software calculates the empirical probability using the Kolmogorov-Smirnov, Anderson-Darling and χ-Square tests and evaluates the correspondence between the theoretical and empirical probability distributions of the measured series.
From the evaluation of the tests, there is the following ranking. We note that Kolmogorov-Smirnov evaluates as the best distribution Pearson 6 (4P), Anderson-Darling evaluates Gen. Logistic while χ2 estimates the Lognormal distribution.

Table 3: Estimating the best distribution

#	Distribution	Kolmogorov		Anderson		Chi-Squared	
		Statistic	Rank	Statistic	Rank	Statistic	Rank
1	Frechet (3P)	0.03828	6	0.08397	3	0.34749	9
2	Gen. Extreme Value	0.03921	8	0.08886	5	0.34853	10
3	Gen. Gamma (4P)	0.0444	11	0.13777	13	0.17205	6
4	Gen. Logistic	0.0472	12	0.07771	1	0.14931	4
5	Gumbel Max	0.04311	10	0.11128	12	0.31036	8
6	Log-Gamma	0.0413	9	0.10638	11	0.38227	14
7	Log-Logistic (3P)	0.04787	13	0.08155	2	0.16217	5
8	Log-Pearson 3	0.03849	7	0.09717	10	0.06562	3
9	Logistic	0.11605	19	0.72767	17	1.4538	16
10	Lognormal	0.03563	3	0.09428	8	0.06444	1
11	Lognormal (3P)	0.03586	4	0.09539	9	0.35462	12
12	Pearson 5 (3P)	0.03703	5	0.08697	4	0.35341	11
13	Pearson 6	0.03542	2	0.08938	6	0.06557	2
14	Pearson 6 (4P)	0.03536	1	0.09257	7	0.3553	13

Based on this assessment, we are choosing as the most representative distribution the Lognormal distribution of the $\chi 2$ test since the determination of the parameters of the series is easier. We will need this later during series generation.

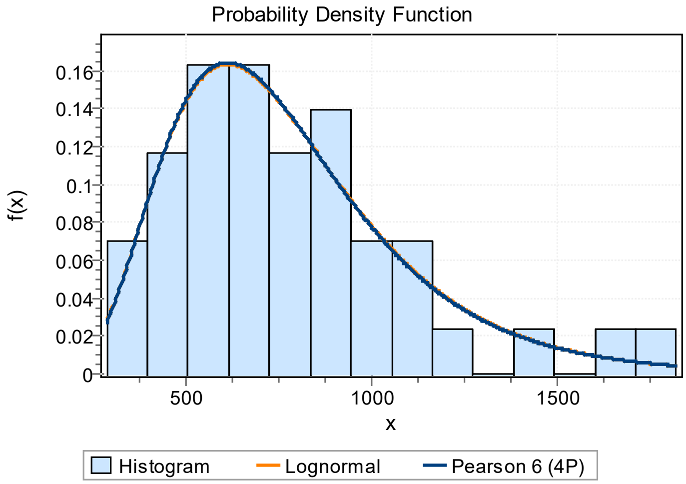


Figure 6: Best distributions according to EasyFit

This choice was also made due to the fact that the lognormal distribution and Pearson 6(4P) of Kolmogorov-Smirnov are very similar, as well as due to the fact that in this test the lognormal distribution is rated as the third best.

In order to further determine the maximum flows with different return periods, we used the software Hyfran-Plus, which simplifies the calculations and gives the flow values as well as the confidence interval limits for different distributions.

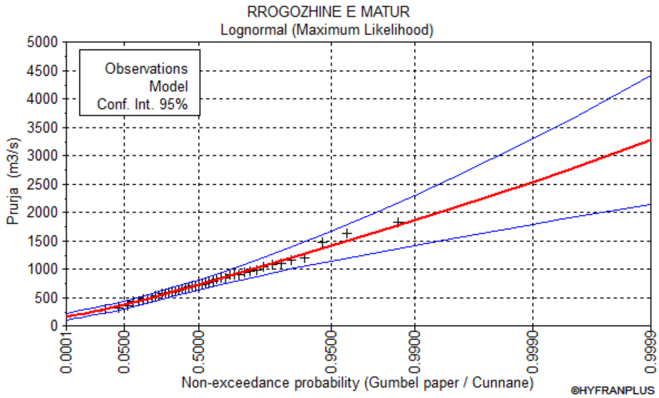


Figure 7: Flow as a function of the probability of non-occurrence according to the Lognormal distribution

For the Lognormal distribution we have these values in graphical and tabular form.

SERIA E MATUR RROGOZHINE				Number of observations 4	
Results of the fitting			Lognormal (Maximum Likelihood)		
Parameters				mu	6.579872
				sigma	0.407447
Quantiles			q = F(X) : non-exceedance probability		
			T = 1/(1-q)		
T	q	XT	andard deviat	Confidence interval (95%)	
10000	0.9999	3280	579	2360	3510
2000	0.9995	2750	438	2050	3010
1000	0.999	2540	383	1920	2800
200	0.995	2060	268	1610	2300
100	0.99	1860	224	1480	2090
50	0.98	1660	184	1340	1870
20	0.95	1410	135	1160	1590
10	0.9	1210	102	1020	1370
5	0.8	1020	73.6	874	1140
3	0.6667	859	55.8	752	967
2	0.5	720	44.8	638	813
1.4286	0.3	582	38.6	512	664
1.25	0.2	511	37.1	441	590
1.1111	0.1	427	36	347	505
1.0526	0.05	369	35.4	275	446
1.0204	0.02	312	34.5	200	388
1.0101	0.01	279	33.7	153	353
1.005	0.005	252	32.9	112	324
1.001	0.001	205	30.8	34.1	269
1.0005	0.0005	188	30	5.81	250
1.0001	0.0001	158	28	-51.3	211

Table 4: Estimation of confidence interval limits with Hyfran-Plus software
From the estimation made with Hyfran-Plus software, we see that the maximum flow for different return periods is as follows:
For return period T=20 years:
The maximum flow is 1410 m3/s and the upper bound of the 95% confidence interval is 1590 m3/s
For return period T=50 years:
The maximum flow is 1660 m3/s and the upper limit of the 95% confidence interval is 1870 m3/s
For return period T=100 years:
The maximum flow is 1860 m3/s and the upper bound of the 95% confidence interval is 2090 m3/s
For return period T=1000 years:
The maximum flow is 2540 m3/s and the upper limit of the 95% confidence interval is 2800 m3/s
For return period T=10000 years:
The maximum flow is 3280 m3/s and the upper limit of the 95% confidence interval is 3510 m3/s.

3.2 *The Monte Carlo method*

In the second method, the assessment of the upper limit of the confidence interval is done by generating synthetic series with the Monte Carlo method, which is based on the following. If the maximum flows are random variables and we are dealing with a stochastic event, then the measured series is a representative sample of the historical series and it represents this series with the probability density distribution.
This means that in the next 43, 50, 100 or 1000 years we will have another series, with other values different from the measured series but its probability distribution will be the same as our series with parameters similar to those of our series. This gives us the right to generate hundreds of series of the same length, where its variables have a Lognormal distribution and the probability a random number from zero to one.
The application of this method requires that the flows of the measured series enjoy the property of data independence as well as the property of stationarity. For this purpose, the test of independence according to Wald-Wolfovitz and the test of stationarity according to Kendall is performed using Hyfran-Plus software. From the evaluation we see that the measured series enjoys these two properties and we can continue with the generation of these series.

Table 5: Test of independence (Wald-Wolfowitz)

Hypotheses	
H0	The observations are independent
H1	Observations are dependent (autocorrelation of order 1)
Results	
Statistics value	$\lvert U \rvert = 0.316$
p-value :	p = 0.752
Conclusion We accept H0 at a significance level of 5 %.	

Table 6: Stationarity Test (Kendall)

Hypotheses	
H0	No trend is apparent in the observations
H1	There is a trend in the observations
Results	
Statistics value	$\lvert K \rvert = 0.230$
p-value :	p = 0.818
Conclusion	
We accept H0 at a significance level of 5 %.	

From the evaluation of the measured series, we found as the best distribution the Lognormal distribution with the following parameters:

Table 7: Parameters of the Lognormal distribution

Lognormal	
Mean	781.3
SD	330.3
Parameters	
Mu	6.57872503
Sigma	0.40548089

Based on the procedure explained in the theoretical part, synthetic series were created using the Monte Carlo method. Below we present the values of the measured series and a generated series, its probability density function (pdf) and its cumulative density function (cdf)

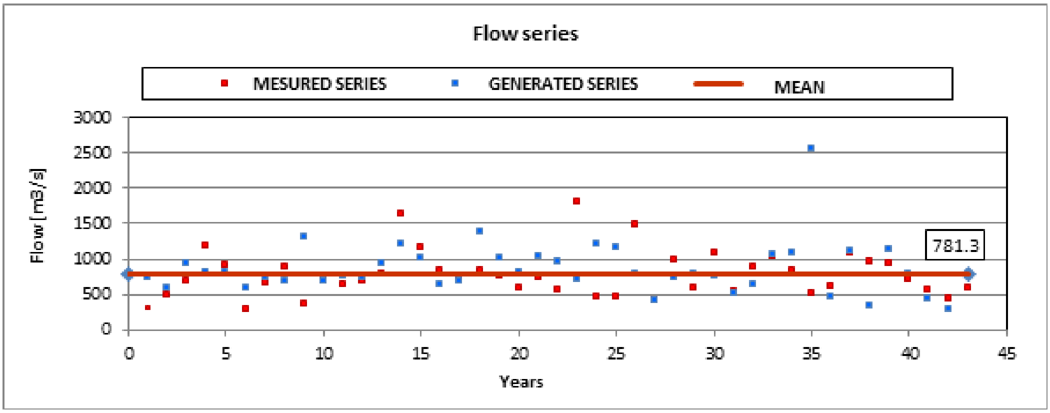
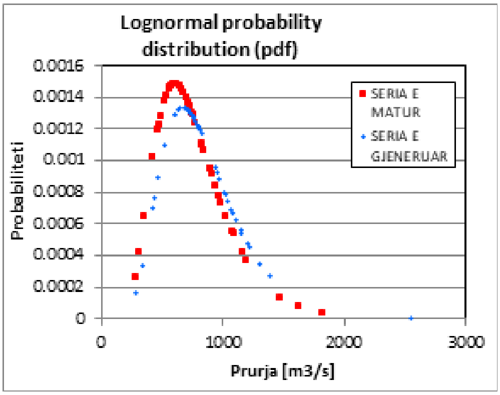


Figure 8: Measured series and one of the generated series



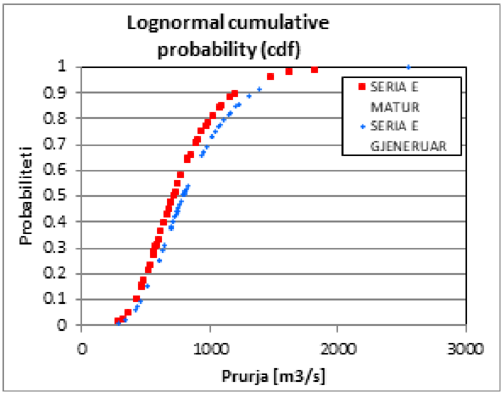


Figure 9: The lognormal probability distribution (pdf)

Figure 10: Cumulative lognormal probability (cdf)

For each synthetic series, we determine the flow with a return period T=20 years. For 10,000 such series, we get 10,000 of Q_{20}, which as a series has a certain probability distribution.

Below we are presenting the Q_{20} flow values for each series generated, the probability density function as well as cumulative probability function, for Log-Pearson 3 as one of the best distributions.

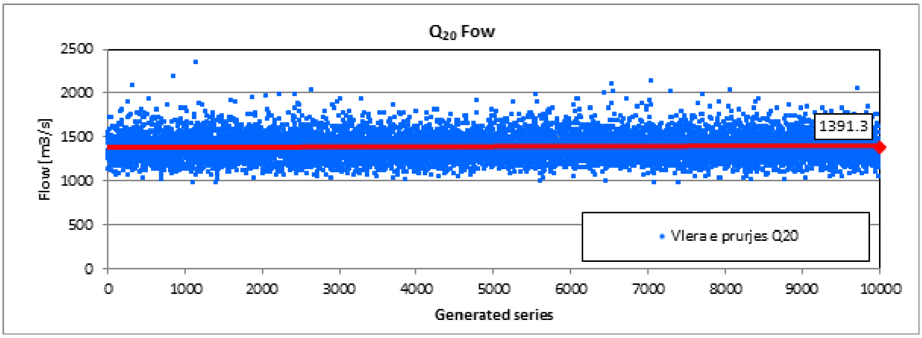


Figure 11: Flow Q_{20} for 10000 generated series

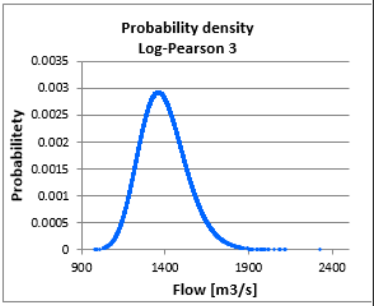


Figure 12: Probability density function

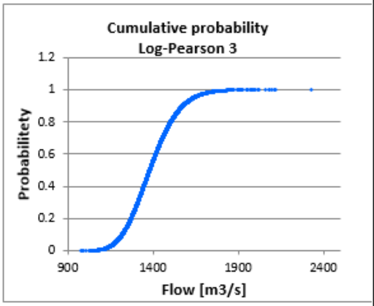


Figure 13: Cumulative probability

After finding 10,000 of Q_{20} flow values with a return period of 1 time in 20 years, we use the Easyfit program to determine which theoretical distribution best represents the probability density distribution for this series.

From the evaluation it was found that the best distributions are Johnson SU, Log-Pearson 3, Lognormal (3P), Pearson 5 (3P) and all four of these distributions are very close to each other.

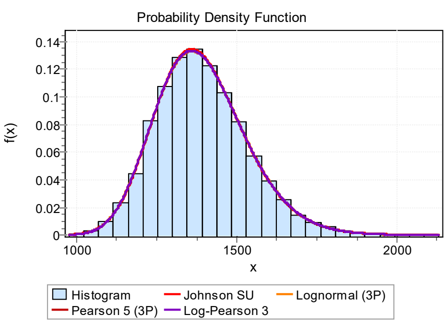


Figure 14: Probability density function for the Q_{20} series

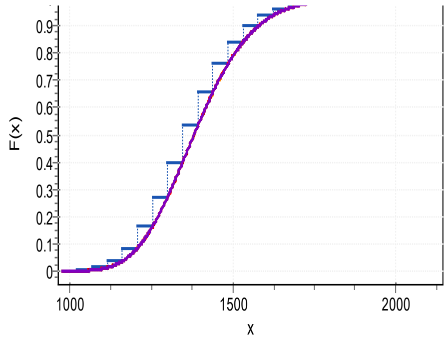


Figure 15: The cumulative probability function for the Q_{20} series.

The top three distributions are shown below:

Table 8: Evaluation of distributions by tests

#	Distribution	Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
		Statistic	Rank	Statistic	Rank	Statistic	Rank
1	Johnson SU	0.005	1	0.152	1	2.536	1
2	Log-Pearson 3	0.006	2	0.329	2	3.126	2
3	Pearson 5 (3P)	0.006	3	0.334	3	3.305	3

The parameters of the best distributions are as follows:

Table 9: Parameters of the best distributions

#	Distribution	Parameters
1	Johnson SU	γ=-4.5515 δ=4.7707 λ=444.37 ξ=889.48
2	Log-Pearson 3	α=165.89 β=0.00788 γ=5.9255
3	Pearson 5 (3P)	α=79.307 β=98241.0 γ=137.09

The value of the upper limit of the 95% confidence interval can be determined by calculating the inverse of the best distribution function. From the calculations we have these values for the three best distributions:

Table 10: The upper limit of the 95% confidence interval for T_{20}

Distribution	Q 20 [m3/s]
Johnson SU	1702
Log-Pearson 3	1701
Pearson 5 (3P)	1700

This procedure was also applied to flows with a return period of 50, 100, 1000 and 10000 years and the results are presented as follows.

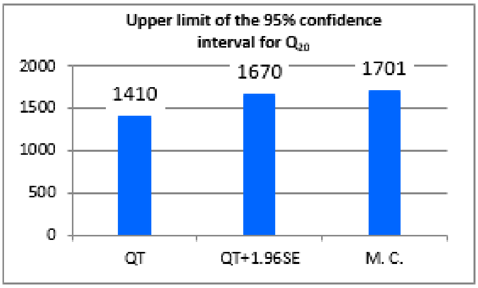


Figure 16: Upper limit of the 95% confidence interval for Q_{20}

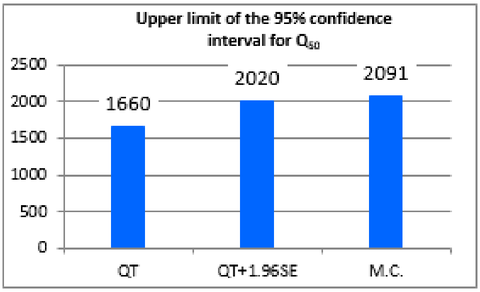


Figura 17: Upper limit of the 95% confidence interval for Q_{50}

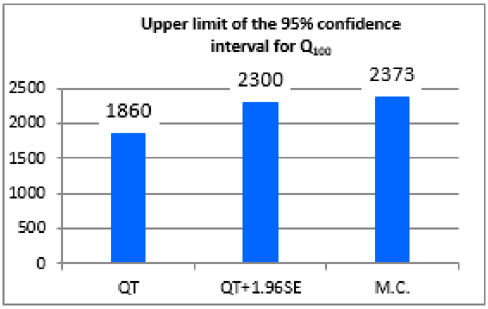


Figure 18: Upper limit of the 95% confidence interval for Q_{100}

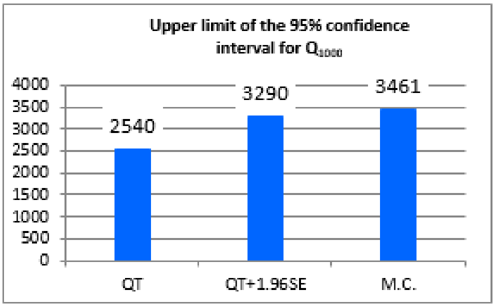


Figure 19: Upper limit of the 95% confidence interval for Q_{1000}

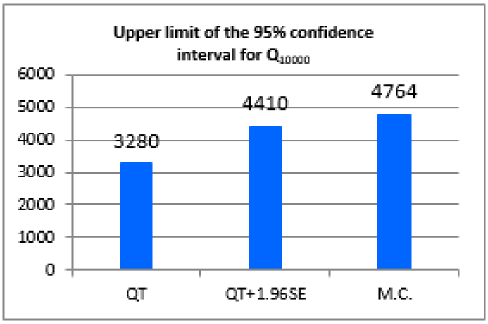


Figure 20: Upper limit of the 95% confidence interval for Q_{10000}

4. Analysis of the results

We note that the upper limit of the confidence interval with 95% certainty found by the Monte Carlo method, for the flow with return period 20 years is 1.85% greater than the upper limit found by the standard error method.

For the flow with return period 50 years, this difference is 3.51% greater.
For the flow with return period 100 years, this difference is 3.17% greater.
For the flow with a return period 1000 years, this difference is 5.19% greater.
For the flow with return period 10,000 years, this difference is 8.02% greater.
From the estimation of the upper limit of the 95% confidence interval, it was observed that the flows are greater than estimated by the standard error method. It was also observed that this difference increased with the increase of the return period. So if the uncertainty with the standard error method is taken into account by choosing the upper limit value as the worst case for the maximum flows, in the case of the Monte Carlo method this limit has a larger value that varies according to the following table:

Table 10: The upper limit of the 95% confidence interval for different return periods

	Standard Error method	Monte Carlo method	The difference %
T_{20}	1670	1701	1.85
T_{50}	2020	2091	3.51
T_{100}	2300	2373	3.17
T_{1000}	3290	3461	5.19
T_{10000}	4410	4764	8.02

5. Conclusion

From the estimation of the upper limit of 95% confidence interval the Monte Carlo method, for different return periods, it was observed that this limit is always greater than that estimated by the Standard Error method. This was the result during the application for a case study where the distribution of probabilities of the measured series was the Lognormal distribution with two parameters. If the upper limit estimated by the Monte Carlo method is always larger regardless of the distribution that best fit the measured series, this is to be evaluated in other case studies.
We can say that the Monte Carlo Method has these positive aspects regarding the uncertainty determination for peak flows.
Firstly, it is related to the fact that the generation of several synthetic series and finding of the maximum flow with a certain return period better represents the nature of the variability of this variable. This highlights how this variable moves against the mean, i.e. is it a symmetric distribution or not against it, and furthermore makes it possible to determine the type of this distribution to proceed further with finding the upper limit of the confidence interval.
Secondly, since when we use Monte Carlo simulation, we do not need to accept a

priori distribution (as is the case when accepting the Normal distribution when using the standard error method), we can say that using Monte Carlo simulation leads to a more accurate approach than using the standard error method.

Thirdly, it is related to the fact that this variable has a non-symmetrical distribution, but more precisely than that assumed by the standard error method, the upper limit for the same level of confidence may turn out to be larger. At least this turned out to be true for our case study, where the measured series had as the best fit distribution the two-parameter Lognormal distribution. However, how the upper limit of the confidence interval changes for cases where the series has a different distribution should be studied on a case-by-case basis. The fact that this method gives us a larger value of peak flow as well as a safer value due to the information it generates about its variability makes it a preferred method by design engineers as it reduces the uncertainty when estimating design parameters. as for our case which was the maximum flow for a given return period.

In conclusion, it can be said that the Monte Carlo Method for considering the uncertainty and estimating the upper limit of the confidence interval is a better than Standard Error method.

References

Bobée, B. and El Adlouni, S. (2008). Hyfran-Plus Version 2.2.

EasyFit Professional Version 5.5. Mathwave Technolologies.

E-Learning Platform for Integrated Flood Management (IFM). Uncertainty in estimating floods by frequency analysis. [Hamburg University of Technology].

Instituti Hidrometereologjik. (1949-1991). Buletinet Hidrologjike. Instituti Hidrometereologjik.

Liu and Frigaard. (2001). Generation and Analysis of Random Waves. Laboratoriet for Hydraulik og HavnebygningInstituttet for Vand, Jord og MiljøteknikAalborg Universitet.

Selenica. A. (2009) Hidrologjia Inxhinierike. Botimet "Dita 2000".