Lidija Tadić and Tamara Brleković

Flood Frequency Analysis

Hydrological processes are present in time and space in a manner that is partly predictable and partly random. We call them stochastic processes. In that case, the value of certain observations is not correlated with adjacent observations. This type of approach is appropriate for observations/measurements of extreme hydrological events such as floods, droughts or hydrological data averaged over longer time period such as annual precipitation. Statistical methods based on mathematical principles are only tools, which can be used to describe their random variations. These methods are focused on data, not on physical processes [7].

Flood is a natural event which cannot be prevented. It is usually defined as a temporary covering by water of land not normally covered by water. This shall include floods from rivers, mountain torrents, Mediterranean ephemeral water courses, and floods from the sea in coastal areas, and may exclude floods from sewerage systems [5]. It is a stochastic event the probability of which may be derived from a number of different sources. It may be derived directly from historic data on water levels or it may be derived indirectly from modelling. In both cases some form of historic data is needed. If modelling is used, then the historic data can be rainfall or river flow. The length of the available record is important in assessing the magnitude of events with small probabilities. Thus it is important to collect data routinely on both rainfall and river flow.

If a sufficiently long length of record is available, then it is possible to estimate the magnitude of floods with different probabilities directly from a historic record. Such historic data cannot be used to assess the impact of proposed works, so if this is required, then some form of modelling would have to be undertaken.

Flood frequency analysis is a hydrological procedure used to determine high flow values of certain probabilities in successive river cross-sections or hydrological profiles (stations).

Flood frequency estimates of recurrence of floods which is used in designing hydraulic structures such as dams, bridges, culverts, dykes, highways, sewage systems, waterworks, etc. In order to achieve the optimum and safe design of hydraulic structures, and to avoid over designing or under designing, it is necessary to apply statistical methods to determine flood frequency. It is also helpful in flood insurance, physical planning of a certain area or maintenance of the hydraulic structures.

If we have sufficiently long data series of flood flows, therefore, the calculation of empirical frequency distribution could be relatively precise under the assumption that natural and anthropogenic processes did not change relationships relevant for flood occurrence. In that case frequency is equal to determination of maximum measured

https://doi.org/10.36250/01052_03

annual discharge over a longer time period which can be relevant for design of flood protection structures. In most of the stations (rivers), there are not many measured data series which could be reliable for optimum and safe designing. Besides, there is always a chance of occurrence of flood greater than the maximum historical flood [9]. Problems arise when the substantial hydraulic structure in the watercourse has been constructed or any other change in the basin has been introduced which significantly changes the hydrological regime and discharges [8].

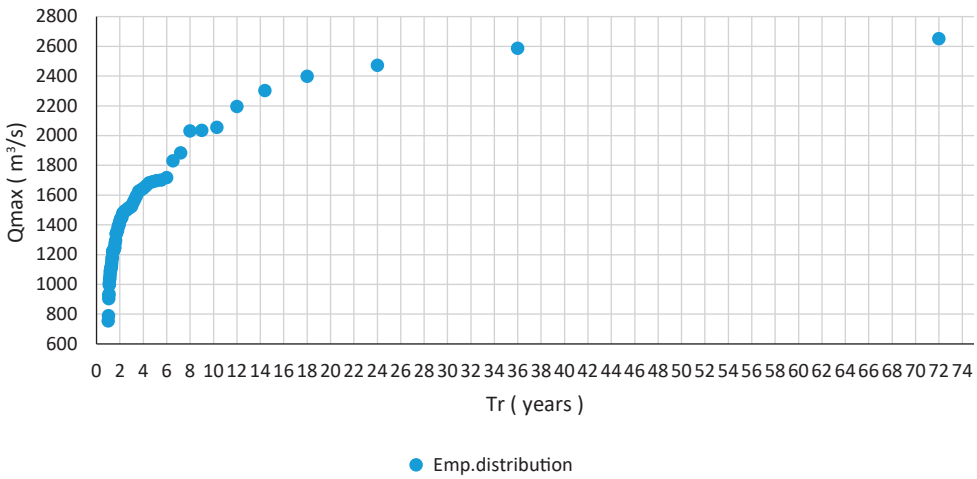

Figure 1. An example of empirical frequency analysis (compiled by the authors)

An example of empirical frequency distribution is presented in Figure 1 calculated on the basis of 71 years long data series of maximum annual discharges recorded on Botovo station (the Drava River in Croatia). In the observed period (1926–1998), the highest discharge occurred in July 1972 (2,652 m^3/s); this is a flood of about 72 years return period.

In order to make flood protection systems safe as much as possible, hydrologists have to use different statistical methods and apply statistical procedures on available data records. Usually only one parameter has been involved in the analysis (water level or discharge) with the following characteristics:

The magnitude of an extreme event is inversely related to the frequency of occurrence. In other words, the most severe floods occur less frequently.

Hydrological data are assumed to be independent and identically distributed [7].

The hydrological regime that produces floods is considered to be stochastic, time and space independent.

The flood frequency curve is used to relate flood discharge values to return periods to provide an estimate of the intensity of a flood event. The discharges are plotted against return periods using either a linear or a logarithmic scale. Generally, the frequency of maximum discharges is more reliable than water levels, because they are less dependent on riverbed deepening or other changes of the watercourse [8].

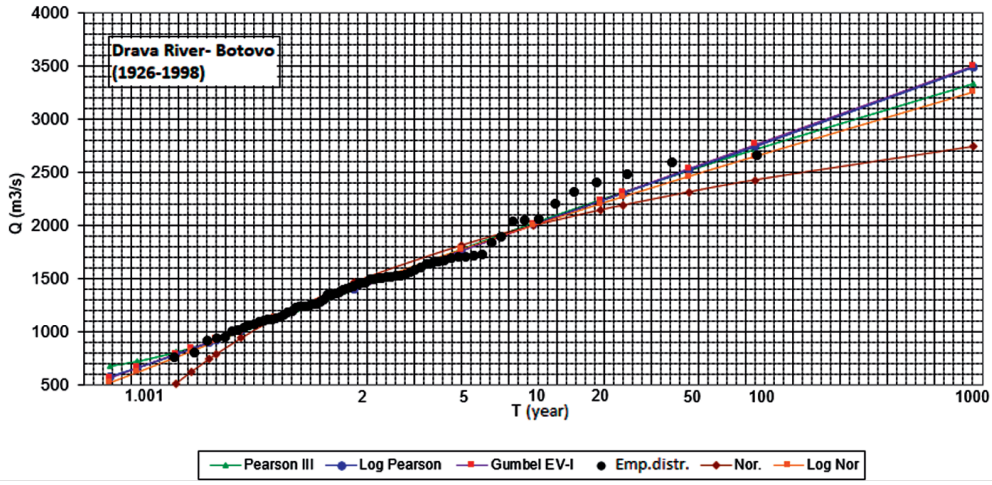

Figure 2. An example of theoretical frequency analysis (compiled by the authors)

An example of application of theoretical frequency distributions is presented in Figure 2. Five of the most common theoretical distributions were tested (Pearson III, Log-Pearson, Gumbel EV-I, Normal and Log-normal) and Log-normal distribution was selected as the most precise. Compared to results presented in Figure 1, the maximum recorded discharge is flood of about 100 years return period. According to presented Log Normal theoretical distribution, flood of 1,000 years return period would have a discharge of 3,245 m^3/s.

As it was illustrated, the result of statistical calculations are floods of different return period. Return period, also referred to as 'recurrence interval' is a term adopted by scientists and policy makers to estimate the likelihood and severity of extreme events (such as cyclones/hurricanes, flooding and earthquakes). It is based on the statistical analysis of data (such as historical climatic records, flood measurements), to provide a probability that an event of any given magnitude will occur in any given year. This probability is often used to assess the risk of these events for human populations. The concept is based on the magnitude-frequency principle, where large magnitude events (such as major cyclones) are comparatively less frequent than smaller magnitude incidents (such as rain showers).

In this approach, which is common in modern flood frequency analysis, it is essential to understand the concept of return period. The theoretical definition of return period is the inverse of the probability (generally expressed in %) that an event will be exceeded in a certain year. For example, the return period of a flood might be 1,000 years, expressed as its probability of occurring it would be 1/1,000, or 0.1% in any year. It means that, in any given year, there is a 0.1% chance that it will happen, regardless of when the last similar event was. Or, it is 10 times less likely to occur than a flood with a return period of 100 years (or a probability of 10%).

The most common misunderstanding about return periods, for example, the 100-year return period is that the flood of this magnitude will only occur once in 100 years. It is essential to understand that if a flood with a 100-year return period occurs now, it does not mean that another flood of this magnitude will not occur in the next 100 years.

EU Flood Directive (Directive/2007/60/EC)

Different countries used to have different approaches to flood frequency analysis as a basis of flood protection measures. The most common return periods used in flood protection are 2, 5, 10, 25, 50, 100, 1,000 and 10,000 years.

Since 2007, members of the EU accepted the common document, Flood Directive [5]. The main reason is the fact that flood risk is best managed on a basin level, not at individual member state level. Without going deeply into the articles of the Flood Directive, its major tasks are determining flood hazard maps of the geographical areas which could be flooded according to the following scenarios:

- floods with a low probability, or extreme event scenarios
- floods with a medium probability (likely return period ≥ 100 years)
- floods with a high probability, where appropriate

For each scenario referred to in the previous paragraph the following elements are important:

- the flood extent
- water depths or water level
- where appropriate, the flow velocity or the relevant water flow

These scenarios are important regarding the problem of return period. According to the Flood Directive, floods of high probability are all flood events with a return period of <100 years. For example, the Danube Flood Risk Atlas (2012) recognised areas along the Danube River affected by floods of a 30 years return period (HQ30). These areas along the river are frequently flooded. Generally flood plains, wetlands, forests and agricultural areas are affected. Usually the inundation areas of a 30-year-flood should serve for retention purposes in order to reduce the overall flood risk and be kept free of settlements buildings. These retention areas are often valuable biotopes, such as in Hungary and the Danube Delta.

The flood event with ≥100 years return period (HQ100) is widely accepted as the design level for flood protection measures along the Danube River. Normally, flood hazard in the areas between the limits HQ30 and HQ100 is known mainly to the residents having lived there for a long time. Agricultural land use is predominant; permission for settlement use should only be given exceptionally and with the provision of preventive construction measures.

During very rare events (HQ1000), flood extents and depths are distinctly larger, respectively higher than what has been observed so far. Existing flood protection works

might be overtopped or might fail to perform, thus describing a residual risk scenario. For the areas between a HQ100 and HQ1000 flood, no direct restrictions of land use arise; however, preventive flood strategies and emergency planning should be accounted for, especially regarding vulnerable objects. As potential preventive measures (such as evacuation plans) are highly dependent on flood depth, not only the limits of flooded areas, but also flood depth classes are illustrated.

According to the Flood Directive, the first scenario is defined as: determining floods with a low probability, or extreme event scenarios are given as a framework, which means that each country can make its own choice within it. The choice usually depends on previously established reference values.

Illustrations of hazard maps of the geographical areas which could be flooded according to chosen return periods are given in Figure 3.

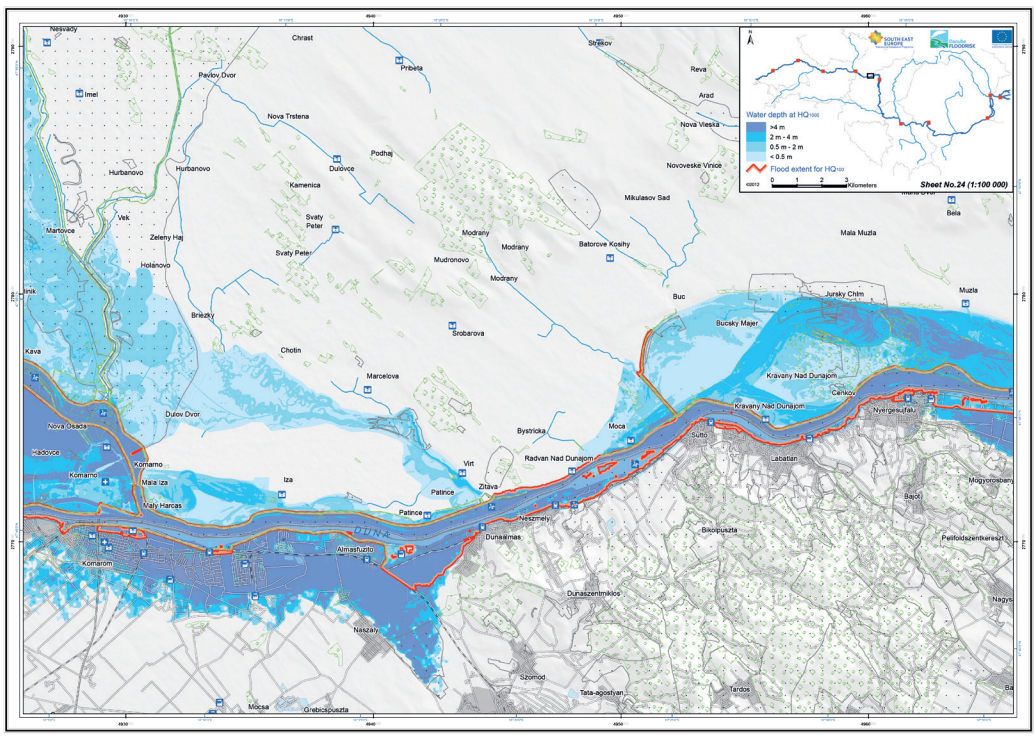

Figure 3. An example of flood hazard map at HQ100 and HQ1000 for one section of the Danube River [4]

Figure 3 presents one section of the Danube River and possible water depth related to discharges of 100 and 1,000 years return period (HQ100, HQ1000). There are dykes along both riversides and their height keeps water in the river flood plain until discharge exceeds HQ100 (red line), or flood with a medium probability (likely return period ≥100 years). Discharges of low probability (HQ1000) will cause floods of an adjacent area with water depth 0.5–4 m, depending on topographical conditions.

Figure 4 presents the relationship between the return period and related flooded areas in the Danube countries. It is clear that floods of low probability (extreme flood scenarios) have the greatest affected area according to the Danube Flood Risk Atlas.

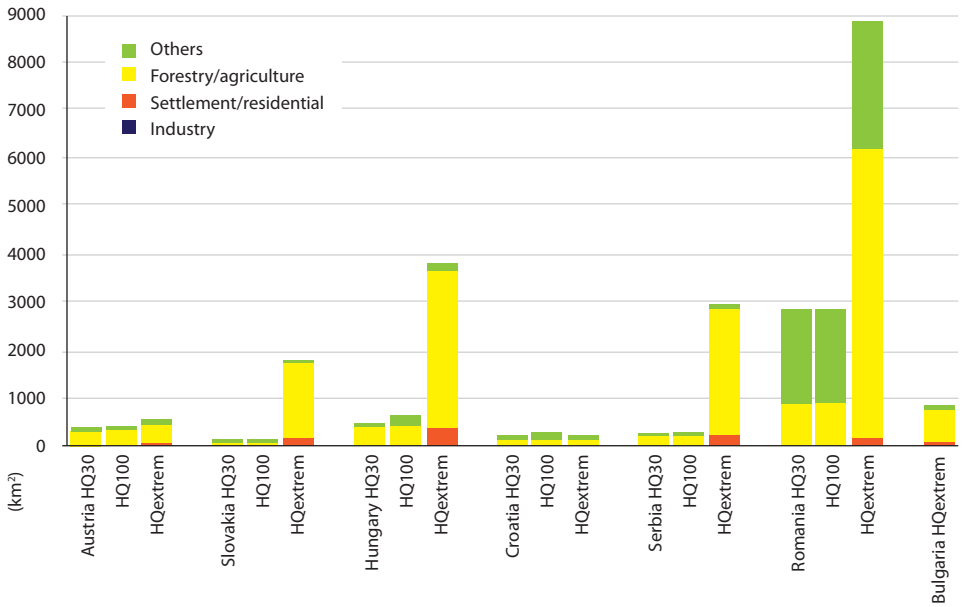

Figure 4. An example of the relationship between the flooded area and return period [4]

Empirical flood frequency

Flood frequency analyses are used to predict design floods for sites along a river. As it was illustrated in Figure 2, in order to estimate the return period of a given discharge or vice versa, the observed data is fitted with a theoretical distribution using a cumulative density function (CDF). This helps the users in analysing the flood frequency curve.

Using the annual peak flow data available for a number of years, flood frequency analysis is used to calculate statistical information such as mean, standard deviation and skewness which is further used to create frequency distribution graphs.

The mostly used frequency distributions in hydrology of extreme events are: Gumbel distribution (in the United Kingdom), Normal distribution, Log-normal distribution and Log-Pearson III distribution (in the USA). After choosing the probability distribution that best fits the annual maxima data, flood frequency curves are plotted. These graphs are then used to estimate the corresponding design flow values.

Procedure:

– using the observed annual maximum discharges of a period as long as it is possible to calculate basic statistical information such as mean values, standard deviations, skewness, etc.

– calculation of recurrence intervals (Figure 1) by using empirical equations such as Weibull equation which is one of mostly used empirical distribution and according to some authors it is the most accurate [6]

It is not the only one, there are a number of empirical distributions as it is presented in Table 1.

Table 1. Several methods of empirical distribution [6]

Method of "RI"	Proponent
m/N + 1	Weibull (1939)
(m-0.31)/(N + 0.38)	Beard (1943)
(m-0.44)/(N + 0.12)	Gringorten (1963)
(m-0.5)/N	Hazen (1914)
(m-0.3)/N + 0.4	Čegodajev (?)

Data records of maximum annual discharges are sorted in descending order and each annual peak have a certain rank, called the magnitude number, "m" (the highest value is ranked as m = 1, and the smallest value is valued as N). The number of items (data points) in the record is "N". The recurrence interval (RI) for a particular river profile (station) gives us information, how often we expect the river to exceed a certain discharge.

After the calculation of basic statistical parameters it is necessary to calculate the maximum annual discharge for different return periods by applying distributions.

Each distribution has its own characteristics and mathematical basis.

Normal (Gauss) distribution

In spite of the fact that Normal (Gauss) distribution is a symmetric two-parameters distribution and flood waves ordinarily have non-symmetric distribution, this distribution is very often used in flood frequency analysis.

The analytical expression is given in the form of frequency density function:

$$p_{(Q_M)} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(Q_M-\overline{Q_M})^2}{2\sigma^2}} \tag{1}$$

p(x) = normal distribution density function

s2 = standard deviation (variance) of the distribution

$\overline{Q_M}$ = mean value of the distribution

Introducing of transformation

$$z = \frac{Q_M - \overline{Q_M}}{\sigma} \tag{2}$$

gives an equation of standard normal distribution:

$$p_{(x)} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2}} \tag{3}$$

If the general variable x designates discharge Q, flood discharge QMp of different return periods (p) can be calculated as:

$$Q_{Mp} = \overline{Q_M} + z\sigma \tag{4}$$

Figure 5 illustrates a standardised normal distribution and its properties. It is clear that the density function is symmetric about the mean value ($\bar{x}$) and the function mode coincides with the mean value. The variance of standardised normal distribution s2 = 1 and the mean value $\overline{x} = 0$. The maximum value of density function is:

$$p(y) = \frac{1}{\sigma\sqrt{2\pi}} = 0.399 \tag{5}$$

$$\sigma = \sqrt{1}$$

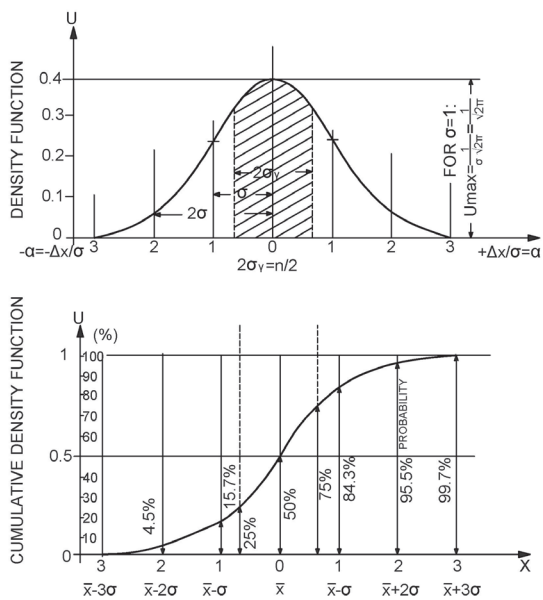

Figure 5. Cumulative density function [8]

The area below the curve presented in Figure 5 equals 1 and represents the number of values N. The total area divided in 2 parts defines 50% of the sample group in the interval:

$$[\overline{x} - 0.6745s, \overline{x} + 0.6745s] \tag{6}$$

The maximum deviation in this case is $s_{max} = 3s$.

Log-normal (Galton) distribution

If some data from the given data series are expressed as log values, in that case Normal distribution changes into Log-normal or Galton frequency distribution.

$$p_{(x)} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(\bar{q}-\bar{q}_o)^2}{\sigma^2}} \tag{7}$$

$$\bar{q} = \log Q_M \tag{8}$$

$$q_o = \frac{\sum \bar{q}}{n} = \frac{1}{2}\log\left(\frac{\overline{Q_M}^4}{\overline{Q_M}^2 + \sigma^2}\right) \tag{9}$$

$\bar{q}$ = log value of maximum discharge

qo = mean value of logarithms of log QM series

In Galton distribution, the reduced deviation is:

$$z = \frac{\bar{q} - \bar{q}_o}{\bar{\sigma}} \tag{10}$$

And the logarithm value of the maximum discharge (of different return periods) will be:

$$\bar{q}_p = \bar{q}_o + z\bar{\sigma} \tag{11}$$

The anti-logarithm of qp will give us the maximum discharge of the given return period (Annex: Table I).

Gumbel distribution

The Gumbel distribution is non-symmetric and two-parametric. According to the Gumbel probability of maximum (flood) discharge occurrence is defined by exponential function:

$$p(_{QM}) = ae^{-a(Q_M - Q^*)}e^{-a(Q_M - Q^*)} \tag{12}$$

where Q* and a presents parameters of the Gumbel distribution. Q* is a mode of Gumbel's curve:

$$Q* = \overline{Q_M} - \frac{0.577}{a} \tag{13}$$

Number 0.577 is Euler's constant and a is the parameter defined as:

$$\frac{1}{a} = 0.780\sigma \tag{14}$$

As in the previous distribution, the introduction of

$$z = a(Q_M - Q^*) \tag{15}$$

into the equation (10) will give

$$p(Q_M) = ae^{-z}e^{-e^{-z}} = p(Q_M) = e^{-e^{-z}} \tag{16}$$

For different return periods, the Gumbel distribution has defined the relationship between p(QM) and z (listed in the Annex: Table II) which can easily lead us to maximum discharges of any return period with solving an equation:

$$Q_{Mp} = Q^* + \frac{1}{a}z \tag{17}$$

Pearson III distribution

The Pearson III distribution is a non-symmetric three-parametric distribution. The original form of this contribution is very complex and using it in practice is time-consuming, so in hydrological problems its modification is more often used proposed by Foster-Ribkin (Annex: Table III) that [8] defined as:

$$Q_{Mp} = (c_v\varphi + 1)\overline{Q_M} \tag{18}$$

$\overline{Q_M}$ = mean value of the distribution

cv = variation coefficient

ϕ = function defined as ϕ = f(cs,p) where cs presents the skew coefficient. Values of ϕ function are listed in Annex: Table IV for different return periods p and skew coefficients cs assuming that the variation coefficient cv = 1.

Testing

The presented frequency distributions will give different maximum discharges for the same return period. An example presented in Figure 2 shows a wide range of Q50 (maximum discharge of 50 years return period), between 2,300 and 2,550 m^3/s.

The decision of the most appropriate, or the most accurate method depends on the result of statistic tests which has to be applied on calculated theoretical distributions to determine if a calculated theoretical distribution matches measured values. There are many statistic tests but, frequently used tests are the Kolmogorov-Smirnov test and the χ^2 test.

Kolmogorov-Smirnov test

The Kolmogorov-Smirnov Goodness-of-Fit Test (K-S test) compares data with a known distribution and lets you know if they have the same distribution. Although the test is nonparametric – it does not assume any particular underlying distribution – it is commonly used as a test for normality to see if your data is normally distributed. More specifically, the test compares a known hypothetical probability distribution (e.g. the normal distribution) to the distribution generated by your data – the empirical distribution function [22].

Measure of tolerance DN is given by equation:

$$D_N = \max\left|\Phi_N(x) - F(x)\right| \quad -\infty < x < +\infty \tag{19}$$

where φ N(x) presents empirical distribution and F(x) is the theoretical distribution.

Predefined confidence level is usually α = 0.05 (5%). Table 2 presents critical values of Do related to the number of data in series (n).

Table 2. Critical values (Do) of K-S test [8]

n	5	10	15	20	25	30	35	40	45	50	>50
Do	0.56	0.41	0.34	0.29	0.27	0.24	0.23	0.21	0.20	0.19	$1.36/n^{1/2}$

If the calculated value of DN<Do theoretical distribution is acceptable. If not, the tested theoretical distribution should be rejected. In a case of testing several distributions, the best one is the one with the smallest DN value.

χ^2 test (Chi-squared test)

The Chi-squared test can also be used to determine how well theoretical distributions, such as normal, binomial, etc. fit empirical distributions, obtained from measured data [10].

In the Chi-squared test, it is necessary to define the nul-hypothesis (Ho), which confirms and the alternative hypothesis (Ha), which rejects the statement. If under hypothesis Ho the computed value of the χ^2 test is greater than some critical value alpha (α), we would reject Ho. Otherwise we would accept it. The critical value is chosen by the researcher. The usual alpha value is 0.05 (5%), but it could also have other levels like 0.01 or 0.10.

In equation:

$$\chi^2 = \sum_j \frac{(o_j - e_j)^2}{e_j} \tag{20}$$

symbols oj and ej are representing respectively observed (measured) data and expected frequency in th j-th cell [10].

Calculation models

In order to perform the flood frequency analysis, the first step is to get the time series of discharges and/or water levels for the hydrological station, which is of interest from water authorities. This type of analysis is usually performed on time series of maximum annual discharges. Then, the record of maximum discharges, which was previously sorted in descending order is fitted with a theoretical distribution using a cumulative density function. This can be done in Excel, but some knowledge of statistic and probability functions is necessary. There are many theoretical distributions integrated within Excel functions. For example, input window for Gamma distribution from Excel is shown in Figure 6.

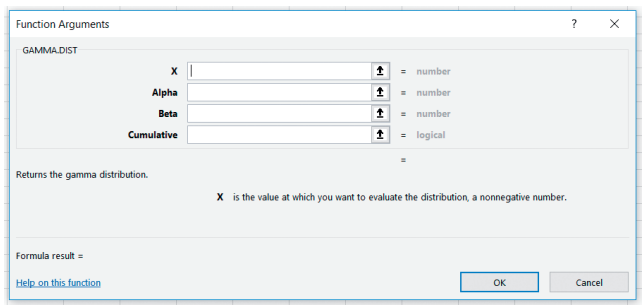

Figure 6. Input parameters for Gamma distribution in Excel (compiled by the authors)

Besides basic functions in Excel, there are many programs or stand-alone applications or add-inns for Excel. One of them is EasyFit. This applications includes goodness-of-fit tests and more than 50 distributions. For each distribution, EasyFit provides several functions to be used in Excel sheets. After the distributions are fitted, EasyFit will display the Fitting Results window (Figure 7) for the distributions comparison and selection of the best model. EasyFit supports all the most popular goodness-of-fit tests, including the Kolmogorov-Smirnov, Anderson-Darling and Chi-squared tests. Once the distributions are fitted, EasyFit displays the goodness-of-fit reports which include the test statistics and critical values calculated for various significance levels.

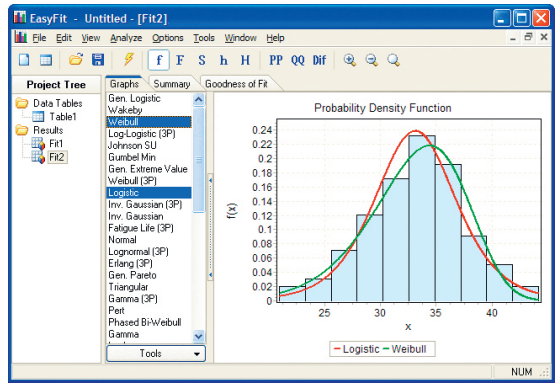

Figure 7. Fitting results window provided by EasyFit (compiled by the authors)

Another software which can be used for distribution fitting is ModelRisk. ModelRisk is a Monte Carlo simulation Excel add-in that allows to include uncertainty in their spreadsheet models. A ModelRisk user replaces uncertain values within their Excel model with special ModelRisk quantitative probability distribution functions that describe the uncertainty about those values. ModelRisk then uses Monte Carlo simulation to automatically generate thousands of possible scenarios. It contains more than 130 probability distributions. The distribution's parameters are estimated using maximum likelihood estimates (MLE) [23].

The fitted distributions are ranked according to the SIC, AIC (Akaike) and HQIC information criteria. For these holds: the lower an information criterion, the better the fit. The advantage of AIC and the other Information Criteria is the fact that they take into account the number of parameters estimated, and penalise for overfitting: a model that has a good fit using fewer parameters is preferred over the one that needs more parameters.[1] The AIC is the least strict of the three in penalising for more parameters, while SIC is the strictest.

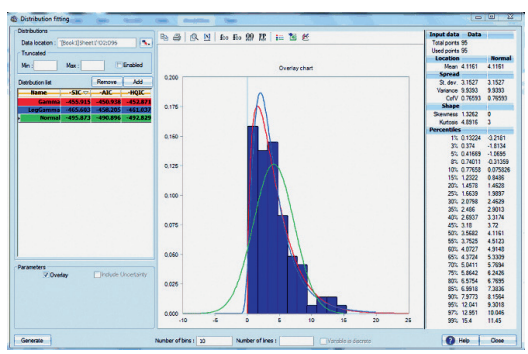

Figure 8. Distribution fit window from ModelRisk (compiled by the authors)

[1] For more information see www.vosesoftware.com/riskwiki/ComparingfittedmodelsusingtheSICHQICorAICinformationcritereon.php

Statgraphics is another software which contains several procedures for manipulating statistical probability distributions. 45 distributions may be plotted, fit to data, and used to calculate critical values or tail areas (Figure 9). Random samples may also be generated from each of the distributions with this stat software. Goodness-of-fit tests used in Statgraphics are ShapiroWilks, the Kolmogorov-Smirnov and Chi-squared tests.

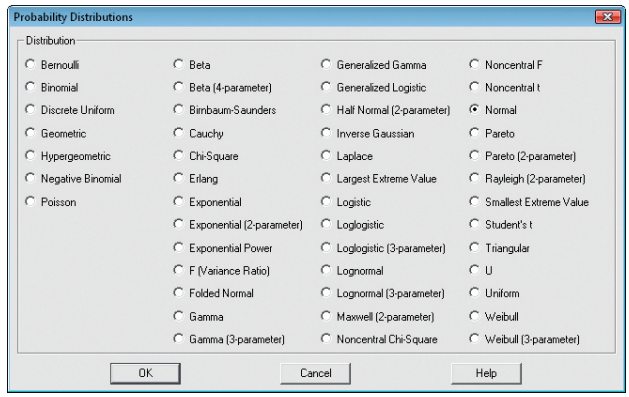

Figure 9. Probability distributions available in Statgraphics (compiled by the author)

These are just some of the software used for distribution fitting. There are also R software, CumFreq, Minitab and others.

When the appropriate function for analysed dataset of maximum annual discharges is obtained, it is necessary to calculate the maximum annual discharge for different return periods. Those values, together with the riverbed and floodplain geometry can be input parameters for hydraulic analysis which can be done in software like HEC-RAS or MIKE. Results of hydraulic analysis can be plotted on maps using GIS software in order to obtain flooded areas according to different return periods, or flood risk maps just like the one shown in Figure 3.

This was done, for example, in the hydrological research of the Kopački rit Nature Park with analysis of its flooding frequency depending on the Danube River water levels and discharges for different return periods [20]. Time series of maximum annual discharges of the Danube River were analysed from the Bezdan station (Serbia) in the period from 1951 to 2008. These data were analysed by several distributions in order to achieve the Danube River discharges and water levels of 5, 10, 25, 50 and 100 return period (Figure 10). As the Log-Pearson III distribution is recommended for flood frequency analysis, it was chosen to be the most appropriate.

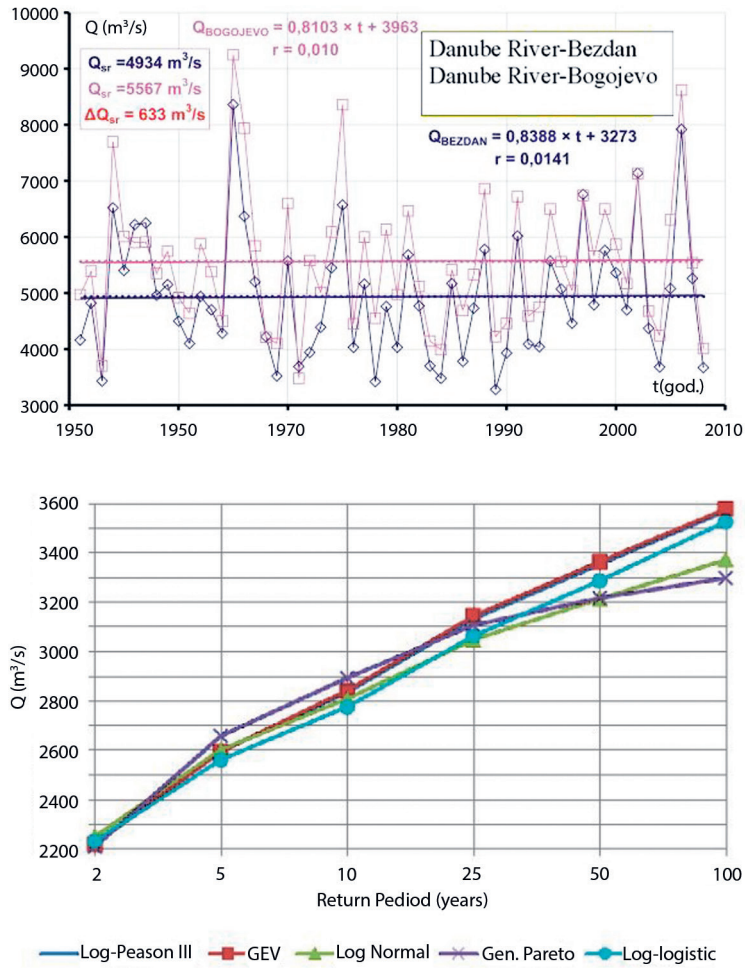

Figure 10. Discharges according to different distributions for different return periods [20]

Hydraulic analysis performed with software HEC-RAS included, besides discharges, water levels at Aljmaš station, discharges of the Drava River, which is the tributary of the Danube River, riverbeds and floodplain geometries and land cover. Obtained results were analysed with GIS in order to see flooded areas according to different return periods. In Figure 11 are shown maps of the flooded area for 5 and 100 years return periods.

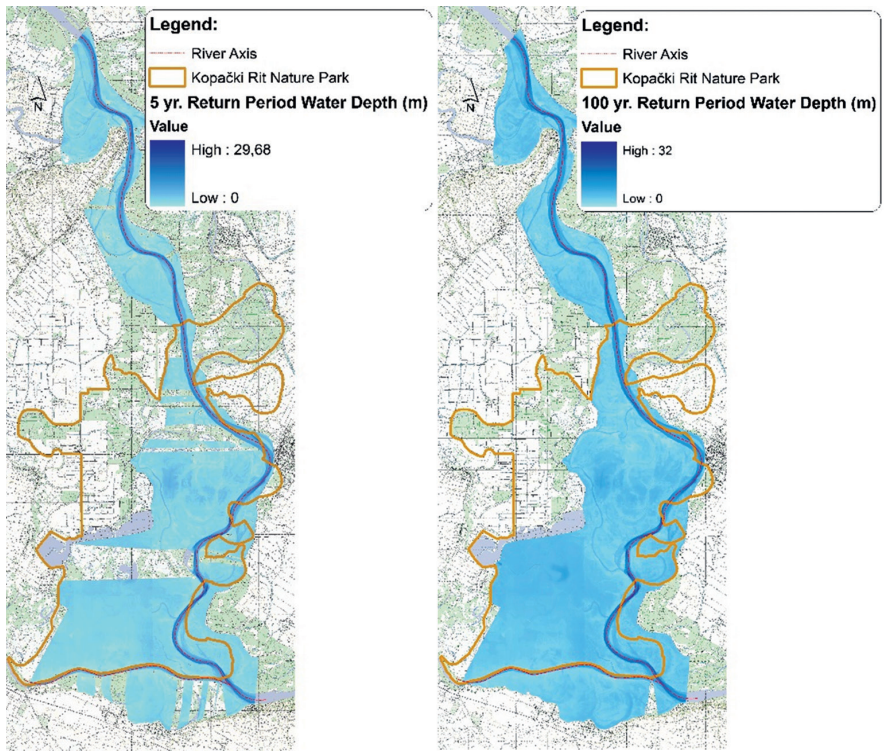

Figure 11. Maps of the flooded area for 5 and 100 years return periods [20]

Flood coincidence

In the previous sections, only the problem of one parameter flood frequency has been analysed, usually the maximum annual water level or discharge assumed to be stochastic. Unfortunately, there are many locations in the world where the flood of one river coincides with the flood of its tributary which enormously increases damages in the given area. In this case, the term "coincidence" presents the occurrence probability of two stochastic events X and Y at the same time (simultaneously), where X presents the event in the main watercourse and Y is the event in its tributary [2].

Only in the Danube River basin there are several potentially critical profiles. In the zone of significant interaction between the mainstream and its tributary, it is recommended to apply flood coincidence methodology which gives a statistically sound analysis concerning an important feature of flood genesis. To judge this, the evaluation of historical data is of great importance. In the case of a complex river system, limited by two inlet profiles (on the mainstream and its tributary) and one outlet profile (on the mainstream), without a significant influence of inflow from the inter-catchment, the relevant combinations of maximum annual discharges and their corresponding (synchronous) discharge values has to be calculated [1].

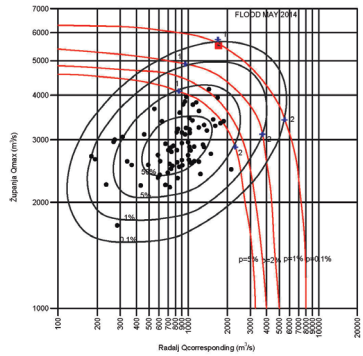

Figure 12. An example of maximum discharges coincidence at Sava and Drina rivers [2]

Figure 12 presents the probabilities of coincidence of two flood waves – on the Sava River (main watercourse) and the Drina River, its tributary. Analysis was done on the basis of historical data. An extreme flood event, occurred in May 2014, was a flood of 1,000 years return period (0.1%).

Statistical calculation is rather complex and it will not be explained in details (see references).

Copulas

The coincidence of two flood waves, of a main river and its tributary, represents a multivariate hydrological event. Another similar problem in flood frequency analysis is to determine the relationship hydrograph – return period. Besides information on flood peak (maximum dischargers), great influence on magnitude of floods have also volume and duration of flood wave. In such cases, the consideration of more than one variable in analyses is reasonable. In order to comprehend and connect these variables, joint cumulative distribution function (cdf) and probability density function (pdf) of involved variables are needed. Because of this, multivariate statistical analyses have to be applied. Some multivariate approaches were introduced in flood frequency analysis during last years, but they all had three limitations [12]:

- all univariate marginal distributions have to belong to the same family, but analysed variables could show different margins
- mathematical formulations become complicated when increasing the number of variables
- it is not possible to distinguish marginal and joint behaviour of studied variables

Copula functions overcome these limitations and present a useful tool in the field of multivariate analyses. The copula actually ‘couples’ the marginal distributions together to form a joint distribution. In analysis of coincidence two flood waves, distribution of maximum discharges of main river represents one marginal distribution and distribution of maximum discharges of tributary represents another one. The copula connects marginal distributions to one joint distribution and gives the probability of their coincidence.

The advantages in using copulas to model joint distributions are [13]:

- flexibility in choosing marginal distributions
- analysis of more than two variables;
- separate analysis of marginal distribution

When analysing two variables, which is the simplest analysis because of the small number of variables, bivariate copula is used. A bivariate copula C is the joint distribution function of two uniform random variables and can be written as [14] [15] [18]:

$$C: [0.1]2 \rightarrow [0.1] \tag{21}$$

Two following conditions must be fulfilled: $C(1,u) = C(u,1) = u$ and $C(u,0) = C(0,u) = 0$ and the second one $C(u1,u2) + C(v1,v2) – C(u1,u2) – C(v1,v2) \geq 0$ if $u1 \geq v1$, $u2 \geq v2$ and $u1$, $u2$, $v1$, $v2 \in [0.1]$. The link between copula and the joint distributions is based on the theorem of Sklar:

$$F_{X,Y}(x,y) = C[F_X(x), F_Y(y)] \tag{22}$$

where

$F_{X,Y}(x,y)$ are the joint cumulative distribution function of the random variables, and F_X, F_Y are marginal distribution functions.

There are two groups of copulas: elliptical and the Archimedean family. Elliptical copulas are the copulas with elliptical distributions, which have an elliptical form and therefore symmetry in the tails. Important copulas in this family are the Gaussian and the student's copula. The Gaussian copula is often used because of his simple form. Archimedean copulas are widely applied, because they are not difficult to construct. Archimedean copulas have only one dependency parameter, instead of a dependency matrix. The most important Archimedean copulas are Gumbel, Clayton and Frank [16]. Different types of copulas are shown in Figure 13.

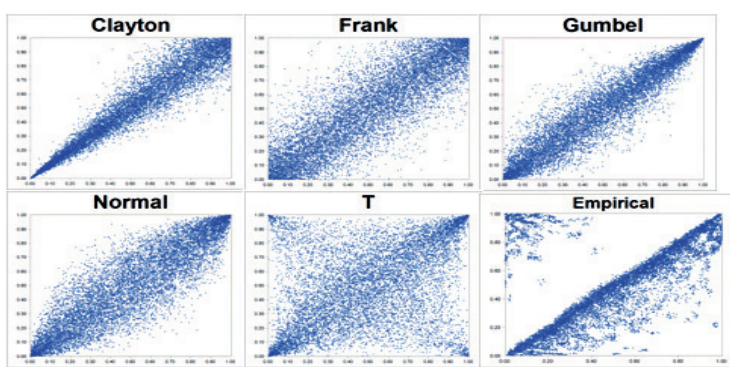

Figure 13. Some types of copula functions [17]

It is evident from Figure 13 that different copulas have different appearances and characteristics. The main characteristic is tail dependence. It is a very important feature of copulas, which has a great effect on how well the joint distribution captures the behaviour of extreme events. The tail dependence is a measure of extreme correlation, which marks the probability of an extreme occurrence for a variable when the extreme value of another variable occurs. When the variables of the marginal distribution in the upper tail (top right corner) of a copula are dependent on each other, we can say that the upper tail of the copula is dependent. From Figure 13 it can be seen that the Clayton copula can capture only lower tail dependence, the Frank copula family cannot exhibit any tail dependence and the Gumbel copula can only capture upper tail dependence [19].

There are different criteria in order to determine which copula is better suited for the analysed problem. The Kolmogorov-Smirnov test and the $\chi 2$ test, which are previously mentioned and explained, can be used. The Anderson-Darling goodness-of-fit and the Bayesian copula selection method can also be mentioned. Statistical measures of fit called information criteria such as the Schwarz Information criterion (SIC), known as the Bayesian information criterion or BIC, Akaike information criterion (AIC) and Hannan-Quinn information criterion (HQIC) can also be used [17].

As mentioned earlier, flood frequency analysis based on copula is mostly used in areas where confluences of tributaries can be found. For example, in [11] bivariate frequency analysis using a copula function is used to calculate the probability of coincidence of maximum water levels in the Drava and the Danube rivers. Results showed a 0.7% probability that highest water levels occur simultaneously in both rivers, which can be seen in the upper right corner of Figure 14. This is important information for future flood risk management because their coincidence would be disastrous for citizens in surrounding areas.

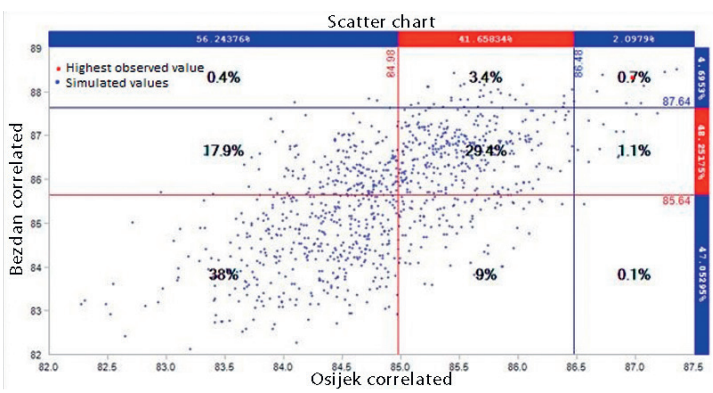

Figure 14. Probability of coincidence of maximum water levels of the Drava and Danube rivers [11]

Another example is from the Sava River basin [21]. The Sava River is the right-hand tributary of the Danube. Two tributaries of the Sava River were selected for analyses and it was shown that the proposed copula approach estimates recent flood events more accurately than the univariate flood frequency analysis based on the observation data.

Case study

In order to show how to fit distribution to observed discharges and how to calculate discharges for different return periods, time series of the Danube River discharges measured at Bogojevo station in Serbia will be analysed.

Maximum annual discharges of the Danube River in the period from 1950 to 2017 are analysed (Figure 15). Basic statistical parameters of this time series are shown in Table 3. Measured values for the years 1996, 1997 and 2010 are missing, so frequency analysis is done without them.

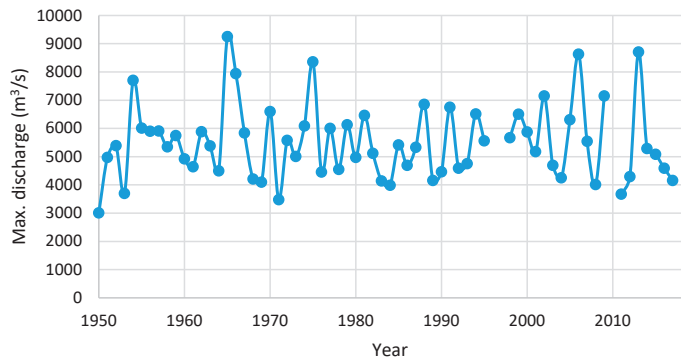

Figure 15. Maximum annual discharge of the Danube River at Bogojevo station (Serbia) (compiled by the authors)

Table 3. Basic statistical parameters of maximum annual discharges (compiled by the authors)

Number of observations	65
Mean value	5,493.862
Minimum value	3,010
Maximum value	9,250
Sum	357,101
Standard deviation	1,325.226
Variance	1,756,224
Skew	0.82733

The next step is to put discharge values in descending order and assign each value a rank. The highest value (in this case 9,250) has rank 1 and the smallest one, 3,010, has rank 65 which is the total number of values. After this, it is possible to fit distributions to observed time series.

Two empirical distributions, Weibull and Čegodajev, and three theoretical are calculated. Normal and Gamma distributions are calculated using Excel functions only and the Log-Pearson III distribution is determined also in Excel but with formulas and coefficients for this distribution.

Empirical distribution

Probabilities of occurrence for Weibull and Čegodajev distribution are calculated according to formulas shown in Table 1.

Results are shown in Figure 16. Probabilities of occurrence according to these two empirical distributions are the same.

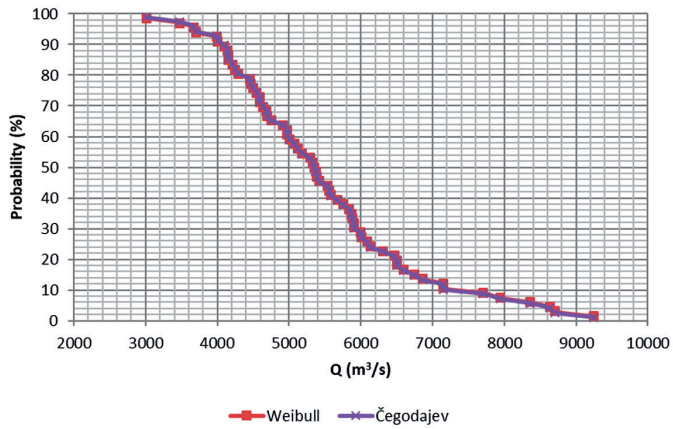

Figure 16. Weibull and Čegodajev distributions

Theoretical distribution

Log-Pearson III distribution is highly recommended for flood frequency analysis. The main advantage of this method is successful application on relatively short time series in order to obtain floods for much longer return periods. The equation and parameters of Log-Pearson III distribution are:

$$log\, x_{PR} = \overline{log\, x} + K \cdot \sigma\, log\, x \tag{23}$$

where

x_{PR} – is the variable value relevant for different return periods,

x – is the random variable (discharge, water level),

$\overline{\log x}$ – is the mean value of logarithms of random variables,

K – is the frequency coefficient; it is in a function of the skewness coefficient Cs and return period (Annex: Table V),

σ – is the standard deviation.

This distribution is based on logarithmic values of discharges, and not discharges themselves, so the first step is to calculate logarithmic values of discharges and then the skewness coefficient Cs and standard deviation of this logarithmic time series. After this, it is easy to calculate discharges for different return periods. Results are shown in Table 4 and in Figures 17 and 18.

Table 4. Results of Log-Pearson III distribution (compiled by the authors)

Probability	Return period	K1 (C = 0.1)	K2 (C = 0.2)	K	logQ	Q(m^3/s)
100	1	–2.252	–2.178	–2,1826	3.5065670	3210.45848
50	2	–0.017	–0.033	–0.0319	3.7247979	5306.37492
20	5	0.836	0.83	0.83037	3.8123042	6490.89067
10	10	1.292	1.301	1.30043	3.8600018	7244.39027
4	25	1.785	1.818	1.81592	3.9123096	8171.64876
2	50	2.107	2.159	2.15573	3.9467904	8846.88533
1	100	2.4	2.472	2.46747	3.9784235	9515.32264

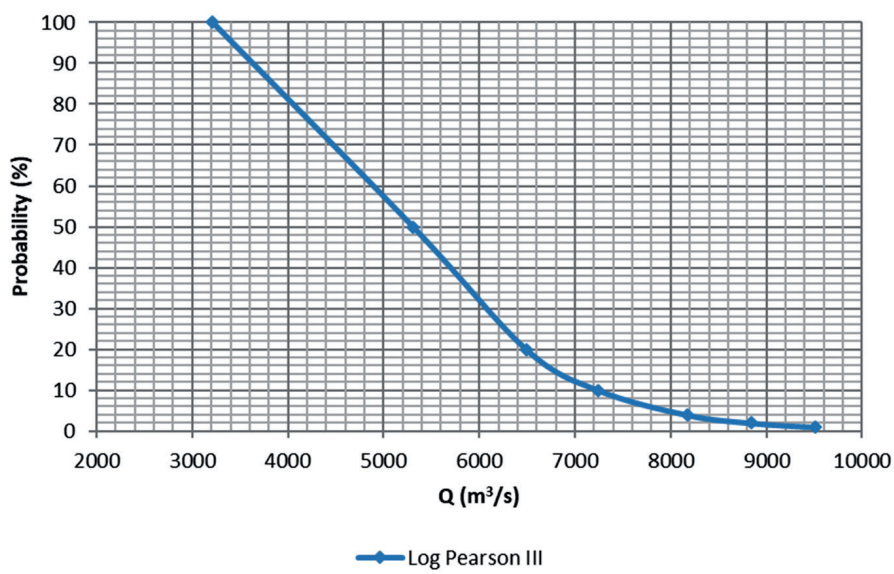

Figure 17. Log-Pearson III distribution (compiled by the authors)

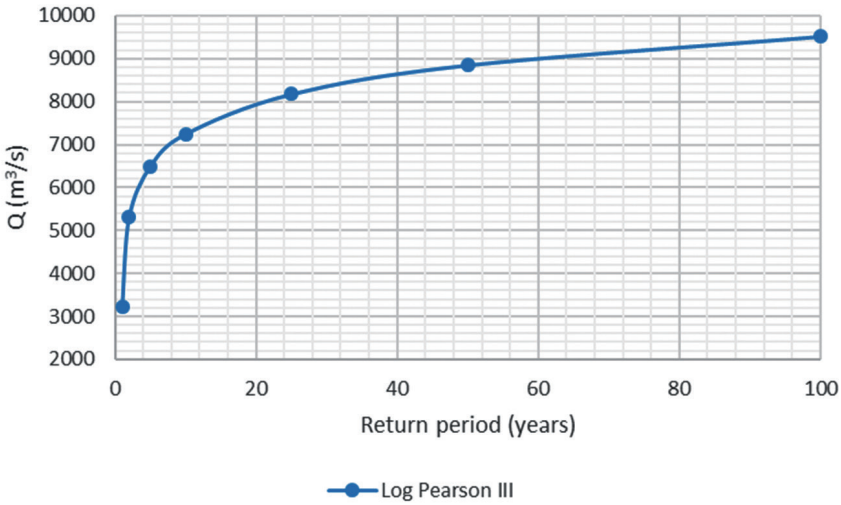

Figure 18. Discharges for different return periods according to Log-Pearson III distribution (compiled by the authors)

Normal distribution is calculated using Excel function NORM.DIST. To use this function, it is necessary to calculate the first mean value and standard deviation of the observed time series. This is already done (Table 3), so below in Figure 19 are the results.

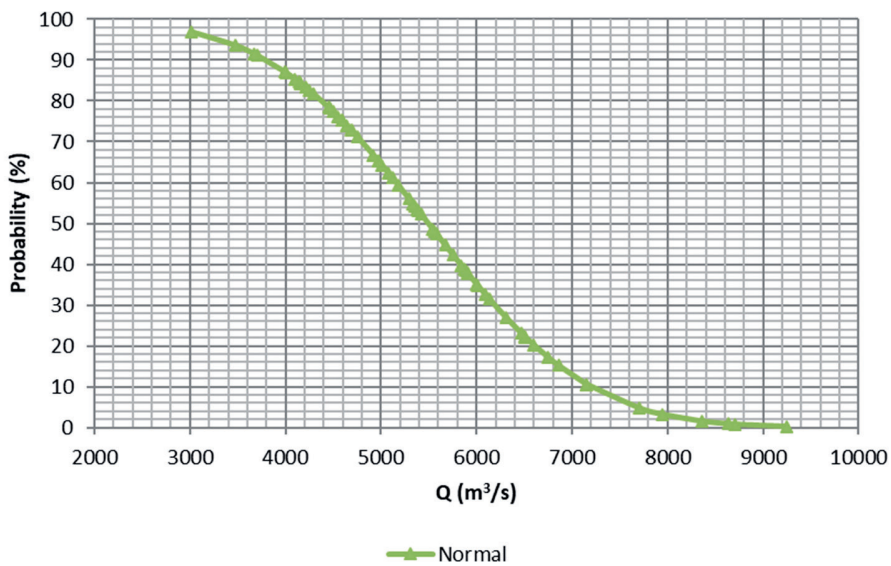

Figure 19. Normal distribution (compiled by the authors)

Gamma distribution is also determined with Excel function for this distribution, but first it is necessary to calculate the α and β parameters:

$$\alpha = E(x)2/Var \quad (24)$$

$$\beta = Var/\ E(x) \quad (25)$$

where

E(x) – is the expected value (mean),

Var – is the variance.

Results of the Gamma distribution are shown in Figure 20.

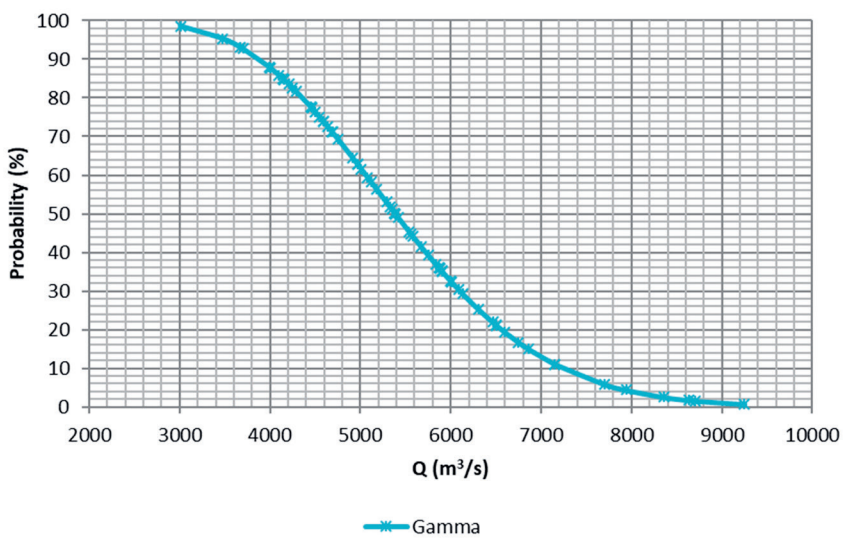

Figure 20. Gamma distributions (compiled by the authors)

All theoretical distributions are compared with empirical ones with goodness-of-fit Chi-squared test. When using this test as an Excel function, two sets must be selected: actual and expected. In this case, the actual range is empirical distribution and the expected range is a theoretical one. Obtained results showed that all distributions are a good fit to both empirical ones. When all distributions are plotted together (Figure 21), this good fit can be seen also.

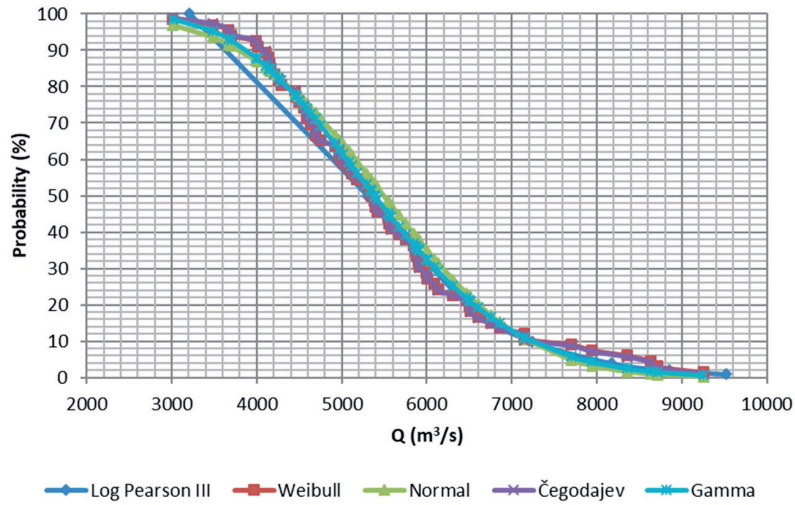

Figure 21. All distributions plotted together (compiled by the authors)

Final remarks

Flood frequency analysis briefly described in the previous chapters presumes the availability of reliable and sufficiently long hydrological data records of maximum discharges or water levels.

Very often, especially on small rivers (catchments) there are no measured data with such characteristics. These catchments are called ungauged and the procedure of flood frequency determination differs compared to these, because the only measured parameter is precipitation. It is necessary, therefore, to apply some of the various hydrological models, rainfall-runoff models (RFRO) or statistical models to determine the hydrological parameters of the flood waves.

Flood is a stochastic event and statistics is the only tool used to give us as much accurate probability of its occurrence as possible. There are many results of methods (empirical and theoretical) but none of them is absolutely accurate.

Besides, each catchment is changing in time, even if there are no drastic human interventions, which additionally makes our calculations and prognosis more complex.

References

[1] Prohaska S, Ilić A (2010). Coincidence of Flood Flow of the Danube River and Its Tributaries. In: Brilly M editor. Hydrological Processes of the Danube River Basin. Perspectives from the Danubian Countries. Dordrecht: Springer. p. 389–428.

[2] Prohaska S, Ilić A (2016). Coincidence of Flood Waves in the Sava and Drina Rivers. Hrvatske vode 24(95):1–18.

[3] EIB Flood Review (2007). Guide for Preparation of Flood Risk Management Schemes. Available from: www.eib.org/attachments/strategies/flood_risk_management_guide_en.pdf
[4] Danube Flood Risk Atlas (2012). Danube Atlas – Flood Hazard and Risk Maps. Available from: www.icpdr.org/main/activities-projects/danube-floodrisk-project
[5] EU Flood Directive (2007). Directive/2007/60/EC. Available from: www.voda.hr/sites/default/files/council_directive_2007-60-ec.pdf
[6] Makkonen L (2005). Plotting Positions in Extreme Value Analysis. J App Meteor Climat. 45(2):334–340. DOI: https://doi.org/10.1175/JAM2349.1
[7] Chow VT, Maidment DR, Mays LW (1988). Applied Hydrology. New York: McGraw-Hill.
[8] Žugaj R (2000). Hidrologija. RGN: Sveučilište u Zagrebu.
[9] Maidment DR (1992). Handbook of Hydrology. New York: McGraw-Hill.
[10] Spiegel MR (1972). Theory and Problems of Statistics. Schaum's Outline Series. New York: McGraw-Hill.
[11] Tadić L, Bonacci O, Dadić T (2016). Analysis of the Drava and Danube Rivers Floods in Osijek (Croatia) and Possibility of Their Coincidence. Environmental Earth Sciences 75. DOI: https://doi.org/10.1007/s12665-016-6052-0
[12] Grimaldi S, Serinaldi F (2006). Asymmetric Copula in Multivariate Flood Frequency Analysis. Adv W Res. 29(8):1155–1167. DOI: https://doi.org/10.1016/j.advwatres.2005.09.005
[13] Li T, Guo S, Chen L, Guo J (2013). Bivariate Flood Frequency Analysis with Historical Information Based on Copula. Journal of Hydrologic Engineering, 18(8):1018–1030.
[14] Šraj M, Bezak N, Brilly M (2014). Bivariate Flood Frequency Analysis Using the Copula Function: A Case Study of the Litija Station on the Sava River. Hydr Process. 29(2):225–238. DOI: https://doi.org/10.1002/hyp.10145
[15] Klein B, Pahlow M, Hundecha Y, Schumann A (2010). Probability Analysis of Hydrological Loads for the Design of Flood Control Systems Using Copulas. Journal of Hydrologic Engineering, 15(5):360–369. DOI: https://doi.org/10.1061/(ASCE)HE.1943-5584.0000204
[16] Habiboellah F (2007). Modeling Dependencies in Financial Risk Management. Master thesis, Vrije Universiteit, Amsterdam.
[17] Vose D (2010). Fitting Distributions to Data and Why You Are Probably Doing It Wrong. Available from: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.695.3357&rep=rep1&type=pdf
[18] Karmakar S, Simonovic SP (2007). Flood Frequency Analysis Using Copula with Mixed Marginal Distributions. Project Report No. 055, University of Western Ontario, Department of Civil and Environmental Engineering.
[19] Huang SH, Hou B, Chang J, Huang Q, Chen Y (2014). Copulas-based Probabilistic Characterization of the Combination of Dry and Wet Conditions in the Guanzhong Plain, China. Journal of Hydrology 519:3204–3213. DOI: https://doi.org/10.1016/j.jhydrol.2014.10.039
[20] Tadić L, Dadić T, Barač B (2013). Flood Frequency Modelling of the Kopački rit Nature Park. Tehnički vjesnik 20(1):51–57.
[21] Gilja G, Ocvirk E, Kuspilić N (2018). Joint Probability Analysis of Flood Hazard at River Confluences Using Bivariate Copulas. Građevinar 70(4):267–275. DOI: https://doi.org/10.14256/JCE.2173.2017
[22] The Kolmogorov-Smirnov Goodness-of-Fit Test. Available from: www.statisticshowto.com/kolmogorov-smirnov-test/
[23] Maximum Likelihood Estimates (MLEs). Available from: www.vosesoftware.com/riskwiki/MaximumLikelihoodEstimatesMLEs.php
[24] Haan, CT (1977). Statistical Methods in Hydrology. Iowa State University Press.

Annex

Table I. Log normal (Galton)distribution – $z\overline{\sigma}$ values for different return periods [8]

Return period (year)	Probability of occurrence p (%)	z	$z\overline{\sigma}$
10,000	0.01	3.715	0.6901
1,000	0.1	3.090	0.5740
100	1	2.326	0.4321
50	2	1.054	0.3816
25	4	1.752	0.3255
10	10	1.281	0.2380
5	20	0.842	0.1564
2	50	0.000	0.000
1.25	80	–0.842	–0.1564
1.1111	90	–1.281	–0.2380
1.0417	96	–1.752	–0.3255
1.0204	98	–2.054	–0.3816
1.0101	99	–2.326	–0.4321
1.0010	99.9	–3.090	–0.5740
1.0001	99.99	–3.715	–0.6901

Table II. Gumbel distribution – "z" values for different return periods [8]

Return period (year)	Probability of occurrence p (%)	P1	z	$\frac{1}{a}z$
10,000	0.01	0.9999	9.21	253.3
1,000	0.1	0.999	6.91	190
100	1	0.99	4.60	126.5
50	2	0.98	3.91	107.5
25	4	0.96	3.20	88.0
10	10	0.90	2.25	61.9
5	20	0.80	1.50	41.3
2	50	0.50	0.37	10.2
1.25	80	0.20	–0.48	–13.2
1.1111	90	0.10	–0.83	–22.8
1.0417	96	0.04	–1.15	–31.6
1.0204	98	0.02	–1.35	–37.1
1.0101	99	0.01	–1.53	–42.1
1.0010	99.9	0.001	–1.94	–53.4
1.0001	99.99	0.0001	–2.20	–60.5

Table III. Pearson III distribution – Probability of occurrence p (%), according to Foster-Ribkin [8]

czM	Probability of occurrence p (%)														
	0.01	**0.1**	**1**	**2**	**4**	**10**	**20**	**50**	**80**	**90**	**96**	**98**	**99**	**99.9**	**99.99**
–2.0		1.00	0.99	0.96	0.96	0.90	0.78	0.31	–0.6	–1.3	–2.3				
–1.5		1.31	1.26	1.23	1.15	1.02	0.82	0.24	–0.7	–1.3	–2.0				
–1.0		1.79	1.59	1.37	1.22	1.13	0.85	0.16	–0.8	–1.3	–2.0				
–0.8		2,02	1.74	1.65	1.42	1.17	0.85	0.13	–0.8	–1.3	–1.9				
–0.6		2.27	1.88	1.76	1.51	1.20	0.85	0.10	–0.8	–1.3	–1.9				
–0.4		2.54	2.03	1.90	1.60	1.23	0.85	0.07	–0.8	–1.3	–1.8				
–0.2		2.81	2.18	1.98	1.67	1.26	0.85	0.03	–0.8	–1.3	–1.8				
0	3.72	3.09	2.33	2.04	1.75	1.28	0.84	0.00	–0.9	–1.3	–1.7				
0.2	4.16	3.38	2.47	2.16	1.81	1.30	0.83	0.00	–0.9	–1.3	–1.6				
0.4	4.61	3.66	2.61	2.26	1.87	1.32	0.82	–0.1	–0.9	–1.2	–1.5				
0.6	5.05	3.96	2.76	2.35	1.94	1.33	0.80	–0.1	–0.9	–1.2	–1.5				
0.8	5.50	4.24	2.89	2.45	2.00	1.34	0.78	–0.1	–0.9	–1.2	–1.4				
1.0	5.96	4.53	3.02	2.54	2.05	1.34	0.76	–0.2	–0.9	–1.1	–1.3				
1.2	6.41	4.81	3.15	2.62	2.09	1.34	0.73	–0.2	–0.8	–1.1	–1.3				
1.5	7.09	5.26	3.33	2.74	2.15	1.33	0.69	–0.2	–0.8	–1.0	–1.1				
2.0	8.21	5.91	3.60	2.91	2.23	1.30	0.61	–0.3	–0.8	–0.9	–1.0				
2.5	9.30	6.60	3.83	3.04	2.28	1.24	0.53	–0.4	–0.7	–0.8	–0.8				
3.0	10.4	7.25	4.02	3.16	2.30	0.42	0.42	–0.4	–0.6	–0.7	–0.7	–0.7	–0.7	–0.7	–0.7

Table IV. Values of ϕ function in Pearson III distribution [8]

Return period (year)	Probability of occurrence p (%)	ϕ
10,000	0.01	5.50
1,000	0.1	4.24
100	1	2.89
50	2	2.45
25	4	2.00
10	10	1.34
5	20	0.78
2	50	–0.13
1.25	80	–0.86
1.1111	90	–1.17
1.0417	96	–1.47
1.0204	98	–1.60
1.0101	99	–1.74
1.0010	99.9	–2.02
1.0001	99.99	–2.18

Table V. Frequency factors K for Gamma and Log-Pearson type III distributions [24]

	Recurrence interval in years							
	1.0101	2	5	10	25	50	100	200
Skew coefficient	Percent chance (> =) = 1-F							
Cs	99	50	20	10	4	2	1	0.5
3	–0.667	–0.396	0.42	1.18	2.278	3.152	4.051	4.97
2.9	–0.69	–0.39	0.44	1.195	2.277	3.134	4.013	4.904
2.8	–0.714	–0.384	0.46	1.21	2.275	3.114	3.973	4.847
2.7	–0.74	–0.376	0.479	1.224	2.272	3.093	3.932	4.783
2.6	–0.769	–0.368	0.499	1.238	2.267	3.071	3.889	4.718
2.5	–0.799	–0.36	0.518	1.25	2.262	3.048	3.845	4.652
2.4	–0.832	–0.351	0.537	1.262	2.256	3.023	3.8	4.584
2.3	–0.867	–0.341	0.555	1.274	2.248	2.997	3.753	4.515
2.2	–0.905	–0.33	0.574	1.284	2.24	2.97	3.705	4.444
2.1	–0.946	–0.319	0.592	1.294	2.23	2.942	3.656	4.372
2	–0.99	–0.307	0.609	1.302	2.219	2.912	3.605	4.298
1.9	–1.037	–0.294	0.627	1.31	2.207	2.881	3.553	4.223
1.8	–1.087	–0.282	0.643	1.318	2.193	2.848	3.499	4.147
1.7	–1.14	–0.268	0.66	1.324	2.179	2.815	3.444	4.069
1.6	–1.197	–0.254	0.675	1.329	2.163	2.78	3.388	3.99
1.5	–1.256	–0.24	0.69	1.333	2.146	2.743	3.33	3.91
1.4	–1.318	–0.225	0.705	1.337	2.128	2.706	3.271	3.828
1.3	–1.383	–0.21	0.719	1.339	2.108	2.666	3.211	3.745
1.2	–1.449	–0.195	0.732	1.34	2.087	2.626	3.149	3.661
1.1	–1.518	–0.18	0.745	1.341	2.066	2.585	3.087	3.575
1	–1.588	–0.164	0.758	1.34	2.043	2.542	3.022	3.489
0.9	–1.66	–0.148	0.769	1.339	2.018	2.498	2.957	3.401
0.8	–1.733	–0.132	0.78	1.336	1.993	2.453	2.891	3.312
0.7	–1.806	–0.116	0.79	1.333	1.967	2.407	2.824	3.223
0.6	–1.88	–0.099	0.8	1.328	1.939	2.359	2.755	3.132
0.5	–1.955	–0.083	0.808	1.323	1.91	2.311	2.686	3.041
0.4	–2.029	–0.066	0.816	1.317	1.88	2.261	2.615	2.949
0.3	–2.104	–0.05	0.824	1.309	1.849	2.211	2.544	2.856
0.2	–2.178	–0.033	0.83	1.301	1.818	2.159	2.472	2.763
0.1	–2.252	–0.017	0.836	1.292	1.785	2.107	2.4	2.67
0	–2.326	0	0.842	1.282	1.751	2.054	2.326	2.576
–0.1	–2.4	0.017	0.846	1.27	1.716	2	2.252	2.482
–0.2	–2.472	0.033	0.85	1.258	1.68	1.945	2.178	2.388
–0.3	–2.544	0.05	0.853	1.245	1.643	1.89	2.104	2.294
–0.4	–2.615	0.066	0.855	1.231	1.606	1.834	2.029	2.201
–0.5	–2.686	0.083	0.856	1.216	1.567	1.777	1.955	2.108
–0.6	–2.755	0.099	0.857	1.2	1.528	1.72	1.88	2.016
–0.7	–2.824	0.116	0.857	1.183	1.488	1.663	1.806	1.926

	Recurrence interval in years							
	1.0101	2	5	10	25	50	100	200
Skew coefficient	Percent chance (> =) = 1-F							
Cs	99	50	20	10	4	2	1	0.5
–0.8	–2.891	0.132	0.856	1.166	1.448	1.606	1.733	1.837
–0.9	–2.957	0.148	0.854	1.147	1.407	1.549	1.66	1.749
–1	–3.022	0.164	0.852	1.128	1.366	1.492	1.588	1.664
–1.1	–3.087	0.18	0.848	1.107	1.324	1.435	1.518	1.581
–1.2	–3.149	0.195	0.844	1.086	1.282	1.379	1.449	1.501
–1.3	–3.211	0.21	0.838	1.064	1.24	1.324	1.383	1.424
–1.4	–3.271	0.225	0.832	1.041	1.198	1.27	1.318	1.351
–1.5	–3.33	0.24	0.825	1.018	1.157	1.217	1.256	1.282
–1.6	–3.88	0.254	0.817	0.994	1.116	1.166	1.197	1.216
–1.7	–3.444	0.268	0.808	0.97	1.075	1.116	1.14	1.155
–1.8	–3.499	0.282	0.799	0.945	1.035	1.069	1.087	1.097
–1.9	–3.553	0.294	0.788	0.92	0.996	1.023	1.037	1.044
–2	–3.605	0.307	0.777	0.895	0.959	0.98	0.99	0.995
–2.1	–3.656	0.319	0.765	0.869	0.923	0.939	0.946	0.949
–2.2	–3.705	0.33	0.752	0.844	0.888	0.9	0.905	0.907
–2.3	–3.753	0.341	0.739	0.819	0.855	0.864	0.867	0.869
–2.4	–3.8	0.351	0.725	0.795	0.823	0.83	0.832	0.833
–2.5	–3.845	0.36	0.711	0.711	0.793	0.798	0.799	0.8
–2.6	–3.899	0.368	0.696	0.747	0.764	0.768	0.769	0.769
–2.7	–3.932	0.376	0.681	0.724	0.738	0.74	0.74	0.741
–2.8	–3.973	0.384	0.666	0.702	0.712	0.714	0.714	0.714
–2.9	–4.013	0.39	0.651	0.681	0.683	0.689	0.69	0.69
–3	–4.051	0.396	0.636	0.66	0.666	0.666	0.667	0.667