Catalytics Asia Key Insights Series

Mathematically Modeling Flood Frequency and Intensity

FromHistorical, Incomplete and Uncertain Records

(Rev. 1.03)

Catalytics White Paper Written By:

Prof. Andrzej Kijko,
University of Pretoria,
Natural Hazard Centre,
Africa

CATALYTICS PTE LTD
T +65 6248 4738 | F +65 6248 4531 | www.catalytics.asia
UOB Plaza 1, 80 Raffles Place #36-00, Singapore 048624
Business Registration Number: 200918237

Abstract

In order to model flood losses it is necessary to calculate the frequency of large storms and their duration and intensity. Coherent weather measurement is a recent phenomenon and even so measurement has its flaws. For this reason advanced statistical tools are required to evaluate what the extreme values are of these three variables. Catalytics' methodology is laid out in this paper.

The information contained in this document represents the current view of the author on the issues discussed at of the date of publication. Because Catalytics must respond to changing market conditions, no information in this paper should be interpreted to be a commitment on the part of the company, and it does not guarantee the accuracy of any information presented after the date of publication.

Gender disclaimer: Unless otherwise stated, whenever the masculine gender is used, both men and women are included.

This White Paper is for informational purposes only. Catalytics MAKES NO WARRANTIES, EXPRESS OR IMPLIED, IN THIS DOCUMENT.

CATALYTICS PTE LTD
T +65 6248 4738 | F +65 6248 4531 | www.catalytics.asia
UOB Plaza 1, 80 Raffles Place #36-00, Singapore 048624
Business Registration Number: 200918237

Introduction

This paper will assume that the probabilistic flood hazard analysis (PFHA) for a specified site can be characterized by: (1) the mean flood activity rate λ, (2) the distribution of flood level $F_M(m)$ and (3) the site, or gauging station-characteristic, maximum possible flood level $m_{\max}$.

According to above parameterization, the flood activity rate λ describes the mean number of floods expected to occur within the vicinity of the specified site with water level equal or larger than $m_{\min}$, within specified a time interval, usually 1 year. While this paper acknowledges that are no constrains regarding the distribution of the flood level $F_M(m)$, for purposes of illustration it will assume that the functional form of $F_M(m)$is known and has form of negative exponential distribution with an unknown parameter β_{F}. From the physical point of view, the parameter β_{F} describes the ratio between low and high level flood occurrences. The physical mining of the maximum possible flood water level $m_{\max}$ is the same as for example, meaning of the regional characteristic, maximum possible seismic event magnitude $m_{\max}$.

It must be clearly stated, that the proposed procedure of flood hazard assessment is generic and can be applied virtually to any temporal and flood size distributions.

Nature of Input Data

The lack or incompleteness of data in flood records is a frequent issue in a statistical analysis of flood hazard. Contributing factors include the historical and socio–economic context and demographic variations. In general and in most cases, the degree of completeness is a monotonically increasing function of time, i.e. the more recent portion of the catalogue is more complete and includes smaller floods. Our approach and its associated mathematical formalism makes provision for the floods record to contain three types of data: (1) very large prehistoric floods (paleo-floods), dated over the last thousands of years; (2) the deepest historic floods which occurred in the course of the last few hundred years; and (3) complete recent data for a relatively short period of time. Often, the complete part of the record can itself be divided into several sub-records, each of which contains complete flood data above a given flood level $m_{\min}^{(\mathrm{j})}$, and occurring in a certain period of time T_{j} where $j = 1, \dots, s$ and s is the number of complete flood sub-records. Uncertainty in knowledge of flood water level m can also be taken into account by assuming that the observed flood level is also an unknown and that the ‘true’ level is subjected to a random error that follows a mathematical distribution having zero mean and a known standard deviation such as the Gaussian distribution (Eadie *et al.*, 1971; Tinti and Mulargia, 1985). The schematic illustration of such a flood database is illustrated in Figure 1.

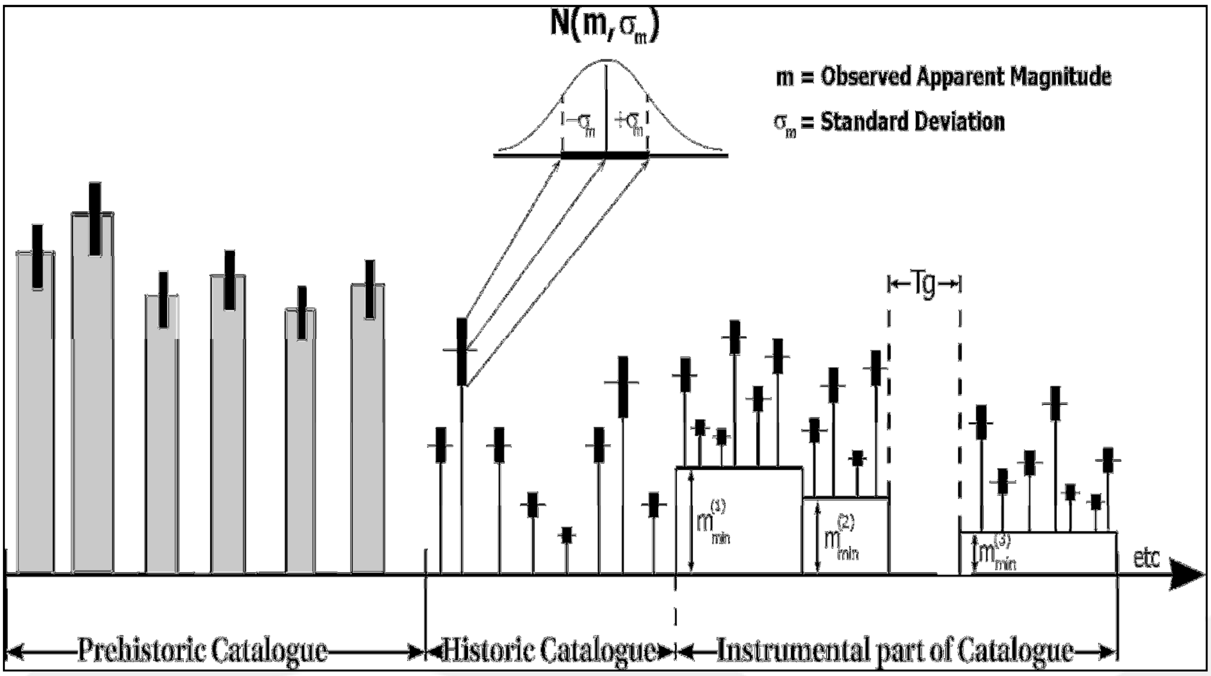

Figure 1. Illustration of data which can be used to obtain reccurence parameters of the proposed flood model. The applied approach permits the combination of the largest (prehistoric/paleo- and historic floods) and complete flood records having variable level of completeness. It accepts 'gaps' (T_g) when records were missing. The procedure is capable of accounting for uncertainties of occurrence time of paleo-floods. Uncertainty in flood records is also taken into account. In that an assumption is made that the observed flood water level is unknown, the "true" water level is subjected to a random error that follows a Gaussian distribution having zero mean and a known standard deviation. (Modified after Kijko and Sellevoll, 1992)

In addition to account of incompleteness and uncertainties of the flood data, the ideal methodology for flood hazard assessment takes into account the inevitable discrepancy between the data and the model describing the flood occurrence. The basic statistical tools applied in the development of such a methodology are briefly described in the following section.

Temporal distribution of floods

It is reasonable to assume that the temporal distribution of the floods level observed in close vicinity of a selected site can be modeled by a Poisson process (Cramér, 1961). Following the above assumption, the probability that at specified site, within specified time interval t, n floods will be observed is

$$p(n|\lambda,t) = P(N = n|\lambda,t) = \frac{(\lambda t)^{\mathrm{n}}}{n!}\mathrm{e}^{-\lambda t} n = 0,1,2 \dots \tag{1}$$

where $\lambda \equiv \lambda(m_{\min})$ denotes the mean occurrence rate of floods with an water level greater than or equal to$m_{\min}$.

Distribution of flood level records

Assuming that flood levels recorded within close vicinity of the selected site or gauging station, are independent, random values distributed according to the cumulative distribution function (CDF) $F_{\mathrm{M}}(m)$ and only for purposes of illustration, assuming that distribution $F_{\mathrm{M}}(m)$ is a shifted negative exponential distribution which is truncated at its top end by the physical characteristics of the site at the level of the maximum possible flood level $m_{\max}$. Acceptance of this assumption is equivalent to the statement that the probability distribution function (PDF) and the cumulative distribution function of flood level *m* are equal to

$$f_M(m|m_{\min}) = \begin{cases} 0 & \text{for} m_{\min} < m \\ \dfrac{\beta_{\mathrm{F}} \exp[-\beta_{\mathrm{F}}(m - m_{\min})]}{1 - \exp[-\beta_{\mathrm{F}}(m_{\max} - m_{\min})]} & \text{for } m \leq m \leq m_{\max} \\ 0 & \text{for} m > m_{\max} \end{cases} \tag{2}$$

and

$$F_{\mathrm{M}}(m|m_{\min}) = \begin{cases} 0 & \text{for} m < m_{\min} \\ \dfrac{1 - \exp[-\beta_{\mathrm{F}}(m - m_{\min})]}{1 - \exp[-\beta_{\mathrm{F}}(m_{\max} - m_{\min})]} & \text{for} m_{\min} \leq m \leq m_{\max} \\ 1 & \text{for} m > m_{\max}, \end{cases} \tag{3}$$

where β_{F} is the parameter. In (2) and (3) the flood level *m* is considered a continuous variable that may assume any value between $m_{\min}$ and the maximum possible flood water level $m_{\max}$ dictated by the site's characteristics.

The functional form of distributions (2)-(3) is well known. This is special case of the gamma distribution and since it describes events recurring 'at random in time' (Johnson and Kotz, 1970). It is widely applied in reliability theory (Barlow and Proschan, 1975; Kalbfleisch and Prentice, 2002), in description of frequency and size of natural catastrophes and in quantification of hazard, risks and losses (Klugman *et al*., 2008). Aki (1965) provides a simple derivation of the special case of the above distributions when the flood level *m* is not truncated from the top. Page (1968) and Cosentino *et al*. (1977) provide distributions equivalent to (2) and (3), when the random variable *m* denotes earthquake magnitude. Also, similar functional forms of these distributions are common in the context of frequency e.g. by Geist and Parsons (2006) tsunami size distributions quantification.

Account of discrepancy between the data and the flood occurrence model

An explicit assumption underpinning most probabilistic flood models is that parameters - the mean flood activity rate λ as well as parameters of frequency-size flood distribution $F_{\mathrm{M}}(m)$, remain constant in time. However, examination of most of natural disasters records indicates that the sequence of natural catastrophes is composed of temporal trends, cycles, oscillations and pure random fluctuations (Pisarenko and Rodkin, 2010).

One of the most efficient ways to account for temporal and special fluctuation of parameters λ and β_{F} (or in general, parameters of flood distribution $F_{\mathrm{M}}(m)$), can be done by the introduction of so called compound distributions. Compound distributions are often called Bayesian distributions because they strongly resemble the Bayesian formalism and have many of the same benefits of the original formalism, which naturally governs the whole of this paper. These distributions are obtained by compounding one distribution with another and they offer a powerful tool to account for cases in which a parameter of the distribution is itself a random variable (DeGroot 1970). Treating both parameters λ and β_{F} as random variables distributed according to respective gamma distributions appropriately accounts for their uncertainty. The choice of the gamma distribution does not introduce much limitation, since the gamma distribution can fit an extremely large variety of shapes (Johnson and Kotz, 1970).

The gamma distribution, given its flexibility, is used to model the distribution of various natures of random variables and is given by:

$$f_{\mathrm{X}}(x) = (x)^{\mathrm{q}-1}\frac{p^{\mathrm{q}}}{\Gamma(q)}\mathrm{e}^{-px}, \qquad x > 0 \tag{4}$$

where$\Gamma(q)$ is the gamma function defined as:

$$\Gamma(q) = \int_{0}^{\infty} y^{\mathrm{q}-1}\mathrm{e}^{-\mathrm{y}}\mathrm{d}y. \qquad q > 0 \tag{5}$$

The distribution parameters p and q are related to the mean μ, and variance σ^2, of the distribution according to:

$$\mu_{\mathrm{x}} = \frac{q}{p} \tag{6}$$

and

$$\sigma_{\mathrm{x}}^{2} = \frac{q}{p^{2}}. \tag{7}$$

The coefficient of variation expresses the uncertainty related to a random variable *x*, and is given by:

$$COV_x = \frac{\sigma_x}{\mu_x}. \tag{8}$$

Thus equation (8) describes the variation of a random variable *x* relative to its mean value, with a higher value indicating greater dispersion of the parameter.

After combining the Poisson distribution (1) together with the gamma distribution (4) with parameters p_λ and q_λ, the probability of observing *n* floods within specified time interval t, for temporary, randomly varying flood frequency takes the form:

$$\begin{aligned} P(n|t) &= \int_0^\infty p(n|\lambda, t) f_\Lambda(\lambda) \mathrm{d}\,\lambda \\ &= \frac{\Gamma(n+q_\lambda)}{n!\,\Gamma(q_\lambda)} \left(\frac{p_\lambda}{t+p_\lambda}\right)^{q_\lambda} \left(\frac{t}{t+p_\lambda}\right)^n \end{aligned} \tag{9}$$

where $p_\lambda = \bar{\lambda}/\sigma_\lambda^2$,$q_\lambda = \bar{\lambda}^2/\sigma_\lambda^2$and $\Gamma(\cdot)$ is the Gamma function (7). Parameter $\bar{\lambda}$ denotes the mean value of the distribution parameter λ.

Similarly, combining the flood level distribution (2) with the gamma distribution for β_{F} with parameters p_β and q_β, the CDF of flood level *m* takes the form:

$$F_M(m|m_{\min}) = C_\beta \left[1 - \left(\frac{p_\beta}{p_\beta + m - m_{\min}}\right)^{q_\beta}\right], \tag{10}$$

where $p_\beta = \bar{\beta}_F/\sigma_\beta^2$ and $q_\beta = \bar{\beta}_F^2/\sigma_\beta^2$. The symbol $\bar{\beta}_{\mathrm{F}}$ denotes the mean value of parameter β_{F}, σ_β denotes the standard deviation of $\bar{\beta}_{\mathrm{F}}$ and the normalizing coefficient C_β is given by:

$$C_\beta = \left[1 - \left(\frac{p_\beta}{p_\beta + m_{\max} - m_{\min}}\right)^{q_\beta}\right]^{-1}. \tag{11}$$

Noting that $q_\lambda = \bar{\lambda} \cdot p_\lambda$ and $q_\beta = \bar{\beta}_F \cdot p_\beta$, equations (9) and (10) may alternatively be written respectively as:

$$P(n|t) = \frac{\Gamma(n+q_\lambda)}{n!\,\Gamma(q_\lambda)} \left(\frac{q_\lambda}{\bar{\lambda}t + q_\lambda}\right)^{q_\lambda} \left(\frac{\bar{\lambda}t}{\bar{\lambda}t + q_\lambda}\right)^n \tag{12}$$

and

$$F_M(m|m_{\min}) = C_\beta \left[1 - \left(\frac{q_\beta}{q_\beta + \bar{\beta}_F\ (m - m_{\min})}\right)^{q_\beta}\right] \tag{13}$$

with

$$C_\beta = \left[1 - \left(\frac{q_\beta}{q_\beta + \bar{\beta}_F\ (m_{\max} - m_{\min})}\right)^{q_\beta}\right]^{-1} \tag{14}$$

Note that $q_\beta = \left(COV_\beta^{-1}\right)^2$ and $q_\lambda = \left(COV_\lambda^{-1}\right)^2$. Upon specification of the COV the parameters $\bar{\lambda}$ and $\bar{\beta}_{\mathrm{F}}$, referred to as hyper-parameters of the respective distributions, are estimated on the basis of observed data by applying the maximum likelihood procedure.

One has to be aware that disregarding the temporal and spatial variations of the parameters λ and β_{F} leads to biased estimates of the flood hazard. The compound distributions arise from many probabilistic models applied in the engineering (Hamada *et al.*, 2008), insurance and risk industries (Klugman *et al.*, 2008). In the related field of seismic hazard assessment the first application of the compound distributions was done by Benjamin (1968).

Extreme flood level distribution as applied to paleo- and historic records

This we will present a construction of the likelihood function of desired flood hazard parameters $\boldsymbol{\theta} = \left(\bar{\lambda}, \bar{\beta}_F\right)$ and $m_{\max}$, when prehistoric (i.e. paleo-) and historic floods records are available (Cox *et al.*, 2002). It is assumed that prehistoric and historic flood records contain only the strongest events. The mathematical formalism will be restricted only to the construction of the likelihood function based on historic flood occurrences, since with the exception of a few details, the likelihood function based on prehistoric floods record is built in a similar manner.

By the Theorem of the Total Probability (e.g. Mood *et. al.*, 1974), the probability that in time interval t either no flood occurring, or all occurring floods have flood levels not exceeding m, can be expressed as (Epstein and Lomnitz, 1966; Gan and Tung, 1983; Gibowicz and Kijko, 1994)

$$F_M^{max}(m|m_0, t) = \sum_{i=0}^{\infty} P(i|t)\ [F_M(m|m_0)]^i \tag{15}$$

Relation (15) can be expressed in a much simpler form (e.g. Campbell, 1982;1983), which may be written as

$$F_M^{max}(m|m_0, t) = \left[\frac{q_\lambda}{q_\lambda + \bar{\lambda}_0 t[1 - F_M(m|m_0)]}\right]^{q_\lambda} \tag{16}$$

In relations (15) and (16), m_0 is the flood threshold level for the prehistoric or historic part of the flood records ($m_0 \geq m_{\min}$). Flood level $m_{\min}$ denotes the 'total' threshold flood level and has a rather formal character. The only restriction on the choice of its value is that $m_{\min}$ may not exceed the level of completeness of any part, prehistoric, historic or complete floods records.

It follows from relation (16) that the PDF of the largest flood level m within a period t is:

$$f_M^{\max}(m|m_0,t) = \frac{\bar{\lambda}_0 t q_\lambda f_M(m|m_0) F_M^{\max}(m|m_0,t)}{q_\lambda + \bar{\lambda}_0 t[1 - F_M(m|m_0)]}, \qquad (17)$$

where $\bar{\lambda}_0$ represents the mean of the distribution of the mean activity rate for floods with flood level not less than m_0, and is given by:

$$\bar{\lambda}_0 = \bar{\lambda}[1 - F_M(m|m_0)], \qquad (18)$$

$\bar{\lambda}$ denotes the mean activity rate corresponding to flood levelm_{min}, i.e. $\bar{\lambda} \equiv \bar{\lambda}(m_{min})$.

The function $F_M(m|m_0)$is the CDF of flood level as defined by (13) and $f_M(m|m_0)$ is the PDF of the flood level equal to:

$$f_M(m|m_0) = C_\beta \bar{\beta}_F, \left(\frac{q_\beta}{q_\beta + \bar{\beta}_F(m - m_0)}\right)^{q_\beta + 1}. \qquad (19)$$

After introducing the PDF (17) of the largest flood level m within a time period t, the likelihood function of unknown parameters $\boldsymbol{\theta} = (\bar{\lambda}, \bar{\beta}_F)$ becomes:

$$L_0(\boldsymbol{\theta}|\boldsymbol{m_0}, \boldsymbol{t_0}, \boldsymbol{cov}) = \prod_{i=1}^{n_0} f_M^{\max}(m_{0i}|m_0, t_i). \qquad (20)$$

In order to build the likelihood function (20), three kinds of input data are required: $\boldsymbol{m_0}$, $\boldsymbol{t_0}$, and $\boldsymbol{cov}$, where $\boldsymbol{m_0}$ is vector of the largest flood levels records, $\boldsymbol{t_0}$ denotes vector of the time intervals within which the largest floods occurred, and vector $\boldsymbol{cov} = (cov_\lambda, cov_\beta)$, consists of the coefficients of variation (uncertainty relative to the mean) of the parameters $\boldsymbol{\theta}$.

Combination of flood records with different levels of completeness

If it is assumed that the third, complete part of the flood record can be divided into s sub-records, each of them has a span T_{i} and is complete starting from the known flood level $m_{\mathrm{min}}^{(\mathrm{i})}$. For each sub-record i, $\boldsymbol{m_i}$ is used to denote n_{i} flood levels m_{ij}, where $m_{\mathrm{ij}} \geq m_{\mathrm{min}}^{(\mathrm{i})}$, $i = 1,2,\ldots,s$, and $j = 1,2,\ldots,n_i$. Let $L_i(\boldsymbol{\theta}|\boldsymbol{m}_i)$ denote the likelihood function of the unknown $\boldsymbol{\theta} = (\bar{\lambda}, \bar{\beta}_{\mathrm{F}})$, based on the i-th complete flood sub-record. If the flood levels are independent of their number, the likelihood function $L_i(\boldsymbol{\theta}|\boldsymbol{m}_i)$ is the product of two functions, $L_i(\bar{\lambda}|m_i)$ and $L_i(\bar{\beta}_F|m_i)$.

The assumption that the number of flood events per unit of time is distributed according to the compound Poisson distribution (12) which implies that $L_i(\bar{\lambda}|m_i)$ has the following form:

$$L_i(\bar{\lambda}|\boldsymbol{m}_i) = \left(\bar{\lambda}^{(i)}t + q_\lambda\right)^{-q_\lambda}\left(\frac{\bar{\lambda}^{(i)}t}{\bar{\lambda}^{(i)}t + q_\lambda}\right)^{n_i}, \qquad (21)$$

where $\bar{\lambda}^{(\mathrm{i})}$ is the mean activity rate of flood occurrence corresponding to the level of completeness $m_{\mathrm{min}}^{(\mathrm{i})}$ and is given by:

$$\bar{\lambda}^{i} = \bar{\lambda}\left[1 - F_M\left(m_{\mathrm{min}}^{(i)}|m_{\mathrm{min}}\right)\right]. \qquad (22)$$

Following the definition of the likelihood function based on a set of independent observations and PDF of flood levels (19), the likelihood function $L_i(\bar{\beta}_F|m_i)$ takes the form:

$$L_i(\bar{\beta}_{\mathrm{F}}|\boldsymbol{m}_i) = \left[C_\beta \bar{\beta}_{\mathrm{F}}\right]^{n_i} \prod_{j=1}^{n_i}\left[1 + \frac{\bar{\beta}_F}{q_\beta}\left(m_{ij} - m_{min}^{(i)}\right)\right]^{-(q_\beta+1)}. \qquad (23)$$

Relations (21) and (23) define the likelihood function of the unknown parameters $\boldsymbol{\theta} = (\bar{\lambda}, \bar{\beta}_{\mathrm{F}})$ for each complete flood's record.

Finally, $L(\boldsymbol{\theta})$, the joint likelihood function based on all data, is calculated as the product of the likelihood functions based on prehistoric, historic and complete flood records.

The maximum likelihood estimates of the required flood hazard parameters $\boldsymbol{\theta} = (\bar{\lambda}, \bar{\beta}_{\mathrm{F}})$ are given by the value of $\boldsymbol{\theta}$, which for a given maximum, site characteristic, maximum flood level m_{max}, maximizes the likelihood function$L(\boldsymbol{\theta})$. The maximum of the likelihood function is obtained by solving the system of two equations $\frac{\partial \ell}{\partial \bar{\lambda}} = 0$ and $\frac{\partial \ell}{\partial \bar{\beta}_{\mathrm{F}}} = 0$, where $\ell = \ln[L(\boldsymbol{\theta})]$.

A variance-covariance matrix $\mathbf{D}[\boldsymbol{\theta}]$ of the estimated flood hazard parameters, $\hat{\bar{\lambda}}$ and $\widehat{\bar{\beta}}_{\mathrm{F}}$, can be calculated according to the formula (Edwards, 1972):

$$\mathbf{D}[\boldsymbol{\theta}] = -\begin{bmatrix} \dfrac{\partial^2 \ell}{\partial \bar{\lambda}^2} & \dfrac{\partial^2 \ell}{\partial \bar{\lambda}\, \partial \bar{\beta}_{\mathrm{F}}} \\ \dfrac{\partial^2 \ell}{\partial \bar{\beta}_{\mathrm{F}}\, \partial \bar{\lambda}} & \dfrac{\partial^2 \ell}{\partial \bar{\beta}_{\mathrm{F}}{}^2} \end{bmatrix}^{-1} \tag{24}$$

where derivatives are calculated at the point $\bar{\lambda} = \hat{\bar{\lambda}}$ and $\bar{\beta}_F = \widehat{\bar{\beta}}_F$.

Estimation of the Maximum, Site Characteristic Flood Level $m_{\max}$

From a formal point of view, the maximum likelihood estimate of $m_{\max}$ is simply the largest observed flood level $m_{\max}^{\mathrm{obs}}$ within the span of the entire flood record *T*. This follows from the fact that the likelihood function $L(\boldsymbol{\theta})$ decreases monotonically for $m_{\max} \to +\infty$. Therefore, a more realistic estimation of $m_{\max}$ can be provided only by introduction of some additional information. It can be done e.g. by introducing the condition that the largest observed flood $m_{\max}^{\mathrm{obs}}$ within the span of the entire flood record is equal to the largest expected flood level $\mathbf{E}[m_{\max}^{\mathrm{obs}}; T]$. One can show, (Kijko, 2004; Kijko and Singh, 2011) that introduction of such a condition leads to the equation:

$$m_{\max} = m_{max}^{\mathrm{obs}} + \int_{m_{min}}^{m_{max}} [F_M(\zeta | m_{min})]^n d\zeta. \tag{25}$$

where$F_M(\zeta | m_{\min})$ denotes the compound CDF of flood level (13). Unfortunately, the integral $\int_{m_{min}}^{m_{max}} [F_M(\zeta | m_{min})]^n d\zeta$ does not have a simple solution. A more accessible assessment can be obtained through the application of Cramér's approximation. According to Cramér (1961), for large n, the value of $[F_M(\zeta | m_{min})]^n$ is approximately equal to $\exp\{-n[1 - [F_M(\zeta | m_{min})]\}$. After replacement of $[F_M(\zeta | m_{min})]^{\mathrm{n}}$ by its Cramér approximation, equation (25) takes the form (Kijko and Graham, 1998; Kijko, 2004, Kijko and Singh, 2011)

$$m_{\max} = m_{max}^{obs} + \frac{\delta^{1/q}\, exp[nr^q/(1-r^q)]}{\bar{\beta}_T}\left[\Gamma\left(-\frac{1}{q}, \delta r^q\right) - \Gamma\left(-\frac{1}{q}, \delta\right)\right], \tag{26}$$

where $r = \frac{p_\beta}{p_\beta + m_{max} - m_{min}}$, $c_1 = exp\left[-n\left(1 - C_\beta\right)\right], \delta = nC_\beta$, $p_\beta = \bar{\beta}_F / \sigma_\beta^2$, and $\Gamma(\cdot,\cdot)$ is the complementary Incomplete Gamma Function (Abramowitz & Stegun, 1970).

One has to note, that equation (26) does not provide an explicit estimator for $m_{\max}$ since some terms on the right hand side of the equation also contain the unknown$m_{\max}$. The estimator of $m_{\max}$ can therefore be calculated only by iteration. The approximate variance of this $m_{\max}$ estimator is of the form:

$$Var(m_{max}) = \sigma_M^2 + \left[\frac{\delta^{1/q}\, exp[nr^q/(1-r^q)]}{\bar{\beta}_T}\left[\Gamma\left(-\frac{1}{q}, \delta r^q\right) - \Gamma\left(-\frac{1}{q}, \delta\right)\right]\right]^2, \tag{27}$$

where σ_M denotes standard error in the determination of the largest observed flood record $m_{\max}^{\text{obs}}$.

The maximization of the likelihood function $L(\boldsymbol{\theta})$ together with condition (26), provides the maximum likelihood estimates of the flood hazard parameters $\bar{\lambda}, \bar{\beta}_F$ and $m_{\max}$ which solution can be readily obtained by an iterative procedure.

Some alternative techniques for assessment of $m_{\max}$

Procedure based on Order Statistics

Assuming that in the site of concern, there is a record of *n* floods with levels $m_1, m_2, \dots, m_n$. Each flood level m_{i} is greater than or equal to $m_{\min}$ $(i = 1, \dots, n)$, where $m_{\min}$ is a known level of completeness i.e. all floods having flood level greater than or equal to $m_{\min}$ are recorded. The time span of the flood record is denoted as *T*. Assuming that the flood levels *m* are independent, identically distributed, random values having PDF $f_M(m|m_{min})$ and CDF, $F_M(m|m_{min})$ respectively. The parameter $m_{\max}$ is the unknown, the site characteristic upper limit of the flood level.

If the n flood records are arranged in increasing order, that is $m_{min} \le m_1 \le m_2 \le \cdots \le m_{n-1} \le m_n \le m_{\max}$, any empirical distribution function $\hat{F}_M(m|m_{min})$ can be approximated as:

$$\hat{F}_M(m|m_{min}) = \begin{cases} \quad 0 & for\ m < m_1 \\ {}^{i}/_{n} & for\ m_i \le m \le m_{i+1} \\ 1 & for\ m \ge m_n, \end{cases} \tag{28}$$

where $i = 1, \dots, n-1$.

The approximate value of the integral $\int_{m_{min}}^{m_{max}} [F_M(\zeta|m_{min})]^n d\zeta$, (equation 25), is then:

$$\int_{m_{min}}^{m_{max}} [F_M(\zeta|m_{min})]^n d\zeta = \sum_{i=1}^{n-1} \left(\frac{i}{n}\right)^n (m_{i+1} - m_i)$$
$$= m_{max}^{obs} - \sum_{i=0}^{n-1} \left[\left(1 - \frac{i}{n}\right)^n - \left(1 - \frac{i+1}{n}\right)^n\right] m_{n-i}. \tag{29}$$

Since $\lim_{n\to\infty}(1 + 1/n)^n = \mathrm{e}$, and after simple rearrangement, order statistics estimator (29) takes a simple form:

$$\widehat{m}_{max} = m_{max}^{obs} + \left(m_{max}^{obs} - c_1 \sum_{i=0}^{n-1} e^{-i} m_{n-i}\right), \tag{30}$$

where $\mathrm{c}_1 = 1 - 1/\mathrm{e} \cong 0.632$.

Assuming that the standard error in the determination of the flood records $m_1, m_2, \ldots, m_n$ is known and equal to σ_M, for large n, the approximate standard deviation of the estimator (30) is equal to

$$\sigma_{\widehat{m}_{max}} = \sqrt{Var(\widehat{m}_{max})} = \sqrt{c_0\sigma_M^2 + \left(\widehat{m}_{max} - m_{max}^{obs}\right)^2}, \tag{31}$$

where $c_0 = (1 + e^{-1})^2 + e^{-2}\,(1 - e^{-1})/(1 + e^{-1}) \cong 1.93.$

Estimator (30) is very useful. It can be used when the functional form of the CDF of flood level $F_M(m|m_{min})$, is not known or known only approximately. Also, it can be used when only information about the largest floods is available. Despite of the fact that the procedure was not design for small number of observations, because the contribution of low level floods decreases rapidly with increasing i, the procedure relies only on the knowledge of a few of the largest flood events. However, in the next paragraph an alternative technique is discussed designed specifically for the case when only several of the largest flood records are known.

Procedure based on a few largest flood records.

In the language of mathematical statistics, the case in which a known number of observations is missing from either end of the distribution is known as (single) data censoring (David, 1981). The problem of estimating the bounds of random variables when the data is censored, and only a few largest (or smallest) observations are available, has been extensively discussed by Cooke (1980). Theoretical results expressed in terms of determination of the maximum flood level characteristic to the site can be summarized as follows.

Assuming, in the vicinity of the site of concern, only then_0 largest floods are knownout of n occurred flood events with levels $m_i \geq m_{min}$ $(i = 1, \dots, n)$, which occurred. Following Gnedenko's condition (Gnedenko, 1943), that for a very broad class of CDFs, when their argument is near to the upper endpoint and the CDF is linear, one can justify an estimator of $m_{\max}$ of the following form:

$$\widehat{m}_{max} = \sum_{i=1}^{n_0} a_i m_{n-i+1} \tag{32}$$

where a_i $(i = 1, \dots, n_0)$ are the coefficients to be determined. Concentrating on the most important case from a practical point of view, viz. when the distribution of the flood levels is truncated from the top, Cooke (1980) has found that for truncated distributions, the minimization of the mean squared error of estimator (32) can be obtained when $a_1 = 1 + 1/n_0$, $a_2 = \dots = a_{n_0-1} = 0$, and $a_{n_0} = -1/n_0$, that is $\varDelta = \left(m_{max}^{obs} - m_{n-n_0+1}\right)/n_0$ and therefore:

$$\widehat{m}_{max} = m_{max}^{obs} + \frac{1}{n_0}\left(m_{max}^{obs} - m_{n-n_0+1}\right). \tag{33}$$

Probably, the greatest attraction of estimator (33) relies on its simplicity and that even for a small number of observations n_0, the estimator is nearly optimal (in the sense of its mean squared errors). This emanates as a result of the fact that for large$n$, the largest few observations carry most of the information about its endpoint. It is interesting to note that the value of $m_{\max}$ estimated according to formula (33) is based only on two observations: the n_0th largest flood level m_{n-n_0+1}, and the largest observed flood level $m_{\max}^{\text{obs}}$. Clearly, the better estimator of $m_{\max}$ can be obtained by inclusion of the remaining $n_0 - 2$ largest observations. This can be done by application of the Quenouille's technique, originally developed for averaging of the bias of an estimator (Quenouille, 1956). Averaging the correcting factor $\varDelta = \left(m_{max}^{obs} - m_{n-i+1}\right)/n_0$ over the $n_0 - 1$ possible choices produces:

$$\varDelta = \frac{1}{n_0}\left(m_{max}^{obs} - \frac{1}{n_0 - 1}\sum_{i=2}^{n_0} m_{n-i+1}\right), \tag{34}$$

and therefore, the estimator of m_{max} takes the form:

$$\widehat{m}_{max} = m_{max}^{obs} + \frac{1}{n_0}\left(m_{max}^{obs} - \frac{1}{n_0 - 1}\sum_{i=2}^{n_0} m_{n-i+1}\right). \tag{35}$$

Assuming that the standard errors in determination of the flood levels m_{n-i+1} are the same and equal to σ_M, the approximate variance of the estimator (35) is:

$$Var(\widehat{m}_{max}) = c_0\sigma_M^2 + \Delta^2, \tag{36}$$

where $c_0 = (n_0^2 + n_0 - 1)/[n_0(n_0 - 1)]$.

The greatest attraction of estimator (36) relies on its simplicity and that it requires only the knowledge of magnitudes of a few largest events.

Robson-Whitlock procedure

Let us assume that the analytical form of the flood level distribution$F_M(m|\mathrm{m_{min}})$is not known and we want to estimate the right end point of the distribution, viz. the maximum flood m_{max}. This can be achieved in several ways. One of them is to apply the classical, (Quenouille, 1956) technique of successive bias reduction, modified to fit the factorial series rather than the power series in$1/n$. Robson and Whitlock (1964) showed that, under very general conditions, and when the data are arranged in ascending order of flood levels, viz. $m_1 \leq m_2 \leq \cdots \leq m_{n-1} \leq m_{max}^{obs}$, Quenouille's approach leads to the following rule in estimation of m_{max}

$$\widehat{m}_{\max} = m_{\max}^{obs} + \left(m_{\max}^{obs} - m_{n-1}\right). \tag{37}$$

Equation (37) was probably first derived by Robson and Whitlock (1964), and so it is often called the Robson-Whitlock estimator. It can be shown that the above estimator is mean-unbiased to the order n^{-2} and asymptotically median-unbiased. The simplicity of the formula (37) makes it very attractive. It can be applied in cases of limited and/or doubtful flood data, when one wants to get quick results without going into sophisticated analysis. Unfortunately, it can be shown, that the bias reduction is achieved at the expense of a high mean-squared error value. In fact, Robson and Whitlock (1964) derived a general formula for an estimator of truncation point, with mean unbiased to any order n^{-k}:

$$\widehat{m}_{max} = \sum_{j=0}^{k} (-1)^j \binom{k+1}{j+1} m_{n-j} \tag{38}$$

where $0 < k < n$. Also, this formula does not provide a guarantee that $\widehat{m}_{max}$, the estimated upper point of flood level distribution, is equal to or exceeds the maximum flood level m_{max}^{obs}, already observed.

The approximate variance of the *Robson-Whitlock* estimator of $m_{\max}$ is of the form:

$$Var(\widehat{m}_{\max}) = 5\sigma_M^2 + \left(m_{\max}^{\mathrm{obs}} - m_{n-1}\right)^2, \tag{39}$$

where $\sigma_M$denotes standard error in the determination of the largest observed flood record m_{max}^{obs}.

Some Special Cases

Only largest historic flood records are available.

For many flood-threatened areas, long flood historic records are available but they are incomplete and contain information only about the largest and catastrophic flood occurrences.

In such a case, the likelihood function $L(\boldsymbol{\theta})$ of the unknown flood hazard parameters $\boldsymbol{\theta} = (\bar{\lambda}, \bar{\beta}_F)$ takes the form (20). Further simplification of the procedure for flood hazard assessment can be obtain if the Poisson-gamma, compound distribution (12) is replaced by the classic Poisson distribution (1) and the negative exponential-gamma distribution of flood level (13), is replaced by the classic negative exponential distribution (3).

If $\ell = \ln[L(\boldsymbol{\theta})]$, by solving the system of two equations $\frac{\partial \ell}{\partial \bar{\lambda}} = 0$ and $\frac{\partial \ell}{\partial \bar{\beta}_F} = 0$, we obtain (Kijko and Dessokey, 1987; Kijko and Sellevoll, 1989):

$$\begin{cases} \dfrac{1}{\lambda} = \dfrac{\langle \boldsymbol{t_0} \rangle A_2 - \langle t_0 A \rangle}{A_2 - A_1}, \\ \dfrac{1}{\beta_F} = \langle m_0 \rangle - \dfrac{\langle \boldsymbol{t_0 A\, m_0} \rangle - \langle \boldsymbol{t}_0 \rangle A_2 m_{max}}{\langle \boldsymbol{t_0 A} \rangle - \langle \boldsymbol{t_0} \rangle A_2} \end{cases} \tag{40}$$

where

- $\langle \boldsymbol{t_0} \rangle = \sum_{j=1}^{n_0} t_{0\,j} / n_0,$

- $\langle \boldsymbol{m_0} \rangle = \sum_{j=1}^{n_0} m_{0j}/n_0$,
- $\langle \boldsymbol{t_0 A} \rangle = \sum_{j=1}^{n_0} t_{0j} A(m_{0j})/n_0$,
- $\langle \boldsymbol{t_0\ m_0 A} \rangle = \sum_{j=1}^{n_0} t_{0j} m_{0j} A(m_{0j})/n_0$,
- $A_1 = exp(-\beta_F m_{\min})$,
- $A_2 = exp(-\beta_F m_{\max})$,

and elements of vector $\boldsymbol{A}$ are equal to $A(m_{0j}) = \exp(-\beta_F\ m_{0j})$, where j = 1,...,n_0.

For the specified value of $m_{\max}$, the solution set of equations (40) provides the maximum likelihood estimates of the required flood hazard parameters λ and β_{F}. It is interesting to note that for $m_{\max} \rightarrow +\infty$, and $T_{\mathrm{j}} = \mathrm{const}$, the system of equations (40) is reduced to the maximum likelihood estimation of the parameters λ and β of the first Gumbel distribution (Kimball, 1946).

Largest historic flood records are not available

In the case, when the paleo- and historic flood records are absent, all the above formalism can be significantly simplified. By replacing the respective compound distributions by their classic counterparts, a simple, overall maximum likelihood estimate of the λ and β_F parameters can be obtained by the application of the additive property of likelihood functions (Rao, 1973). If applied to the current problem, the joint likelihood function of the β_{F} parameter, which utilizes flood records having variable level of completeness (Figure 2), is defined as:

$$L(\beta_F) = \prod_{i=1}^{s} L_i(\beta_F | \boldsymbol{m}_{\mathrm{i}}), \tag{41}$$

where $L_i(\beta_F|\boldsymbol{m}_i)$ is the special case of the likelihood function (23), i denotes index of the i^{th} complete flood record, and $i = 1,2, \ldots, s$.

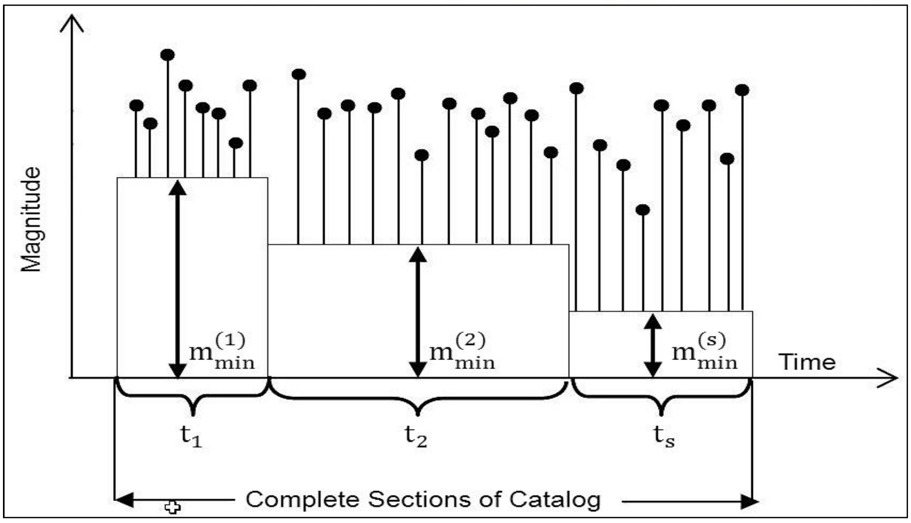

Figure 2. A schematic illustration of a flood record with variable level of completeness.(Modified after Kijko and Sellevoll, 1989).

After assumption that flood levels are independent, identically distributed random variables, following the PDF (3), and for $m_{max} \to +\infty$, the likelihood function for the i^{th} sub-record takes the form:

$$L_i(\beta_F|\boldsymbol{m}_i) = \prod_{j=1}^{n_i} \beta_F \exp\left[-\beta_F\left(m_j^{(i)} - m_{min}^{(i)}\right)\right], \tag{42}$$

where $m_{\mathrm{j}}^{(\mathrm{i})}$ is the sample of n_i flood records recorded during the time span of the i^{th} sub-record. Following equation (23), the joint likelihood function, which utilizes records all floods within the entire span of the record, takes the form:

$$L(\beta_F) = \prod_{i=1}^{s}\prod_{j=1}^{n_i} \beta_F \exp\left[-\beta_F\left(m_j^{(i)} - m_{\min}^{(i)}\right)\right]. \tag{43}$$

One can show (Kijko and Smit, 2012), that maximization of (43) provides the maximum likelihood estimator of β_{F} in the simple form:

$$\hat{\beta}_F = \left(\frac{r_1}{\hat{\beta}_F^{(1)}} + \frac{r_2}{\hat{\beta}_F^{(2)}} + \cdots + \frac{r_S}{\hat{\beta}_F^{(S)}}\right)^{-1}, \tag{44}$$

where $r_i = n_i/n$;$n = \sum_{i=1}^{s} n_i$ is total number of floods in complete flood record with levels equal or exceeding the relevant level of completeness $m_{\min}^{(\mathrm{i})}$, and $\hat{\beta}_{\mathrm{F}}^{(\mathrm{i})}$ are:

$$\hat{\beta}_{\mathrm{F}}^{(\mathrm{i})} \quad = \frac{1}{\langle \boldsymbol{m}_i \rangle - \mathrm{m}_{\min}^{(\mathrm{i})}}, \tag{45}$$

where $\langle \boldsymbol{m}_i \rangle = \sum_{j=1}^{n_i} \frac{m_{ij}}{n_i}$ and denotes the sample mean of the flood records observed within complete part of the flood data *i*. It is easy to note that the parameters $\hat{\beta}_{\mathrm{F}}^{(\mathrm{i})}$calculated in this manner are the simple estimators of the β_{F} parameters, calculated for individual complete flood records *i*, $(i = 1, \dots, s)$. The estimator (33) has exactly the same form as the classic Aki-Utsu estimator of theβ-value of in the frequency-magnitude Gutenberg-Richter relation (Aki, 1965, Utsu, 1965).

One of advantages of application of the maximum likelihood procedure for parameters estimation is fact, that it provides straightforward approximations for the standard errors and confidence intervals. Based on the Central Limit Theorem, it can be shown (e.g. Mood *et. al*., 1974), that under suitable regularity conditions and for a sufficiently large number of events, the estimator (44) is approximately normally distributed about its mean and its sample standard deviation is defined as:

$$\hat{\sigma}_{\hat{\beta}} = \frac{\hat{\beta}_{\mathrm{F}}}{\sqrt{n}} \tag{46}$$

The confidence intervals of estimator (44) are:

$$\hat{\beta}_{\mathrm{F}} \pm z_{\alpha/2} \hat{\sigma}_{\hat{\beta}}. \tag{47}$$

In equation (47), $\mathrm{z}_{\alpha/2}$ denotes the $(1 - \alpha/2)$ quintile of the standard normal distribution. The natural question is: how many events are needed in order to be sure that the estimator of $\hat{\beta}_{\mathrm{F}}$ is distributed normally. According to Jansson (1966), the sum of 12 uniformly distributed random numbers will create a set of random numbers with a bell-shape distribution that is approximately Gaussian. Surprisingly, such an approximation by only 12 numbers fairly sound, especially if only the central part (mean ± SD) of such a Gaussian-like distribution is explored. Obviously, if inference is to be based on the tail of such a distribution, more observations are required.

Once the value of parameter $\hat{\beta}_{\mathrm{F}}$is known, the mean value of flood activity rate can be calculated. Following our notation, $\lambda \equiv \lambda(m_{\min})$ denotes the site-characteristic, flood activity rate with an flood level equal to or greater than $m_{\min}$. It can be shown (Kijko and Sellevoll, 1989; Kijko and Smit, 2012), that if the number of floods per unit of time is a Poisson random variable, the maximum likelihood estimator of $\lambda(m_{\min})$ takes the form:

$$\hat{\lambda}(m_{min}) = \frac{n}{\sum_{i=1}^{s} t_i \cdot \exp\left[-\hat{\beta}_{\mathrm{F}}\left(m_{\min}^{(i)} - m_{\min}\right)\right]}. \tag{48}$$

For only one complete flood record (i.e. where $s = 1;\ m_{\min}^{(1)} = m_{\min}^{(2)} = \cdots = m_{\min}^{(s)} = m_{\min};\ t = t_1$ with $t_2 = t_3 = \cdots = 0$ and $n = n_1$ with $n_2 = n_3 = \cdots = 0$), the estimator (48), reduces to the classic maximum likelihood estimator of parameter of Poisson distribution and takes simple form $n/t.$

This formalism can be applied to flood from rainfall data in a specific subset of cases where reliable flood records are not available. As long as one accepts the assumption that flood levels can be described as a linear function of rainfall, rainfall records can be used as a proxy for the corresponding flood observations generating an index of severity.

REFERENCES

Abramowitz, M. and I. A. Stegun (1970). *Handbook of Mathematical Functions*, 9th ed., Dover, New York.

Aki, K. (1965). Maximum Likelihood estimate of b in the formula log N=a-bM and its Confidence Limits, *Bull. Earthquake Res* Inst., Tokyo Univ., **43**, 237-239.

Barlow R.E. and F. Proschan. (1975). *Statistical Theory for Reliability and Life Testing: Probability Models,* Holt, Reinhart, and Winston, New York.

Benjamin, J.R. (1968). Probabilistic models for seismic forces design, J. Struct. Div., ASCE **94**,(ST5) 1175-1196.

Campbell, K.W. (1982). Bayesian analysis of extreme earthquake occurrences. Part I. Probabilistic hazard model, *Bull. Seism. Soc. Am.*, **72**, 1689-1705.

Campbell, K.W. (1983). Bayesian analysis of extreme earthquake occurrences. Part II. Application to the San Jacinto Fault zone of southern California, *Bull. Seism. Soc. Am.*, **73**, 1099-1115.

Cooke, P. (1980). Optimal linear estimation of bounds of random variables, *Biometrika*, **67**, 257-258.

Cosentino, P., V. Ficara, and D. Luzio (1977). Truncated exponential frequency - magnitude relationship in the earthquake statistics, *Bull. Seism.Am.*,**67**, 1615-1623.

Cox, D.R., V.S. Isham and P.J. Northrop (2002). Floods: some probabilistic and statistical approaches, *Phil. Trans. R. Soc. London. A* 2002, **360**, 1389-1408

Cramér, H. (1961). *Mathematical Methods of Statistics,* Princeton University Press. Princeton.

David, H.A. (1981). *Order Statistics,* John Wiley and Sons, New York.

DeGroot, M.H. (1970), *Optimal Statistical Decisions,* McGraw-Hill, New York.

Eadie, W.T., D. Drijard, F.E. James, M. Roos, and B. Sadoulet (1970).*Statistical Methods in Experimental Physics,* North-Holland Publishing Company. Second reprint 1982.

Edwards, A.W.F. (1972). *Likelihood.* Cambridge University Press, New York, pp. 235.

Epstein, B., and C. Lomnitz (1966). A model for occurrence of large earthquakes, *Nature*, **211**, 954-956.

Gan, Z.J. and C.C. Tung (1983). Extreme value distribution of earthquake magnitude, *Phys. Earth Planet. Inter.,* **32**, 325-330.

Geist, E.L. and T. Parsons (2006). Probabilistic Analysis of Tsunami Hazards, *Natural Hazards*, **37**, 277-314.

Gibowicz, S.J. and A. Kijko (1994). *An Introduction to Mining Seismology*, Academic Press, San Diego, pp. 396.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'unesériealéatoire, *Ann. Math.***44**, 423-254.

Hamada, M.S., A.G. Wilson, C. Shane Reese and H.F. Martz (2008). *Bayesian Reliability,* Springer, New York, pp.430.

Jansson, B. (1966). *Random Number Generators,* Vector Pettersons Bokindustri Aktiebolag, Stockholm, pp. 205.

Johnson, N.L., and S. Kotz (1970). *Continuous Univariate Distributions* -1, John Willey & Sons. New York, pp. 300.

Kalbfleisch, J.D. and R.L. Prentice (2002). *The Statistical Analysis of Failure Time Data,* Second Edition. Wiley Interscience. New Jersey, pp. 439.

Kijko, A. (2004). Estimation of the maximum earthquake magnitude $m_{\max}$, *Pure Appl. Geophys*, **161**, 1-27.

Kijko, A. and M.M. Dessokey (1987). Application of the extreme magnitude distributions to incomplete earthquake files, *Bull. Seism. Soc. Am*. **77**, 1429-1436.

Kijko, A., and G. Graham (1998).Parametric-historic Procedure for Probabilistic Seismic Hazard Analysis. Part I: Estimation of Maximum Regional Magnitude $\mathrm{m}_{\max}$, *Pure Appl. Geophys*. ***152***, 413-442.

Kijko, A. and M.A. Sellevoll (1989). Estimation of earthquake hazard parameters from incomplete data files, Part I, Utilization of extreme and complete catalogues with different threshold magnitudes, *Bull. Seism. Soc. Am*. **79**, 645-654.

Kijko, A. and M.A. Sellevoll (1992). Estimation of earthquake hazard parameters from incomplete data files, Part II, Incorporation of magnitude heterogeneity, *Bull. Seism. Soc. Am*. **82**, 120-134.

Kijko, A., Smit, A. (2012). Extension of the Aki-Utsu b-value Estimator for Incomplete Catalogs, *Bull. Seism. Soc. Am*.**102** nr 3.

Kimball, B. F. (1946). Sufficient statistical estimation functions for the parameters of the parameters of the distribution of maximum values, *Ann. Math. Stat*. **17**, 299-306.

Klugman, S.A., H.H. Panjer, and G.E. Willmot (2008). *Loss Models. From Data to Decisions.* John Willey & Sons, Inc., Hoboken, New Jersey.

Mood, A.M., F. Graybill, and D.C. Boes (1974). *Introduction to the Theory of Statistics,* McGraw-Hill, Auckland, pp.564.

Page, R. (1968). Aftershocks and microaftershocks. *Bull. Seism. Soc. Am*., **58**, 1131-1168.

Pisarenko, V., and M. Rodkin (2010). *Heavy-Tailed Distributions in Disaster Analysis*. Springer. Advances in Natural and Technological Hazards Research, Volume 30. New York, pp. 190.

Rao, C.R. (1973). *Linear Statistical Inference and Its Application,* Edition 2. John Willey and Sons, New York, p.625.

Robson, D.S. and J.H. Whitlock (1964).Estimation of a truncation point, *Biometrika*, **51**, 33-39.

Quenouille, M.H. (1956). Note on bias in estimation, *Biometrica*, **43**, 353-360.

Tinti, S. and F. Mulargia (1985). Effects of magnitude uncertainties in the Gutenberg-Richter frequency-magnitude law, *Bull. Seism. Soc. Am*. **75**, 1681-1697.

Utsu, T. (1965). A method for determining the value of b in the formula log n = a–bM showing the magnitude-frequency relation for earthquakes (with English summary), *Geophys. Bull. Hokkaido Univ*. **13**, 99-103.