Nordic Hydrology, 8, 1977, 257-266

Crossings and Extremes in Dependent Annual Series

Dan Rosbjerg.
Technical University of Denmark.

Crossing characteristics of a discrete stationary process with Markov properties are re-examined. A new extreme value distribution is developed and compared with extreme values obtained from long-term annual streamflow records.

Introduction

In a series of annual mean values X_i, $i = 1,2,....$, we sometimes find the assumptions of independence and stationarity justified, i.e.

$$P\{X_i \leq x \mid X_{i-1} = x_{i-1}, \quad X_{i-2} = x_{i-2}, \ldots\} = P\{X_i \leq x\} = F(x) \quad (1)$$

allowing us to make use of the well known theory of the binominal process, when considering crossings of the level x. For the time period until the first upcrossing of the level x occurs, in the following denoted by N, a geometric distribution with parameter $p = 1 - F(x)$ is obtained. Because of the independence, the same distribution can be applied to the time interval between successive events $X > x$ as well as to the number of time steps in succession where $X \leq x$ (the run-length).
The geometric probability mass function reads

$$P\{N = n\} = [1-F(x)][F(x)]^{n-1}; \qquad n \geq 1 \quad (2)$$

and the probability distribution function

$$P\{N \leq n\} = 1 - [F(x)]^n \tag{3}$$

For the mean and the variance we have

$$E\{N\} = \frac{1}{1-F(x)} \tag{4}$$

and

$$\mathrm{Var}\{N\} = \frac{F(x)}{[1-F(x)]^2} \tag{5}$$

It should be noted that the mean value, $E\{N\}$ is equivalent to the return period of the event $X > x$.

Considering a given interval of time, $1 \leq i \leq n$, it follows directly that the distribution function for the maximum value of X_i in this time period, denoted by X'_n, is

$$P\{X'_n \leq x\} = P\{N > n\} = [F(x)]^n \tag{6}$$

Let us now turn to the more complicated situation where the independence assumption has to be abandoned, and let us assume that a stationary Markov model satisfactorily describes the persistence in the considered series.

The usual Markov condition reads

$$P\{X_i \leq x \mid X_{i-1} = x_{i-1}, \quad X_{i-2} = x_{i-2}, \ldots\} = P\{X_i \leq x \mid X_{i-1} = x_{i-1}\} \tag{7}$$

However, dealing with crossings of the level x, we shall not make use of Eq. (7), which applies to the continuous state space Markov process. Instead we will make a rather crude approximation where the continuous state space is reduced to the two states $X \leq x$ and $X > x$. With this Markov chain approximation Eq. (7) can be replaced by the condition

$$P\{X_i \leq x \mid X_{i-1} \leq x, \quad X_{i-2} \leq x, \ldots\} = P\{X_i \leq x \mid X_{i-1} \leq x\} \tag{8}$$

which leads to great simplifications in the analysis. Fortunately Eq. (8) is approximately valid for the general Markov process, at least for moderately persistent processes, see e.g. Saldarriaga and Yevjevich (1970)

The marginal distribution function and the simultaneous distribution function of two successive events in the process will be denoted $F_1(\cdot)$ and $F_2(\cdot,\cdot)$, respectively, i.e.

$$P\{X_i \leq x\} = F_1(x) \tag{9}$$

and

$$P\{X_i \leq x, \; X_{i+1} \leq x\} = F_2(x,x) \tag{10}$$

The correlation coefficient ρ_1, between successive events in the process is defined by

$$\rho_1 = \frac{E\{(X_i - \mu_x)(X_{i+1} - \mu_x)\}}{\sigma_x^2} \tag{11}$$

where $\mu_x = E\{X_i\}$ and $\sigma_x^2 = \mathrm{Var}\{X_i\}$ denote the mean value and the variance in the

process respectively. ρ_1 is a measure of the degree of linear dependence between successive years and therefore an important parameter in the description of the persistence in the series.

We will now proceed with a more detailed analysis of the following random variables:

K: the time interval between successive events $X > x$.
M: the number of time steps in succession where $X \leq x$ (the run-length).
N: the time interval until the first upcrossing of the level x occurs.
X_n': the maximum value of the process in the time interval $1 \leq i \leq n$.

In this paper only crossings of the level x are considered, and we will accordingly facilitate the notation by writing F_1 for $F_1(x)$ and F_2 for $F_2(x,x)$.

Distribution of the Run-Length

As a beginning, we shall give two simple probability formulas to be used in the following. According to Fig. 1, we have

$$P\{X_0 > x,\ X_1 \leq x\} = P\{X_0 < \infty,\ X_1 \leq x\} - P\{X_0 \leq x,\ X_1 \leq x\} = F_1 - F_2 \tag{12}$$

and

$$\begin{aligned} &P\{X_0 > x,\ X_1 > x\} \\ &\quad = 1 - P\{X_0 \leq x,\ X_1 > x\} - P\{X_0 > x,\ X_1 \leq x\} - P\{X_0 \leq x,\ X_1 \leq x\} \\ &\quad = 1 - 2(F_1 - F_2) - F_2 = 1 - 2F_1 + F_2 \end{aligned} \tag{13}$$

The probability of obtaining an event $X > x$ followed by two events $X \leq x$ can now be found. Combining Eqs. (8) and (12) gives

$$\begin{aligned} &P\{X_0 > x,\ X_1 \leq x,\ X_2 \leq x\} = P\{X_2 \leq x \mid X_0 > x,\ X_1 \leq x\} P\{X_0 > x,\ X_1 \leq x\} \\ &\quad = P\{X_2 \leq x \mid X_1 \leq x\} P\{X_0 > x,\ X_1 \leq x\} = \frac{F_2}{F_1} \quad (F_1 = F_2) \end{aligned} \tag{14}$$

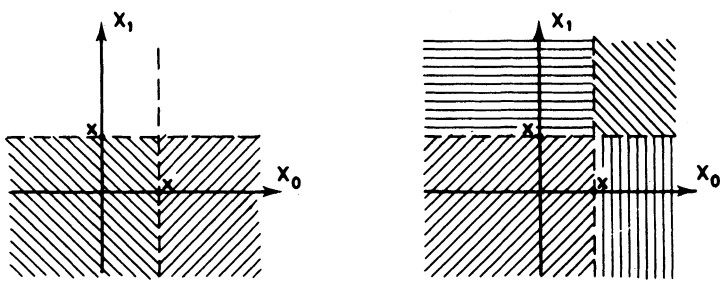

Fig. 1. Sketches corresponding to Eqs. (12) and (13).

By applying this procedure successively we can derive the probability of obtaining m events $X \leq x$ in succession preceded and followed by events $X > x$. This compound event is illustrated in Fig. 2. For the probability we get

$$P\{X_0 > x,\ X_1 \leq x,\ X_2 \leq x, \ldots, X_m \leq x,\ X_{m+1} > x\}$$
$$= (F_1 - F_2)\left(\frac{F_2}{F_1}\right)^{m-1} \frac{F_1 - F_2}{F_1} \equiv \frac{(F_1 - F_2)^2}{F_1} \left(\frac{F_2}{F_1}\right)^{m-1}; \qquad m \geq 1 \qquad (15)$$

Eq. (15) which previously has been stated by Gottschalk (1976) forms the basis from which the wanted probability distributions shall be deduced.

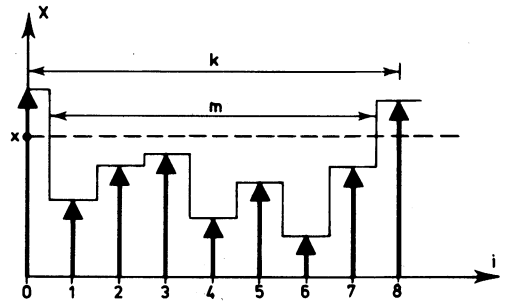

Fig. 2. m events $X<x$ preceded and followed by events $X>x$.

Combining Eqs. (13) and (15) and utilizing $k = m+1$ we find

$$P\{K = k,\ X_0 > x\} = \begin{cases} 1 - 2F_1 + F_2 & k = 1 \\ \dfrac{(F_1 - F_2)^2}{F_1} \left(\dfrac{F_2}{F_1}\right)^{k-2} & k > 1 \end{cases} \qquad (16)$$

The probability mass function for the time interval between successive events $X > x$ is now easily derived from Eq. (16). We get

$$g(k) = P\{K = k \mid X_0 > x\} = \begin{cases} \dfrac{1 - 2F_1 + F_2}{1 - F_1} & k = 1 \\ \dfrac{(F_1 - F_2)^2}{F_1(1 - F_1)} \left(\dfrac{F_2}{F_1}\right)^{k-2} & k > 1 \end{cases} \qquad (17)$$

The probability distribution function then becomes

$$G(k) = \sum_{j=1}^{k} g(j) = 1 - \frac{F_1 - F_2}{1 - F_1} \left(\frac{F_1}{F_1}\right)^{k-1} \qquad (18)$$

For the mean and the variance we obtain

$$E\{K\} = \sum_{k=1}^{\infty} k\, g(k) \equiv \frac{1}{1 - F_1} \qquad (19)$$

and

$$\mathrm{Var}\{K\} = \sum_{k=1}^{\infty} [k - E\{K\}]^2 g(k) \equiv \frac{2F_1^2}{(F_1 - F_2)(1 - F_1)} - \frac{F_1}{(1 - F_1)^2} \qquad (20)$$

Note that Eq. (19) is expectedly equivalent to Eq. (4) in spite of the introduced persistence in the series. The return period of an event $X > x$ can only be a function of the marginal distribution.

In order to obtain the probability distribution of the run-length (the number of time steps in succession where $X \leq x$), we write Eq. (15) as

$$P\{M = m,\ X_0 > x,\ X_1 \leq x\} = \frac{(F_1 - F_1)^2}{F_1} \left(\frac{F_2}{F_1}\right)^{m-1}; \qquad m \geq 1 \qquad (21)$$

The probability mass function now follows immediately from Eq. (21)

$$h(m) = P\{M = m \mid X_0 > x,\ X_1 \leq x\} = \frac{F_1 - F_2}{F_1} \left(\frac{F_2}{F_1}\right)^{m-1}; \qquad m \geq 1 \qquad (22)$$

which is the geometric distribution with parameter $p = (F_1 - F_2)/F_1$. Accordingly, the probability distribution function, the mean and the variance become

$$H(m) = 1 - \left(\frac{F_2}{F_1}\right)^m \qquad (23)$$

$$E\{M\} = \frac{F_1}{F_1 - F_2} \qquad (24)$$

$$\mathrm{Var}\{M\} = \frac{F_1 F_2}{(F_1 - F_2)^2} \qquad (25)$$

It was to be expected to find a geometric distribution of the run-length because of the Markov assumption (i.e. the future state is only dependent on the present), which implies that the remaining part of the run-length at an arbitrarily chosen time has to be distributed exactly as the total run-length. No other discrete probability distribution, except the geometric one satisfies this condition.

The distribution of the length of an excursion *above* the level x, denoted by L, will consequently be a geometric one too, the parameter being $p = (F_1 - F_2)/(1 - F_1)$. The mean value therefore becomes

$$E\{L\} \equiv \frac{1 - F_1}{F_1 - F_2} \qquad (26)$$

a formula previously given by Nordin and Rosbjerg (1970). In this paper some crossing properties of a number of long-term annual stremflow records were analysed, one of them being the mean length of excursions above different crossings levels.

The standard normal distribution function $\Phi(\cdot)$ was used as an approximation to the marginal distribution of the standardized series (i.e. average value and standard devia-

tion equal to 0 and 1, respectively), and the standard normal bivariate distribution function $\Phi_2(\cdot,\cdot;\rho_1)$ was utilized as an approximation to the simultaneous distribution of successive events.

Considering for example crossings of the level 1.0 in the standardized series of the rivers Danube ($\rho_1 \equiv 0.1$) and Mississippi ($\rho_1 = 0.3$), the following observed and theoretical mean values were obtained:

Table 1 - Empirical and theoretical mean lengths of excursions above the level 1.0.

		$\bar{\ell}$	$E\{L\} \equiv \frac{1-\Phi(1)}{\Phi(1)-\Phi_2(1,1;\rho_1)} \equiv \frac{1}{p}$
Danube	$\rho_1 = 0.1$	1.33	1.25
Mississippi	$\rho_1 = 0.3$	1.42	1.40

The comparison of mean values given in Table 1 can now be extended by means of the theory outlined above. In Fig. 3, the observed frequences are compared with the theoretical ones obtained from the geometric distribution $h(\ell) = p(1-p)^{\ell-1}$. As seen from the figure, the agreement is fairly good.

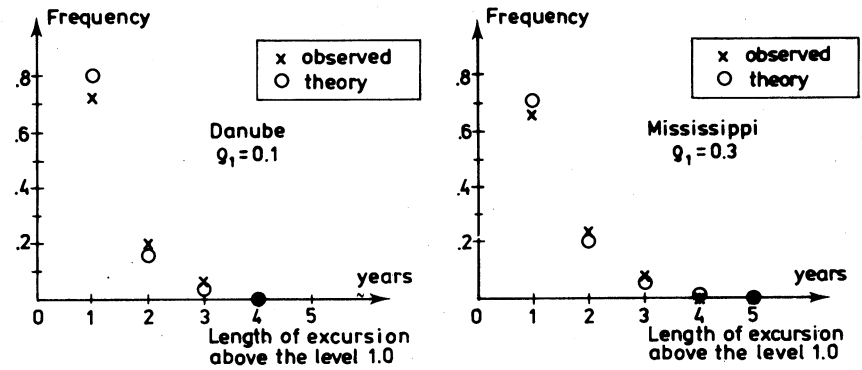

Fig. 3. Distributions of the lengths of excursions above the level 1.0 in the standardized series of Danube (ρ_1=0.1) and Mississippi (ρ_1=0.3).

Extreme Value Distribution

Before the extreme value distribution can be stated, the probability distribution of the time period until the first upcrossing of the level x occurs must be developed. In the case where $X_1 > x$, we immediately have the conditional probability

$$P\{N = n \mid X_1 > x\} = \begin{cases} 1 & n = 1 \\ 0 & n > 1 \end{cases} \qquad (27)$$

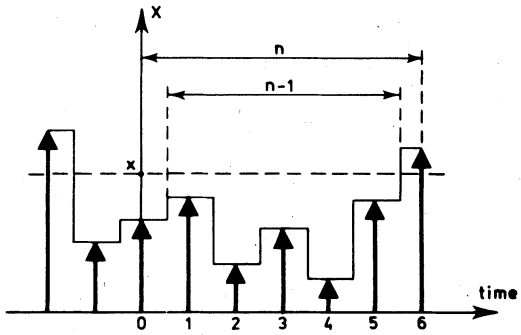

Fig. 4. Run below the level x with a randomly chosen zero time.

The case where $X_1 \leq x$ is sketched in Fig. 4.
Utilizing that the remaining part of the run-length, denoted by M_*, is distributed as the total one, the conditional probability becomes

$$P\{N = n \mid X_1 \leq x\} = \begin{cases} 0 & n = 1 \\ P\{M_* = n-1\} = h(n-1) & n > 1 \end{cases} \tag{28}$$

in which we on the basis of Eq. (22) can insert

$$h(n-1) = \frac{F_1 - F_2}{F_1} \left(\frac{F_1}{F_1}\right)^{n-2} \tag{29}$$

The probability mass function of N is now easily obtained

$$\begin{aligned} \theta(n) &= P\{N = n\} \\ &= P\{N = n \mid X_1 > x\}P\{X_1 > x\} + P\{N = n \mid X_1 \leq x\}P\{X_1 \leq x\} \\ &= \begin{cases} 1 - F_1 & n = 1 \\ (F_1 - F_2)\left(\frac{F_2}{F_1}\right)^{n-2} & n > 1 \end{cases} \end{aligned} \tag{30}$$

The probability distribution function then becomes

$$\Theta(n) = P\{N \leq n\} = 1 - F_1 \left(\frac{F_2}{F_1}\right)^{n-1} \tag{31}$$

Calculation of the mean and the variance gives

$$E\{N\} = 1 + \frac{F_1^2}{F_1 - F_2} \tag{32}$$

and

$$\mathrm{Var}\{N\} = \frac{F_1^2(F_1 + F_2 - F_1^2)}{(F_1 - F_2)^2} \tag{33}$$

The extreme value distribution, i.e. the distribution function for the maximum value of the process in the time period $1 \leq i \leq n$ follows immediately from Eq. (31)

$$\Psi(x) = P\{X'_n \leq x\} = P\{N > n\} = 1 - \Theta(n) = F_1\left(\frac{F_1}{F_2}\right)^{n-1} \tag{34}$$

When dealing with a standardized process $\{Y_i\}$ ($E\{Y_i\} = 0$ and $\mathrm{Var}\{Y_i\} = 1$) with normal marginal and bivariate distribution, the extreme value distribution turns out to be

$$\Psi(y) = \Phi(y)\left[\frac{\Phi_2(y,y;\rho_1)}{\Phi(y)}\right]^{n-1} \tag{35}$$

Eq. (35) is shown in Fig. 5 for different values of the correlation coefficient ρ_1 and a fixed number of time periods, $n = 30$. The figure shows how the probability of obtaining a maximum value above a given level in a given period of time decreases for increasing values of ρ_1.

By differentiation of Eq. (35) the corresponding probability density function can be obtained. Utilizing that

$$\Phi_2(y,y;\rho_1) = [\Phi(y)]^2 + \frac{1}{2\pi}\int_0^{\rho_1} e^{-y^2/(1+z)} \frac{1}{\sqrt{1-z^2}}\, dz \tag{36}$$

and suppressing argument y, the formula for the density becomes

$$\psi(y) = \left(\frac{\Phi_2}{\Phi}\right)^{n-2}\left[(2-n)\frac{\Phi_2}{\Phi}\phi + (n-1)\left(2\Phi\phi - \frac{1}{\pi}\int_0^{\rho_1} e^{-y^2/(1+z)} \frac{y}{(1+z)\sqrt{1-z^2}}\, dz\right)\right] \tag{37}$$

in which ϕ denotes the standard normal probability density function.

Returning to the longterm annual streamflow records considered ny Nordin and Rosbjerg (1970), we get the possibility of comparing the new extreme value distribution with observations. In Fig. 6, the theoretical and observed distributions of 5-year maximum values in the standardized series of Danube (ρ_1=0.1) and St. Lawrence

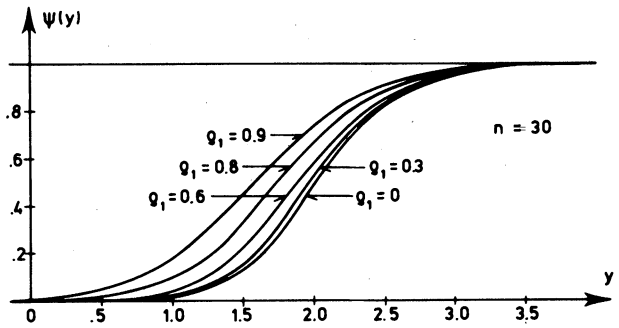

Fig. 5. Extreme value distribution for different values of the correlation coefficient ρ_1.

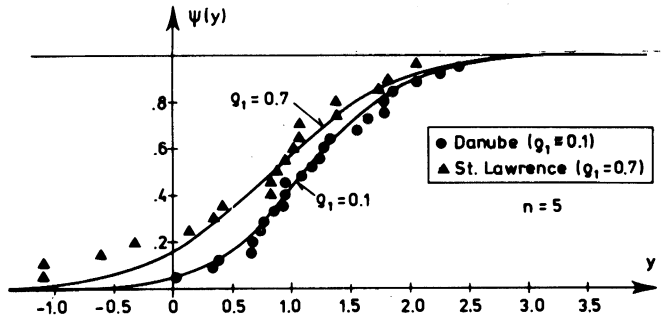

Fig. 6. Theoretical and observed distributions of 5-year maximum values in the standardized series of St. Lawrence (ρ_1=0.7) and Danube (ρ_1=0.1).

(ρ_1=0.7) are given. As seen from the figure, good agreement is obtained between the theory and the observations, emphasizing the applicability of the theory.

Let us finally compare the obtained extreme value distribution with the extreme value distribution developed by Ditlevsen (1971) corresponding to a stationary continuous standard normal process $\{Y_t\}$, ($E\{Y_t\}=0$ and Var $\{Y_t\}=1$). For large values of y, Ditlevsen found the following approximate distribution function

$$\Psi(y) = \Phi(y)\, e^{-\gamma[\phi(y)/\Phi(y)]\tau} \; ; \quad \gamma = \sqrt{\frac{-\rho''(0)}{2\pi}} \qquad (38)$$

in which τ is the considered period of time and $\rho''(0)$ the second derivative of the correlation function $\rho(s) = E\{Y_t\, Y_{t+s}\}$ in the zero point.

In Fig. 7, Eqs. (35) and (38) are compared using the approximation $\rho''(0) = 2(\rho_1 - 1)$ in Eq. (38). Again the number of time intervals is chosen to be n = 30. The figure illustrates the very good agreement between the formulas for large values of ρ_1. For decreasing values, the agreement becomes poorer.

The agreement for larger ρ_1-values is remarkable taking the applied simple Markov chain assumption into account, whereas the decreasing correspondence between the equations was to be expected because of the increasing diversity between the discrete and the continuous process for decreasing values of ρ_1.

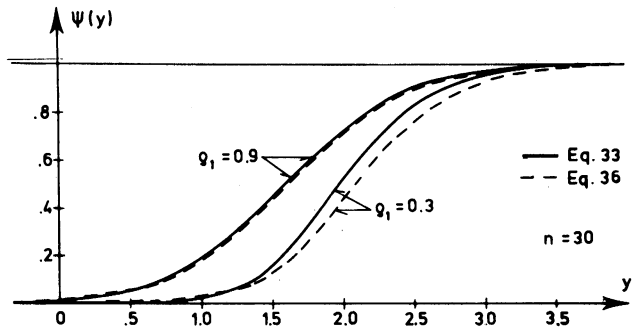

Fig. 7. Comparison of Eqs. (35) and (38).

Conclusions

In a stationary discrete process with Markov properties random variables in connection with crossings of the level x are considered. Distributions of the variables K (the time interval between successive events $X > x$), M (the number of time steps in succession where $X \leq x$, i.e. the run-length), N (the time interval until the first upcrossing of the level x occurs) and X'_n (the maximum value of the process in the time interval $1 \leq i \leq n$) are developed. Table 2 summarizes the distribution functions of K, M, N and X'_n. Utilizing the fact that the independent process appears when $F_2 = F_1^2$, we can make a simple check on the results, which is also shown in Table 2.

Special attention should be given to the distribution of X'_n, the extreme value distribution in the considered process. Good agreement is found between the theoretical extreme value distribution and extreme values obtained from long-term annual streamflow records.

Table 2 - Distribution functions of K, M, N and X'_n

	Persistence type in the series $\{X_i\}$	
Random variable	Markov	Independent
K	$1 - \frac{F_1 - F_2}{1 - F_1} \left(\frac{F_2}{F_1}\right)^{k-1}$	$1 - (F_1)^k$
M	$1 - \left(\frac{F_2}{F_1}\right)^m$	$1 - (F_1)^m$
N	$1 - F_1 \left(\frac{F_2}{F_1}\right)^{n-1}$	$1 - (F_1)^n$
X'_n	$F_1 \left(\frac{F_2}{F_1}\right)^{n-1}$	$(F_1)^n$

References

Ditlevsen, O. (1971) Extremes and first passage times. Technical University of Denmark, Copenhagen.

Gottschalk, L. (1976) Frequency of dry years. Nordic hydrological conference, 1976. Reykjavik. p. II 75-46.

Nordin, C.F., and Rosbjerg, D.M. (1970): Application of crossing theory in hydrology. Bul. Internat. Assoc. Scientific Hydrology, v. 15, no. 1, pp. 27-43.

Saldarriaga, J., and Yevjevich, V. (1970) Application of run-lengths to hydrologic series. Hydrology paper No. 40. Colorado State University, Fort Collins.

Address: Institute of hydrodynamics and hydraulic engineering,
ISVA, Technical University of Denmark,
Bldg. 115,
DK-2800, Lyngby, Denmark.

Received: 17 June, 1977