PEAKS OVER THRESHOLD METHOD IN COMPARISON WITH BLOCK-MAXIMA METHOD FOR ESTIMATING HIGH RETURN LEVELS OF SEVERAL NORTHERN MORAVIA PRECIPITATION AND DISCHARGES SERIES

DANIELA JARUŠKOVÁ[1)], MARTIN HANEK[2)]

[1)]Department of Mathematics, Faculty of Civil Engineering, Czech Technical University, Thákurova 7, CZ – 166 29 Praha 6, Czech Republic; mailto: jarus@mat.fsv.cvut.cz
[2)]Department of Statistics and Probability, Faculty of Mathematics and Physics, Charles University, Sokolovská 83, CZ – 186 75 Praha 8, Czech Republic.

The peaks over threshold method (POT method) is an alternative to the block-maxima method for estimating return levels (these are the levels that are exceeded by a daily precipitation, resp. by a daily average discharge, only with a given small probability; in statistical language they are called high quantiles) when the studied series are not long enough. The paper compares both methods for precipitation and discharges series of Northern Moravia. It is shown how to overcome problems in the POT method caused by autocorrelation and seasonality.

KEY WORDS: Precipitation and Discharges Series, High Return Levels, Peaks Over Threshold Method, Block-Maxima Method, Choice of Threshold, Declustering, Splitting the Series into More Homogeneous Parts.

Daniela Jarušková, Martin Hanek: POROVNÁNÍ ODHADU KVANTILŮ S DLOUHOU DOBOU OPAKOVÁNÍ METODOU ŠPIČEK NAD PRAHEM S METODOU BLOKOVÝCH MAXIM PRO SRÁŽKOVÉ A PRŮTOKOVÉ ŘADY ZE SEVERNÍ MORAVY. Vodohosp. Čas., *54*, 2006, 4; 17 lit., 2 obr., 13 tab.

Metoda špiček nad prahem může sloužit při odhadování vysokých kvantilů (to znamená takových hodnot, že je denní úhrnná srážka, respektive denní průměrný průtok, překročí jen s malou předem danou pravděpodobností). Jedná se o alternativu k metodě blokových maxim, a to zvláště tam, kde studované řady nejsou příliš dlouhé. Článek porovnává výsledky obou metod na příkladech srážkových a průtokových dat ze severní Moravy. Ukazuje dále, jak lze v metodě POT překonat problémy spojené s autokorelací a sezónností řad.

KLÍČOVÁ SLOVA: srážkové a průtokové řady, vysoké kvantily (odpovídající dlouhým dobám opakování), metoda špiček nad prahem, metoda blokových maxim, výběr prahu, metoda odstranění shluků, metoda štěpení řady na homogenní části.

Introduction

Estimating high annual return levels of precipitation and water discharges series is one of basic problems of statistical hydrology. A typical feature of this kind of problems is that the available series are very often shorter than hundred years and the objective is to estimate 50 and 100 years return levels, i.e. in the statistical language 98 % or 99 % quantiles of annual maximal values. Clearly, the series are not long enough to apply some nonparametric estimators and hence, some parametric model has to be chosen.

The central result of mathematical statistics, the so called Fisher – Tippett theorem, claims that after a linear transformation, the distribution of maxima of independent identically distributed random variables are asymptotically, i.e. for a large number of observations, distributed according to a generalized extreme value distribution, provided the limit exists. The distribution function of a generalized extreme value distribution (GEV) has the following form:

$$H_{\xi,\mu,\sigma}(x) = \exp\left\{-\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-1/\xi}\right\} \text{ if } \xi \neq 0 \quad (1)$$

$$H_{\xi,\mu,\sigma}(x) = \exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\} \text{ if } \xi = 0. \qquad (1)$$

Therefore, the annual maxima are usually supposed to follow a GEV distribution and its parameters are estimated either by the maximum likelihood method or the method of probability-weighted moments. In the statistical literature this approach is called a block-maxima method, see *Embrechts* et al. (1997) and *Beirlant* et al. (2004).

In hydrology and meteorology the block maxima are frequently annual maxima. Estimation of high quantiles using GEV distributions for annual maxima is routinely used by researchers of Czech Hydrometeorological Institute. Sometimes, a statistical inference is simplified by choosing a Gumbel distribution as a theoretical model, in other words it is supposed a priori $\xi = 0$. However, many researchers oppose this simplification claiming that precipitation as well as discharges data of smaller rivers and creeks are heavy tailed and using a general form of an extreme value distribution yields a better fit.

Despite the fact that the maximum likelihood method as well as the method of probability-weighted moments are both consistent, i.e., the estimates converge as the number of observations increases, for short series the variance of estimates of parameters of a GEV distribution may be large and the question arises why not to use "all large daily values" rather than the maximal annual values only. The "Peaks Over Threshold" method (the POT method) is a method where estimates of high annual return levels (high quantiles) are based on all values that exceed a certain value u which is called a threshold. Threshold methods have been used by hydrologists for a long time. The first systematic development was in the works of *Todorovic* and *Zelenhasic* (1970) and *Todorovic* and *Rouselle* (1971). Our approach is based on application of the generalized Pareto distribution. The idea of modeling exceedances over a threshold by a generalized Pareto distribution was first suggested by *Pickands* (1975) and developed later by *Smith* (1987), *Hosking* and *Wallis* (1987) and *Joe* (1987) amongst others. *Davidson* and *Smith* (1990) presented an overview paper that summarized all basic knowledge about the method and stressed importance of an appropriate modeling of seasonality and serial dependence. When applying the POT method the user has to choose some parameters subjectively, e.g., a threshold or a declustering parameter. There are no general rules how to choose them as the parameters are dependent on the character of the studied series. Therefore, many statisticians and hydrologists present their experience with the method so that their followers can use the method more easily. The papers by *Engeland* et al. (2004) and *Coles* et al. (2003) may serve as examples. Our paper belongs to those papers too and we hope that it will encourage the applied statisticians to use the method as in some cases it may give better results than the block-maxima method.

POT method

The theoretical background of the POT method is based on the following facts which are true asymptotically, i.e. for a high threshold u:

1. Excesses over a high threshold u can be modeled by a generalized Pareto distribution with the following distribution function:

$$G_{\xi,\beta}(x) = 1 - \left(1 + \xi\frac{x}{\beta}\right)^{-1/\xi} \text{ if } \xi \neq 0, \qquad (2)$$

$$G_{\xi,\beta}(x) = 1 - \exp\left(-\frac{x}{\beta}\right) \text{ if } \xi = 0.$$

2. The number of exceedances over a high threshold follows a Poisson process.
3. The number of exceedances over a high threshold u and the values of excesses are independent.

The procedure consists of choosing the subsequence $\{X_j\}$ from the basic sequence that exceeds a threshold u, calculating the values $\{X_j - u\}$ for those values that exceed the threshold u and estimating parameters ξ and β either by the maximum likelihood method or by the method of probability weighted moments, see *Embrechts* et al. (1997) and *Beirlant* et al. (2004). An estimator of the probability that at randomly chosen day the observation will be smaller than x ($x > u$) can be calculated as follows:

$$p(x) = 1 - \frac{n_u}{n}\left(1 + \hat{\xi}\frac{x-u}{\hat{\beta}}\right)^{-1/\hat{\xi}}, \qquad (3)$$

and an estimator of a quantile x_p, i.e. 1/(1-p) return level, has the form:

$$\hat{x}_p = u + \frac{\hat{\beta}}{\hat{\xi}}\left(\left(\frac{n}{n_u}(1-p)\right)^{-\hat{\xi}} - 1\right), \qquad (4)$$

where n is the number of all data, n_u – the number of the data that exceed the threshold u, and finally $\hat{\beta}$ and $\hat{\xi}$ are the obtained estimates of the parameters β and ξ. Notice that in case we wish to estimate annual return levels we have to replace p by $p^{1/365}$. That means it is assumed that the probability of exceedances of a certain value x is the same for all days, i.e. there is no seasonal effect, and moreover the events that in two subsequent days a value x is exceeded are independent.

When the POT method is applied to real data it is sometimes difficult to choose a right threshold u. In practice the threshold value u should be chosen so that a slight change of its value does not affect values of estimated quantiles significantly. In other words the value u should be chosen in a region where the tail behavior of the analyzed data is stable. This is easy to say but not so easy to do. In statistical textbooks and papers it is recommended to choose a value u above a 90 % quantile of all data.

The POT method is based on the assumption that the observed data can be considered to be a random sample, i.e., they can be modeled by a sequence of independent identically distributed random variables. That is hardly to expect of precipitation and water discharges series. In case the dependence is short one can use a declustering technique that consists in deleting a certain number of data around local maximal values. The number of the deleted data is a parameter of a declustering procedure that has to be chosen by users. Clearly, a declustering technique works well if the series has a "short memory", especially in high values. This is more typical for a precipitation series than for a discharges series.

If the behavior of a sequence is affected by a seasonality it is reasonable to split the series into smaller, more homogeneous sets, and to apply the POT method to these subsets. For instance, we can split the series into January daily values, February daily values etc. Let $p_i(x)$ denote a probability that in a randomly selected day of the i-th month, $i = 1, ..., 12$, a observation was smaller than a real value x ($x > u_i$) then the estimation of $p_i^{(x)}$ may be calculated as follows:

$$p_i(x) = 1 - \frac{n_u(i)}{n(i)}\left(1 + \hat{\xi}\frac{x-u_i}{\hat{\beta}_i}\right)^{-1/\hat{\xi}_i}, \qquad (5)$$

where u_i is a chosen threshold, $\hat{\beta}_i$ and $\hat{\xi}_i$ – the estimated parameters of a generalized Pareto distribution, $n(i)$ – the number of all data in all i-th months and $n_u(i)$ – the number of these values that are above the threshold u_i. The p *100 %* quantile is estimated as a solution of the equation:

$$\prod_{i=1}^{12}\left(p_i(x)\right)^{m_i} = p, \qquad (6)$$

where m_i is the number of days in the i-th month, e.g. $m_1 = 31$, $m_4 = 30$. It occurs very often that estimated quantiles are larger when the series is split than when the POT method is applied to all data.

Data

We have obtained a data set from the Czech Hydrometeorological Institute detailing 5 series of daily discharges averages of the Opava River and its tributary Opavice measured at Karlovice (KAVA I, KAVA II), Krnov (KRVA, KRCE) and Opava (OPAVA). The three last series started 1. 11. 1959 and ended 31. 10. 2003, while the first series in Karlovice (KAVA I) started 1. 11. 1963 and ended 31. 10. 1979 and the second one (KAVA II) started 1. 11. 1979 and ended 31. 10. 2003. For our statistical inference we combined the both Karlovice series. Further, we have obtained daily precipitation values of 9 meteorological stations in the Northern Moravia – Heřmanovice (HE), Karlovice (KA), Krnov (KR), Lichnov (LI), Opava (OP), Praděd (PR), Rejvíz (RE), Vidly (VI), Albrechtice – Žáry (ZY). The longest record spans 45 years (1/1/1960 – 6/2/2005) but most of the records are shorter with missing data.

For the obtained discharges series the basic characteristics are presented by Tab. 1 while for the precipitation series by Tab. 2.

T a b l e 1. Basic descriptive characteristics of discharges series.
T a b u l k a 1. Základní popisné statistiky průtokových řad.

	Number of data	Mean	90% quantile	Maximum
		[m^3 s^{-1}]	[m^3 s^{-1}]	[m^3 s^{-1}]
KAVA	14610	2.80	5.27	157
KRVA	16071	4.16	8.38	289
KRCE	16071	1.39	2.98	118
OPAVA	16071	6.96	14.00	553

T a b l e 2. Basic descriptive characteristics of precipitation series.
T a b u l k a 2. Základní popisné statistiky srážkových řad.

	Number of data (non-missing daily values)	Percentage of positive data (days with precipitation larger than 0.1 mm)	Mean	90% quantile	Max
			[mm]	[mm]	[mm]
HE	15 221	45.4 %	2.51	13.2	196.5
KA	16407	40.0 %	2.00	12.7	124.2
KR	16194	39.0 %	1.65	11.3	59.2
LI	16437	40.0 %	1.68	11.2	110.0
OP	16108	39.8 %	1.59	11.0	62.0
PR	13758	53.6 %	3.13	14.1	139.4
RE	14730	43.8 %	2.83	15.8	214.2
VI	15098	52.6 %	3.13	15.2	199.3
ZY	16347	48.7 %	2.06	11.2	125.0

Results of POT method and comparison with block-maxima method

Choice of threshold

Tabs 3 and 4 present the results of the POT method with thresholds corresponding to the 90 %, respectively to 95 % empirical quantiles of all discharges and precipitation data. The tables show the chosen threshold, the number of observations that exceed the threshold, the maximum likelihood estimates of the parameters β and ξ and the 50 and 100 years return levels, i.e. the estimates of the 98% and 99 % quantiles of the annual values. It seems that the estimates of return levels based on the thresholds corresponding to the 90 %, respectively 95 % empirical quantiles differ less for the precipitation series than for the discharges series, where the estimates of the return levels based on the thresholds corresponding to the 95 % empirical quantiles are in all cases larger than those based on the smaller thresholds corresponding to the 90 % empirical quantiles. As all our series are long we would recommend here to use a larger threshold. (Notice that we have to keep in balance the bias and the variance of the estimate. If we choose a larger threshold the bias decreases but the variance increases and vice versa. Many authors suggest to use a mean excess function for a right decision, for details see *Embrechts* et al. (1997)) .

Declustering

As the daily values of precipitation as well as discharges are dependent, we used also a declustering technique in a way that on both sides of a local maxima we deleted few neighboring observations. Tabs 5 and 6 present several first values of autocorrelation functions of excesses over a high threshold corresponding to 95 % empirical quantile of all observations. It seems that in the case of precipitation series it is enough to delete one neighboring observation (k = 1) while for discharges series we recommend to delete 3 neighboring observations (k = 3) because precipitation series have substantially shorter memory than discharges series. Tabs 7 and 8 present results of declustering.

T a b l e 3. Results of the POT method: estimated 50 and 100 years discharge based on the thresholds corresponding to the 90 % and 95 % empirical quantiles.
T a b u l k a 3. Výsledky metody špiček nad prahem: odhadnuté maximální průtoky s dobou opakování 50 a 100 let, založené na prazích odpovídajících 90% a 95% empirickým kvantilům.

	Threshold	Number of data	$\hat{\beta}$	$\hat{\xi}$	50 years discharge	100 years discharge
					[$m^3 s^{-1}$]	[$m^3 s^{-1}$]
KAVA	5.27	1457	2.242	0.356	90	115
KRVA	8.38	1604	3.820	0.337	139	176
KRCE	2.98	1599	2.009	0.374	87	113
OPAVA	14.00	1581	7.429	0.345	278	353
KAVA	6.98	724	2.711	0.414	110	146
KRVA	11.50	787	4.411	0.404	171	225
KRCE	4.58	800	2.431	0.425	102	138
OPAVA	19.70	794	8.397	0.435	373	502

T a b u l k a 4. Výsledky metody špiček nad prahem: odhadnuté maximální srážky s dobou opakování 50 a 100 let, založené na prazích odpovídajících 90 % and 95 % empirickým kvantilům.
T a b l e 4. Results of the POT method: estimated 50 and 100 years precipitation based on the thresholds corresponding to the 90 % and 95 % empirical quantiles.

Station	Threshold	Number of data	$\hat{\beta}$	$\hat{\xi}$	50 years precipitation	100 years precipitation
					[mm]	[mm]
HE	7.2	1511	7.157	0.273	184	227
KA	6.3	1640	6.617	0.126	89	101
KR	5.0	1611	6.820	0.035	63	70
LI	5.1	1640	6.341	0.103	77	87
OP	4.7	1602	6.556	0.095	76	86
PR	9.4	1368	7.198	0.217	145	172
RE	8.0	1472	8.574	0.219	171	204
VI	9.6	1497	8.087	0.175	135	157
ZY	6.3	1619	6.660	0.133	92	105
HE	12.4	758	9.175	0.240	170	206
KA	11.4	809	6.612	0.177	99	115
KR	9.7	809	7.710	–0.037	56	60
LI	9.5	821	7.492	0.045	69	76
OP	9.3	795	7.802	0.021	66	73
PR	14.7	688	8.337	0.223	148	177
RE	14.4	736	10.117	0.208	166	198
VI	15.6	749	8.796	0.199	142	167
ZY	11.0	813	7.894	0.081	83	92

T a b l e 5. Several first values of the autocorrelation function of the excesses over the thresholds corresponding to the 95% quantiles for the precipitation series when $k = 0$ and $k = 1$.
T a b u l ka 5. Několik prvních hodnot výběrové autokorelační funkce spočtené z velikostí přesahů prahů odpovídajících 95% kvantilům pro srážkové řady, jestliže $k = 0$ a $k = 1$.

Station	ar (0)	ar (1)	ar (2)	ar (3)
HE	1	0.265	0.010	0.150
KA	1	0.180	0.058	–0.020
KR	1	0.067	0.014	–0.025
LI	1	0.050	0.104	0.079
OP	1	0.093	0.062	0.013
PR	1	0.288	0.138	0.077
RE	1	0.315	0.034	0.038
VI	1	0.297	0.089	0.023
ZY	1	0.234	0.093	0.018
HE	1	–0.105	–0.024	0.042
KA	1	0.038	–0.089	–0.057
KR	1	–0.0857	0.034	0.013
LI	1	0.088	0.061	–0.052
OP	1	0.065	–0.093	0.031
PR	1	0.041	–0.086	0.041
RE	1	–0.0096	–0.032	–0.082
VI	1	–0.059	–0.096	0.055
ZY	1	0.036	–0.020	–0.034

T a b l e 6. Several first values of the autocorrelation function of the excesses over the thresholds corresponding to the 95% quantiles for the discharges series when $k = 0$ and $k = 3$.
T a b u l k a 6. Několik prvních hodnot výběrové autokorelační funkce spočtené z velikostí přesahů prahů odpovídajících 95% kvantilům pro průtokové řady, jestliže $k = 0$ a $k = 3$.

Station	ar (0)	ar (1)	ar (2)	ar (3)
KAVA	1	0.694	0.309	0.112
KRVA	1	0.700	0.293	0.097
KRCE	1	0.670	0.269	0.066
OPAVA	1	0.758	0.381	0.134
KAVA	1	–0.046	0.018	–0.069
KRVA	1	–0.051	0.117	–0.008
KRCE	1	–0.014	0.015	0.064
OPAVA	1	–0.050	0.067	–0.018

Seasonal adjustment

We split the original series into twelve data sets, each corresponding to one month, and modeled the tail behavior of all series by the POT method with the thresholds corresponding to the monthly 95 % empirical quantiles and the declustering parameter $k = 1$. To get a feeling how much the estimates of the parameters differ from month to month, Tab. 9

T a b l e 7. Results of the POT method: estimated 50 and 100 years discharge based on the threshold corresponding to the 95 % quantile using a declustering technique with $k = 0, 3$.
T a b u l k a 7. Výsledky metody špiček nad prahem: odhadnuté maximální průtoky s dobou opakování 50 a 100 let, založené na prazích odpovídajících 95% empirickým kvantilům při použití metody vedoucí k odstranění shluků s parametry $k = 0, 3$.

Station	Threshold	Number of data	$\hat{\beta}$	$\hat{\xi}$	50 years discharge	100 years discharge
					[$m^3 s^{-1}$]	[$m^3 s^{-1}$]
KAVA	6.98	724	2.711	0.414	110	146
KRVA	11.50	787	4.411	0.404	171	225
KRCE	4.58	800	2.431	0.425	102	138
OPAVA	19.70	794	8.397	0.435	373	503
KAVA	8.81	100	3.708	0.643	109	161
KRVA	14.00	108	7.605	0.498	184	273
KRCE	5.60	113	4.706	0.460	99	143
OPAVA	24.20	111	17.084	0.400	344	502

T a b l e 8. Results of the POT method: estimated 50 and 100 years precipitation based on the threshold corresponding to the 95 % quantile using a declustering technique with $k = 0, 1$.
T a b u l k a 8. Výsledky metody špiček nad prahem: odhadnuté maximální srážky s dobou opakování 50 a 100 let, založené na prazích odpovídajících 90% a 95% empirickým kvantilům při použití metody vedoucí k odstranění shluků s parametry $k = 0, 1$.

Station	Threshold	Number of data	$\hat{\beta}$	$\hat{\xi}$	50 years precipitation	100 years precipitation
					[mm]	[mm]
HE	12.4	758	9.175	0.240	170	206
KA	11.4	809	6.612	0.177	99	115
KR	9.7	809	7.710	-0.037	56	60
LI	9.5	821	7.492	0.045	69	76
OP	9.3	795	7.802	0.022	66	73
PR	14.7	688	8.337	0.223	148	177
RE	14.4	736	10.117	0.208	166	198
VI	15.6	749	8.796	0.199	142	168
ZY	11.0	813	7.894	0.081	83	92
HE	23.1	214	11.939	0.251	167	203
KA	18.5	255	8.887	0.113	89	101
KR	17.2	252	9.576	-0.202	49	51
LI	17.2	253	7.814	0.079	73	81
OP	17.0	246	10.175	-0.126	58	61
PR	24.6	186	10.690	0.236	145	175
RE	25.4	216	12.060	0.243	168	204
VI	26.8	205	11.208	0.174	130	152
ZY	20.0	228	9.192	0.069	82	91

presents the estimated parameters $\hat{\beta}_i$ and $\hat{\xi}_i$ of a generalized Pareto distribution for the series KAVA. It can be seen that the estimated parameters for the different months differ substantially and it means that the method of series splitting would be very useful. However, after a data splitting the estimates of "months parameters" together with the corresponding return levels are based on substantially less observations and that is why they are affected very strongly by outliers unless the data set is not very large. Tab. 10 shows the estimated 50 years return levels when the series was split into the months data sets.

T a b l e 9. Estimated parameters of generalized Pareto distribution for modeling tails for KAVA station.
T a b u l k a 9. Odhadnuté parametry zobecněného Paretova rozdělení pro modelování chvostů řady získané ve stanici KAVA.

Month	Threshold	$\hat{\beta}_i$	$\hat{\xi}_i$
January	3.50	1.379	0.341
February	4.53	1.310	0.007
March	6.66	1.606	0.050
April	10.7	4.936	-0.161
May	9.52	3.729	0.146
June	7.72	4.706	0.249
July	11.51	6.522	0.523
August	8.70	5.020	0.595
September	5.88	2.801	0.373
October	4.50	1.643	0.066
November	4.53	1.886	0.000
December	4.15	1.099	0.450

T a b l e 10. Estimated 50 years return level without and with a splitting.
T a b u l k a 10. Odhadnutý kvantil odpovídající době opakování 50 let, nepoužijeme-li štěpení řady, respektive použijeme-li štěpení řady.

Station	50years discharge without splitting	50 years discharge with splitting	50 years discharge with splitting excluding the year 1997
	[$m^3 s^{-1}$]	[$m^3 s^{-1}$]	[$m^3 s^{-1}$]
KAVA	96	114	82
KRVA	145	185	113
KRCE	94	109	87
OPAVA	299	320	220

Comparison of POT method with block-maxima method

To compare the POT method to the block-maxima method we also calculated the estimates of 50 and 100 years return levels using the annual maxima and the GEV (generalized extreme values) distribution. Tab. 11 shows the maximum likelihood estimators of the parameters μ, σ, ξ of the GEV distribution for the precipitation series and Tab. 12 shows the same for the discharges series. Comparing the estimates of the return levels for the precipitation series by the POT method (presented by Tabs 3 and 4) and the block-maxima method, we see a good agreement (all differences are in absolute value less than 11 mm) with an exception of the series HE and ZY.

We would like to mention that in a case that the analyzed series is very long, which enables to create very long blocks, respectively to choose a very high threshold, the estimates of the parameter ξ obtained by the block-maxima method and the POT method should be very close to each other. Fig. 1 shows the dependence between the altitude of the station and the estimated parameter $\hat{\xi}$ when the POT method was applied and Fig. 2 shows the same dependence when the block-maxima method was applied. Looking at Fig. 2 we observe that the values of $\hat{\xi}$ for HE and ZY series are outliers. Therefore, it seems that these estimates are not reliable.

In the case of the discharges series the block-maxima method based on the annual maximal values gives systematically higher estimates of the return levels that the POT method. Clearly, the block-maxima method is much more sensitive to one large observation (outlier) because we have to estimate parameters from a small set of data. Tab. 13 shows the change in the estimates of the return levels using the block maxima method when we exclude the values corresponding to the year 1997 of a huge flood in the Northern Moravia.

T a b l e 11. Results of block-maxima method applied to annual maxima: estimated 50 and 100 years precipitation.
T a b u l k a 11. Výsledky metody blokových maxim použité na roční maxima: odhadnuté maximální srážky s dobou opakování 50 a 100 let.

Station	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	50 years precipitation	100 years precipitation
				[mm]	[mm]
HE	48.127	23.635	0.030	146	165
KA	36.245	9.772	0.251	101	121
KR	30.073	5.282	0.066	53	58
LI	30.905	7.882	0.175	75	87
OP	31.983	7.090	0.048	62	69
PR	47.481	14.510	0.243	142	170
RE	51.718	17.065	0.266	169	206
VI	48.594	15.425	0.187	138	163
ZY	33.889	6.931	0.450	108	141

T a b l e 12. Results of block-maxima method applied to annual maxima: estimated 50 and 100 years discharge.
T a b u l k a 12. Výsledky metody blokových maxim použité na roční maxima: odhadnuté maximální průtoky s dobou opakování 50 a 100 let.

Station	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	50 years discharge	100 years discharge
				[$m^3 s^{-1}$]	[$m^3 s^{-1}$]
KAVA	11.881	7.936	0.446	95	132
KRVA	20.468	12.748	0.425	148	202
KRCE	9.115	6.847	0.534	99	146
OPAVA	36.617	23.742	0.451	290	403

T a b l e 13. Results of block-maxima method applied to annual maxima: estimated 50 and 100 years discharge with the value corresponding to the year 1997 omitted.
T a b u l k a 13. Výsledky metody blokových maxim použité na roční maxima: odhadnuté maximální průtoky s dobou opakování 50 a 100 let, jestliže byl z analýzy vypuštěn rok 1997.

Station	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	50 years discharge	100 years discharge
				[$m^3 s^{-1}$]	[$m^3 s^{-1}$]
KAVA	11.785	7.372	0.330	70	91
KRVA	20.653	12.058	0.258	102	127
KRCE	9.029	6.466	0.439	76	105
OPAVA	36.852	22.417	0.294	201	256

The POT method assumes that the data may be considered to be identically distributed. That is neither true for precipitation series nor for discharges series where the distribution is affected by a seasonality. In the Czech Republic the amount of precipitation is larger in summer months and that is why discharges are also higher in the summer. Moreover, discharges are also affected by spring snowmelt.

Conclusion

We have described two methods for estimating annual return levels (high quantiles): the block-maxima method dealing with annual maximal values and the peaks over threshold method.

The block maxima method is more straightforward because it is applied to the variables of interest – annual maxima. It consists in modeling ob-

served maximal values by a generalized extreme value distribution. As the number of the observed annual maxima is often small, i.e. less than 100, the obtained estimates may be sensitive to outliers. The method is asymptotic. That means it works well if the number of observations, from which a maximum is taken, is large. Many authors have shown that its convergence to a limit GEV distribution is often slow. In case the annual maxima are of interest, the maxima of 365 daily values are considered. This seems to be a long sequence but the convergence is slowed down by dependence between observations. The seasonal effect may also play an important role.

To apply the POT method is more difficult because a user himself has to choose a threshold *u*. In this paper we recommended to choose *u* equal to a 95 % quantile of all daily values. Sometimes, it is recommended to calculate estimates of interesting return levels for several threshold values, e.g. to set *u* equal to a 90 %, 95 %, 98% quantile and to check a stability of the obtained results. The quality and stability of estimated return levels is affected by many different properties of the studying data set. First, the method was suggested for the data that can be supposed to be independent variables. If the data are dependent a declustering technique may be applied. The technique works well for short memory series but it is not so reliable for series with a longer memory. Second, the seasonality may also affect the obtained estimates. If the series is long enough we may get better results if the series is split into several more homogeneous sets, e.g. one month periods, three months periods etc., and the POT method is applied to these data subsets. Third, the important property is a "smooth behavior of the data tail" which affects the speed of a convergence to a limit Pareto distribution.

The both described methods are asymptotic. Despite a theoretical proof that the assumption of independence may be relaxed, i.e. the asymptotics holds true even for short memory series, a positive autocorrelation slows down the speed to a limit distribution. The speed of a convergence is also largely affected by a "tail behavior" of the series. These all are reasons why application of statistical methods of theory of extremes has many opponents, see *Klemeš* (2000). We realize that their objections are reasonable and we do not claim that the described methods are always absolutely correct. They can rather help hydrologists to understand better the behavior of extremes of the observed data and they can provide them with an "objective assess" of large annual return levels.

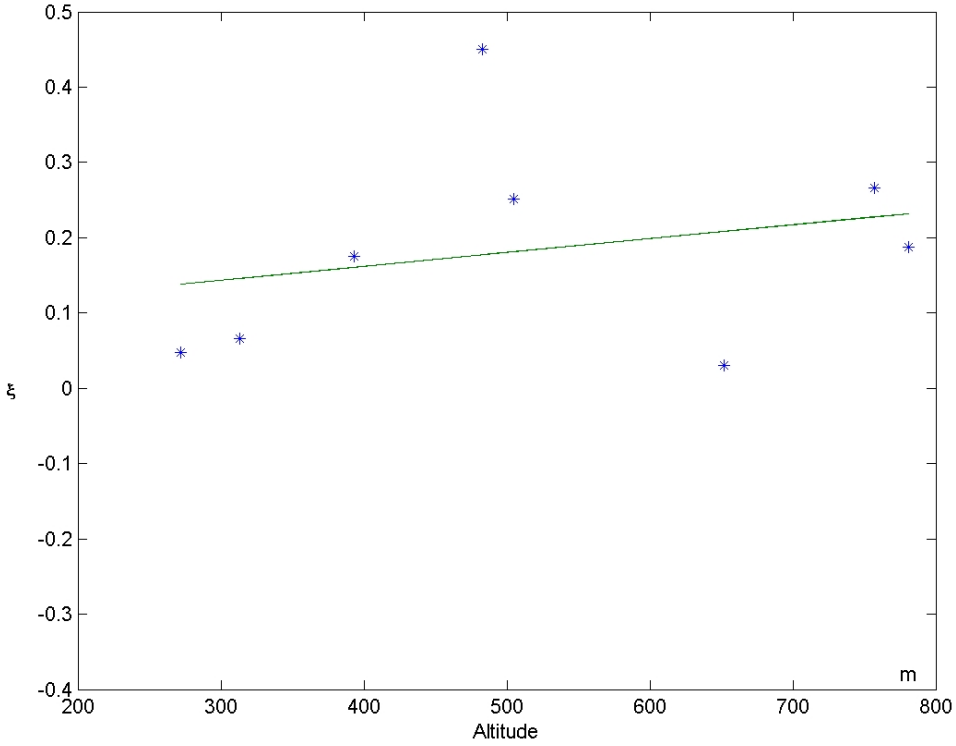

Fig. 1. Values of estimates of parameter ξ versus altitude of stations for the block-maxima method.
Obr. 1. Hodnoty odhadů parametru ξ v závislosti na nadmořské výšce stanice pro metodu blokových maxim.

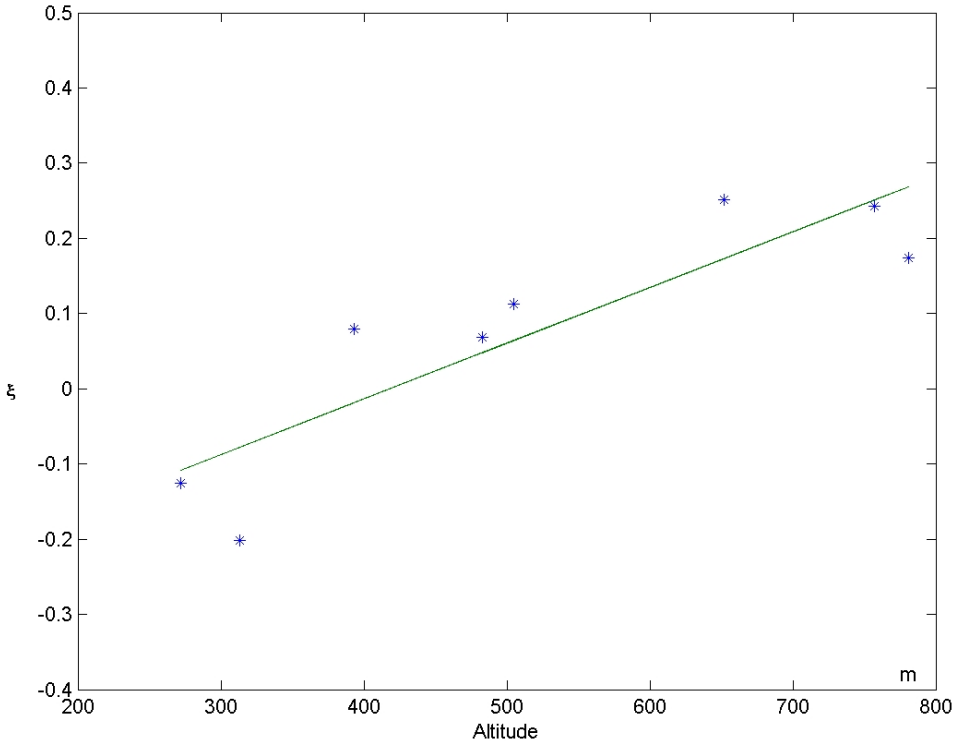

Fig. 2. Values of estimates of parameter ξ versus altitude of stations for the POT method.
Obr. 2. Hodnoty odhadů parametru ξ v závislosti na nadmořské výšce stanice pro metodu špiček nad prahem.

Acknowledgement. The study presented was partly carried out within the framework of the project MSM6840770002.

List of symbols

$H_{\xi,\mu,\sigma}(x)$ – distribution function of generalized extreme value distribution,
$\hat{\xi},\hat{\mu},\hat{\sigma}$ – estimates of parameters ξ,μ,σ of a generalized extreme value distribution,
$G_{\xi,\beta}(x)$ – distribution function of a generalized Pareto distribution,
$\hat{\xi},\hat{\beta}$ – estimates of parameters ξ, β of a generalized Pareto distribution,
u – threshold,
n – number of all observations,
n_u – number of the observations above a threshold u,
$p(x)$ – estimated probability that a variable distributed according to a distribution with a generalized Pareto tail exceeds a value x,
$\hat{x}_p$ – estimated $p100$ % quantile.

REFERENCES

BEIRLANT J., GOEGEBEUR Y., TEUGELS J., SEGERS J., DE WAAL D., FERRO CH., 2004: Statistics of extremes – theory and applications. Wiley, 490 p.

BOHÁČ M., BLAŽEK V., MACHÁČEK L., ŘEZÁČOVÁ D., KVĚTOŇ V., BLAŽKOVÁ Š., KULASOVÁ B., ŠERCL P., 2001: Vývoj metod pro stanovení extrémních povodní. [Souhrnná závěrečná zpráva grantového projektu VaV/510/3/97]. Ministerstvo životního prostředí, Praha.

CARTER D. J. T., CHALLENOR P. G., 1981: Estimating return values of environmental parameters. Q. J. R. Meteorolog. Soc., 108, 975–980.

COLES S., PERICCHI L. R., SISSON S., 2003: A fully probabilistic approach to extreme rainfall modeling. J. Hydrol., 273, 35–50.

DAVISON A.C., SMITH R. L., 1990: Models for exceedances over high thresholds. J. R. Statist. Soc. B, 52, 393–442.

EMBRECHTS P., KLUPPELBERG C., MIKOSCH T., 1997: Modelling extremal events. Springer, 645 p.

ENGELAND K., HISDAL H., FRIGESSI A., 2004: Practical extreme value modelling of hydrological floods and droughts: a case study. Extremes, 7, 5–30.

HOSKING J. R. M., Wallis J.R., 1987: Parameter and quantile estimation for the generalized Pareto distribution. Technometrics, 29, 339–349.

JARUŠKOVÁ D., 2004: Extrémy gaussovských posloupností a procesů. Robust 2004, vyd.. JČMF, 139–168.

KLEMEŠ V., 2000: Tall tales about tails of hydrological distributions. J. Hydrologic. Engng, 5, 227–239.

LEADBETTER M.R., ROOTZEN H., 1988: Extremal theory for stochastic processes. Annals of probability, *16*, 2, 431––478.

LEADBETTER M.R., LINDGREN G., ROOTZEN H., 1986: Extremes and related properties of random sequences and processes. Springer.

PICKANDS J., 1987: Statistical inference using extreme order statistics. Ann. Statist., 3, 119–131.
SMITH R. L., 1987: Estimating tails of probability distribution. Ann. Statist., 15, 1174–1207.
TODOROVIC P., ZELENHASIC E., 1970: A stochastic model for flood analysis. Water Resour. Res., 6, 1641–1648.
TODOROVIC P., ROUSSELLE J., 1971: Some problems of flood analysis. Water Resour. Res., 7, 1144–1150.
WAYLEN P., WOO M.K., 1982: Prediction of annual floods generated by mixed processes. Water Resour. Res., 18, 4, 1283–1286.

Received 7. November 2005
Scientific paper accepted 6. September 2006

POROVNÁNÍ ODHADU KVANTILŮ S DLOUHOU DOBOU OPAKOVÁNÍ METODOU ŠPIČEK NAD PRAHEM S METODOU BLOKOVÝCH MAXIM PRO SRÁŽKOVÉ A PRŮTOKOVÉ ŘADY ZE SEVERNÍ MORAVY

Daniela Jarušková, Martin Hanek

Metoda špiček nad prahem (metoda POT) je důležitá metoda k odhadování pravděpodobností výskytu velmi vysokých hodnot nebo k odhadování vysokých kvantilů. Tak například 98 % kvantilu ročních maximálních průtoků odpovídá padesátiletý průtok, podobně 99 % kvantilu stoletý průtok apod.

K odhadování těchto kvantilů se obvykle používá metoda blokových maxim, kde blokem jsou denní hodnoty během jednoho kalendářního roku. Metoda tak pracuje přímo s ročními maximálními hodnotami, jejichž statistické chování modeluje pomocí zobecněného extremálního rozdělení (GEV rozdělení) nebo někdy zjednodušeně pomocí Gumbelova rozdělení, které je speciálním případem GEV rozdělení, kde $\xi = 0$, viz (1). Parametry GEV rozdělení lze odhadnout buď metodou maximální věrohodnosti nebo metodou vážených momentů. Příslušné teoretické kvantily GEV rozdělení jsou pak odhady hledaných kvantilů.

Metoda POT využívá pro odhad vysokých kvantilů nikoliv maximální roční hodnoty, nýbrž všechny hodnoty, které překročí určitou vybranou mez, které se říká práh. Metoda spočívá v modelování chvostu rozdělení denních hodnot pomocí zobecněného Paretova rozdělení. Její výhodou je, že k odhadování obvykle používá větší množství dat. Subjektivním prvkem metody je volba prahu.

Článek porovnává odhadnuté hodnoty kvantilů pomocí metody POT s odhady získanými metodou blokových maxim pro srážková a průtoková data z jedné oblasti Severní Moravy. Zároveň ukazuje, jak lze metodu POT, která byla navržena pro nezávislé stejně rozdělené náhodné veličiny, aplikovat i na řady s krátkou pamětí, a jak se získané odhady liší. Pokud je řada denních hodnot silně ovlivněna sezónností, tj. ročním chodem, je možno řadu rozdělit na několik více homogenních částí, například měsíců, a aplikovat metodu POT pro tyto části. Výsledkem štěpení řady ovlivňované sezónností jsou poněkud výše odhadnuté hodnoty kvantilů.

Metoda POT je alternativní metodou k metodě blokových maxim. S výhodou se používá tam, kde jsou studované řady krátké. Jedná se o objektivní odhad vysokých kvantilů (odpovídající dlouhým dobám opakování) s určitými subjektivními prvky, které spočívají ve volbě prahu, a v případě, že se používá metoda odstranění shluků, též ve volbě parametru této metody. Je zřejmé, že metoda je založena na předpokladu „hladkosti chování chvostu rozdělení“, která umožňuje odhadnout extrémní kvantily, které leží mimo oblast dat, na základě naměřených velkých hodnot, tj. hodnot nad některým subjektivně zvoleným prahem.

Metoda POT vychází z asymptotických teoretických výsledků. Lze ji tudíž chápat jen jako metodu přibližnou, vycházející pouze z naměřených dat a nezahrnující jakoukoliv další hydrologickou informaci. Z těchto důvodů by měla být používána jen jako metoda pomocná.

Seznam symbolů

$H_{\xi,\mu,\sigma}(x)$ – distribuční funkce zobecněného extremálního rozdělení,
$\hat{\xi}, \hat{\mu}, \hat{\sigma}$ – odhady parametrů ξ, μ, σ zobecněného extremálního rozdělení,
$G_{\xi,\beta}(x)$ – distribuční funkce zobecněného Paretova rozdělení,
$\hat{\xi}, \hat{\beta}$ – odhady parametrů ξ, β zobecněného Paretova rozdělení,
u – práh,
n – počet všech pozorování,
n_u – počet pozorování nad prahem u,
$p(x)$ – odhadnutá pravděpodobnost toho, že proměnná s rozdělením, jehož chvost je modelován Paretovým rozdělením, překročí hodnotu x,
$\hat{x}_p$ – odhadnutý p100% kvantil.