PAPER • OPEN ACCESS

A new threshold selection method for peak over for nonstationary time series

To cite this article: C R Zhou *et al* 2016 *IOP Conf. Ser.: Earth Environ. Sci.* **39** 012071

View the article online for updates and enhancements.

A new threshold selection method for peak over for non-stationary time series

C R Zhou[1], Y F Chen[1, 3], S H Gu[2], Q Huang[1], J C Yuan[1] and S N Yu[1]

[1]Hohai University, College of Hydrology and Water Resources, Nanjing 210000, China
[2]Shanghai Hydrological General Station, Shanghai 200000, China

E-mail: 19870056@hhu.edu.cn

Abstract. In the context of global climate change, human activities dramatically damage the consistency of hydrological time series. Peak Over Threshold (POT) series have become an alternative to the traditional Annual Maximum series, but it is still underutilized due to its complexity. Most literature about POT tended to employ only one threshold regardless of the non-stationarity of the whole series. Obviously, it is unwise to ignore the fact that our hydrological time series may no longer be a stationary stochastic process. Hence, in this paper, we take the daily runoff time series of the Yichang gauge station on the Yangtze River in China as an example, and try to shed light on the selection of the threshold provided non-stationarity of our time series. The Mann-Kendall test is applied to detect the change points; then, we gave different thresholds according to the change points to the sub-series. Comparing the goodness-of-fit of the series with one and several thresholds, it clearly investigates the series that employs different thresholds performs much better than that just fixes one threshold during the selection of the peak events.

1. Introduction

In hydrology, the term non-stationarity refers to hydrological properties that are different from similar properties in the past. It has drawn wide attention, and many studies have contributed to non-stationary hydrological analysis[1-2]. Non-stationarity in floods can result from a variety of anthropogenic processes, including urban land-use modifications, climate change, and modifications to water infrastructure[3]. The interactions among these processes make it difficult to determine the cause of the non-stationarity.

Usually, flood frequency analysis (FFA) has a local basin perspective and is based on the assumption that floods are the consequences of random processes that are independent and identically distributed[4]; however, considering the changing environment, this assumption is no longer valid[5]. Therefore, to obtain a reliable flood estimation, new statistical models are needed to dynamically capture the changes of probability density functions over time. Numerous approaches[6-7] have been developed for FFA by using non-stationary observed data, namely, backward and forward restorations[8], the spatial Bayesian hierarchical approach[9], the climate-informed approach[10], and the Time- Varing Moments method[11].

Additionally, efficiently planning, designing, and operating hydrotechnical works require an in-depth understanding of the probabilistic behavior of extreme events[12]. Thus, the prerequisit lies in making a sound selection of extreme events. Extreme event series vary from each other due to

different sampling techniques. An annual maximum series (AMS) is extracted from a flow series. A POT series tries to retain all peak values that exceed a certain threshold value, so it is not confined to only one event per year.

The differences between the two methods have been frequently discussed, and it is generally agreed that the POT approach gives more robust results as long as certain pre-conditions are met, such as peak event independence and selecting appropriate threshold levels[13].

Compared with the AMS sampling method, POT enlarges the sample number; at the same time, it can reflect both the flood magnitude as well as the flood process. Moreover, POT modeling provides more flexibility in representing floods, and a more comprehensive description of the flood generation process [14].

The extreme events of discharge observations can be described by many candidate distributions. For POT series, the Generalized Pareto distribution (GPD) has been proved to be ideal for POT samples[15-16]. Thus, the GPD was used in this paper.

2. Materials and methods

2.1. Generalized pareto distribution

Picklands first introduced GPD to describe exceedances over high thresholds. The GPD cumulative distribution function (with the 3 parameters μ, σ, ζ) is given below:

$$G(\mu, \sigma, \zeta) = \begin{cases} 1 - (1 + \zeta \dfrac{x-\mu}{\sigma})^{-\frac{1}{\zeta}}, \zeta \neq 0 \\ 1 - \exp(-\dfrac{x-\mu}{\sigma}), \zeta = 0 \end{cases} \tag{1}$$

Where ζ is the shape parameter, σ is the scale parameter, and μ is the location parameter.

The distribution moments (order = 0, 1, 2) can be easily deduced as follows:

$$E(X) = \mu + \frac{\sigma}{1-\zeta} \tag{2}$$

$$Var(X) = \frac{\sigma^2}{(1-\zeta)^2(1-2\zeta)} \tag{3}$$

$$Skew(X) = \frac{2(1+\zeta)\sqrt{1-2\zeta}}{1-3\zeta} \tag{4}$$

From the Equation(4), we can find that if we use moment methods (regardless of the method of moments or the probability weighted moments), they are theoretically applicable only for $\zeta \neq \frac{1}{3}$. Besides, ζ cannot exceed 0.5, because when ζ tend to be 0.5, the variance will tend to be infinite. Some researchers have suggested using the moment estimators for $\zeta < \frac{1}{4}$[17].

The quantile function, which we pay more attention to, especially in FFA, is given by Equation (5):

$$x(G) = \begin{cases} \mu - \dfrac{\sigma}{\zeta}\left[1 - (1-G)^{-\zeta}\right] \zeta \neq 0 \\ \mu + \sigma\left[-\ln(1-G)\right], \zeta = 0 \end{cases} \tag{5}$$

2.2. The peak-over-threshold modeling

Because the POT approach focuses only on higher maximum values, it allows for a wider range of events that can be considered floods; therefore, it can capture more information concerning the whole flood phenomena than can its AMS counterpart. In terms of information about the flood process, the POT approach is more advisable in FFA.

Typically, a POT series is obtained by setting a threshold and selecting the highest values from a continuous hydrological time series. So, a POT sample can be described with the following statement:

If X is a random variable, we define X_s as the maximum value of X in an episode. An episode is defined as a function of a threshold level "s"; it begins when $X_{(t)}$ exceeds "s" and ends when $X_{(t)}$ falls below "s". All $X_{(t)}$ in each episode will form a POT series.

There is still no universal method for choosing an available threshold. Referring to the existing research, the POT series should meet two criteria: independence criteria and threshold selection rules. The two criteria are described below.

2.2.1. Independence criteria. Meeting the independence condition is a prerequisite of any statistical frequency analysis. Regarding the POT series, we assume that the mean occurrence number of POT events per year meets the Poisson process. Researchers raised several independence criteria, such as those proposed by Water Resources Council (1976) (see Equation(6)) and Cunnane (1979).

$$\theta < 5days + \log(A) \, or \, X_{\min} > 0.75 * \min[Q_1, Q_2] \tag{6}$$

Where θ is the time interval between two peak events, $days$; A is the basin area, $mile^2$; $X_{\min}$ is the minimum runoff, m^3/s; Q_1, Q_2 are the runoff of two peak events, m^3/s.

2.2.2. Threshold selection. Selecting the threshold should ensure that the peak values meet independence condition, but a considerably low threshold will lead to the selection of minor peaks that are not always independent. Raising the threshold will obviously increase the series independence. However, if the threshold is too high, fewer peak events will be retained.

Some rules are imposed on threshold selection, such as on the basis of a given return period. For instance, researchers [18] in China concluded that if the mean number of POT per year varied within a range of 2-3, the POT series rivers in China can be described by a Poisson process; Begueria insisted that the number should be within 1.2-5.0[19]; Cunnane suggested that if we choose the POT approach rather than the AMS method, the average number of POT samples must exceed 1.65[20].

2.3. A new method of threshold selection for non-stationary time series
Previous studies tended to employ only one threshold to the time series regardless of whether the series is stationary or not. To our knowledge, the properties of hydrological time series have changed significantly, so it is not advisable to use the same threshold as the non-stationary series.

In this paper, we suggest using different thresholds when the series is no longer stationary, and it is especially important to the change points in the time series. Through this approach, we can obtain a POT series with different thresholds due to a thorough consideration of the changes of characteristics in the series. The steps are listed below.

- Step 1: Use a Mann-Kendall test to detect whether or not the hydrological time series is stationary. If the time series is stationary, use the same threshold to the series. If the results indicate that the points have changed, proceed to the next step.
- Step 2: Assuming λ change points are detected, divide the series into $(\lambda+1)$ parts ($(\lambda+1)$ sub-series).
- Step 3: Give different thresholds to these sub-series according to the rule that ensures the average occurrence number of peak events per year in each sub-series within the range of 2-3 (referring to the results of Chinese scholars). Thresholds that are relatively lower are preferable, as they obtain more peak samples.

Additionally, independence criteria proposed by Water Resources Council was used in this paper. For the sake of comparison, we use two POT series for FFA. The first POT series was selected by employing the same threshold to the whole series, and the second series was given special attention to considering the non-stationarity in the series due to the existence of change points.

After the POT series was selected, we used the retained peak events to estimate the parameters in GPD and compare the goodness-of-fit of two POT series. All the parameters were estimated by the

Probability Weighted Moments (PWM) method. Mean relative errors were used to estimate the adequacy of the two theoretical curves.

3. Case study

3.1. The introduction to the study area

The Yangtze River is China's longest river. The Yichang gauge station is a hydrological station on the main stream of the Yangtze River. The Gezhouba Water Control Project and the Three Georges Dam are all on the upper stream of this river. They are 6 and 44 kilometers away from the Yichang gauged station, respectively (see figure 1). Due to inconsistent recorded data prior to the year 1946, this study only examined runoff data collected from the year 1946-2012.

Figure 1. Location of Yichang gauged station.

It is an undeniable fact that human activities have significantly affected hydrological properties. Because of the construction of the Three Georges Dam, the hydrologic cycle in the local watershed has been changed dramatically, especially downstream of the Yangtze River [21]. We applied the new method detailed in 2.3 to the peak runoff of Yichang gauge station.

3.2. Results

3.2.1. Results: peak over threshold series selection and change points detection. If we ignore the existence of the possible change points in the time series, we can obtain the first POT series. From figure 2, we can see that when the threshold equals $35000\,m^3/s$, the average number of peak events per year is close to 3, so $35000\,m^3/s$ is the minimum threshold. Thus, we get 182 peak events from the year from 1946 to 2012.

Additionally, figure 2 also reveals the absolute value of the average relative errors (ARE) between the empirical points and the theoretical curve (the parameters are estimated by PWM) of the series corresponding to different thresholds (35000, 36000, 37000, 38000, 39000, $40000\,m^3/s$). When the threshold is $37000\,m^3/s$, the absolute value of ARE reaches the minimum. Hence, $37000\,m^3/s$ is the perfect threshold with the best goodness-of -fit to the whole series.
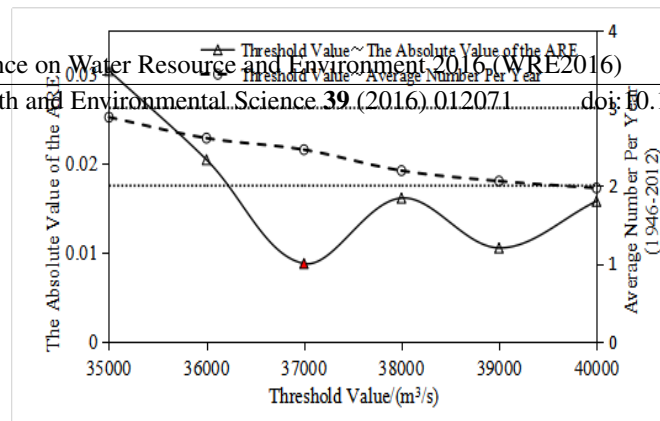
Figure 2. Relative curves between thresholds and
(a) the absolute value of ARE; (b) average number
(b) of peak events per year.

However, given human activities, some properties of our hydrological time series may have changed, so estimating hydrology information becomes a priority. The Mann-Kendall test was applied to test the mutability (or variant properties) of the POT series.

Researchers in China revealed through repeated trials that when the average occurrence per year was within 2~3[19], the theoretical curve fit with the observed data. Our hope is that the average occurrence per year is the largest one within the range of 2-3. Thus, we set $35000 \, m^3/s$ as the threshold (nearly 2.98 peaks per year) because $35000 \, m^3/s$ is the minimum threshold that can obtain as many peak events as possible.

Figure 3 below illustrates that two change points, in 1954 and 1984 respectively, separate the time series into three parts (1946-1954, 1955-1983, and 1984-2012). In each part, we can test its stationarity and the possible trend. The results show that each part of the series can be regarded as stationary.
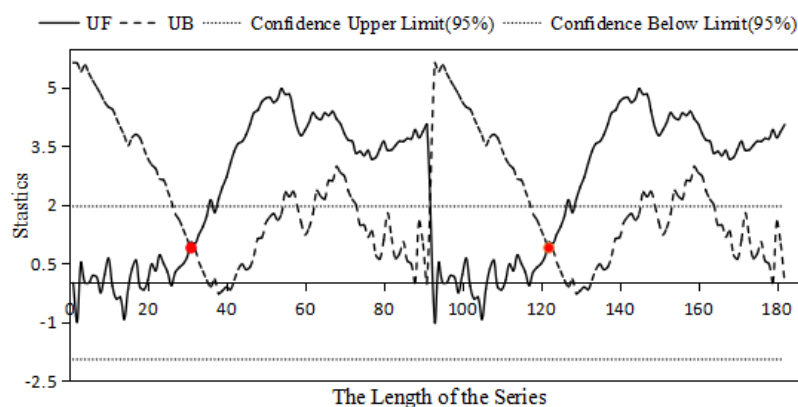

Figure 3. The results of the Mann-Kendall test (with a 95% significance level).

We take each part as a new population and study its threshold and peak event occurrence per year. Figure 4 illustrates how the mean numbers of over-threshold events per year change with the thresholds in the three sub-series. From figure 4, we can recognize the average occurrence per year of the first sub-series fell from 3.98 to 2.73, while the threshold increased from $35000 \, m^3/s$ to 40000 m^3/s, so we selected 38000 as the threshold (the average number of peak events per year was 2.889, the largest number within 2 to 3). The second sub-series showed a similar decrease as the first one; thus, we selected 36000 as the threshold to this sub-series (the average number of peak events per year is 2.9); $35000 \, m^3/s$ was selected as the threshold in the third sub-series. In conclusion, we gave three different thresholds ($38000 \, m^3/s$, $36000 \, m^3/s$, and $35000 \, m^3/s$) to three sub-series (1946-1953, 1954-1983 and 1984-2012, respectively).
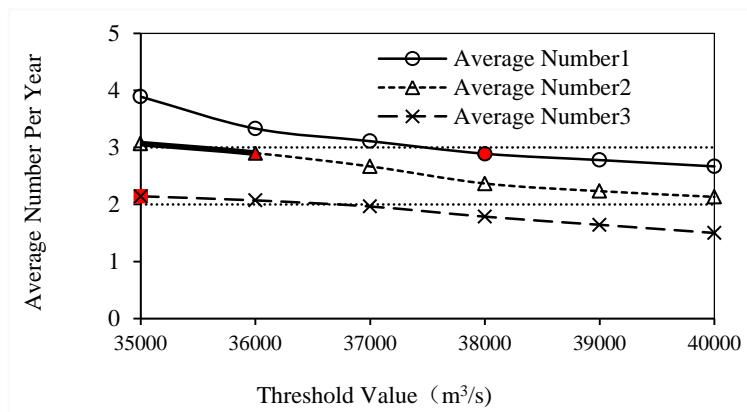
Figure 4. Selecting the threshold to the three sub-series according to how the average number of peak events per year changes with the thresholds.
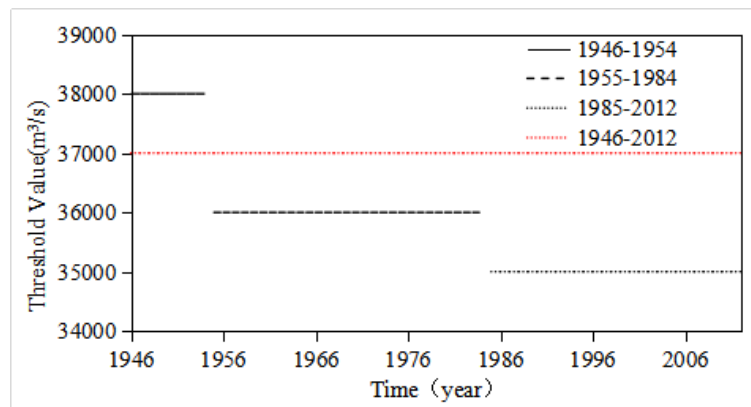

Figure 5. Comparison between the series threshold (daily runoff time series of 1946-2012, which ignores the non-stationarity) and the three sub-series (daily runoff time series of 1946-1953, 1954-1983 and 1984-2012, which takes change points into account).

3.2.2. Results: comparison of the goodness of fitting with the empirical points. According to the above threshold selection, we obtained two different POT series. Two POT series were used to FFA.

Samples of two series were applied to estimate the parameters in the GPD, and the PWM method is employed here [16].

In terms of frequency calculation, we should notice that the second POT series comprises data from three different sub-populations, so this data is less likely to be identically-distributed. For the sake of comparing the goodness-of-fit, the same frequency calculation method should be used. Consequently, the traditional frequency formula is applied here.

$$P(x=i) = \frac{i}{n+1}, i = 0,1,...,n \qquad (7)$$

Where n is the number of peak events.

Figure 6(a) and 6(b), respectively, show how the first and the second POT series fit with their theoretical curves. In figure 6(a), the theoretical curve is slightly above the empirical points. In figure 6(b), the theoretical curve fits much better than that in figure 6(a); most of the theoretical curve fits with the empirical points.
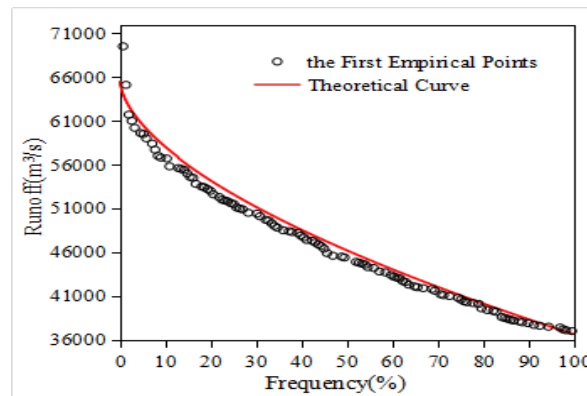
(a) The POT samples when the threshold equals $37000^{m^3/s}$.
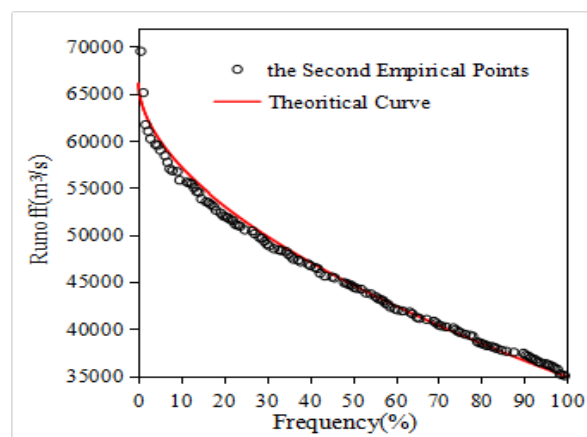
(b) The POT samples with different thresholds.

Figure 6. The comparison between the empirical curves and corresponding theoretical curves.

The average relative error (ARE) is used to compare the goodness-of-fit of two POT series, and the expression is given below (take μ as an example).

$$\delta = \frac{\mu_{obs} - \mu_{sim}}{\mu_{obs}} \tag{8}$$

Where μ_{obs} is the observed data of μ, and μ_{sim} is the theoretical value of μ.

Table 1 below shows the results. It is clear that regardless of the flood frequency, the theoretical curve of the POT series with full consideration to the change points shows a better fit with the empirical points. The theoretical curve obtained by the second POT series fits with the observed data especially well with the high frequency floods.

Table 1. The average relative errors between empirical points and the theoretical curves.

δ	The first POT series	The second POT series
Total	0.017253538	0.005606272
Low frequency flood	0.018960411	0.013587264
High frequency flood	0.015115555	-0.002906787

*Low frequency flood indicates a flood frequency higher than 50% ($P > 50\%$); if $P \le 50\%$, this indicates high flood frequency.

4. Conclusions

The above study sheds light on a new method to select a threshold when the POT approach is used to FFA. The conclusions are given below.

- Due to climate change and human activities, the characteristics of hydrological data may encounter dramatic changes. Estimating hydrology information enables us to have a deeper insight into the change.
- Quantile estimation by other POT series, which gives different thresholds to sub-series according to the change points, performs better compared to the POT series without considering the change points detection. Besides, the theoretical curve corresponding to the peak data extracted on the basis of different thresholds shows superior goodness-of-fit to the POT points, especially in the upper tail.

If a POT series is non-stationary, it is more advisable to distribute different thresholds according to the results of the Mann-Kendall test, rather than to employ the same one.

References

[1]   Tan X and Gan T Y 2015 Nonstationary analysis of annual maximum streamflow of Canada *Journal of Climate* **28**(5) 1788-805

[2]   Vasiliades L, Galiatsatou P and Loukas A 2015 Nonstationary frequency analysis of annual maximum rainfall using climate covariates *Water Resources Management* **29**(2) 339-58

[3]   Richard M V, Chad Y and Meghan W 2011 Nonstationarity: Flood magnification and recurrence reduction factors in the United States *Journal of the America Water Resources Association* **47**(3) 464-74

[4]   Lima C H, Lall U, Troy T J and Devineni N A 2015 Climate informed model for nonstationary flood risk prediction: Application to Negro River at Manaus, Amazonia *Journal of Hydrology* **522** 594-602

[5]   Tamer A G and Van Thanh Van N 2016 An at-site flood estimation method in the context of nonstationarity 1. A simulation study *Journal of Hydrology* **535** 710-21

[6]   Ye C Q, Chen X H, Zhang L J, *et al.* 2013 Changing ruled influence of threshold for peak-over-threshold flood series under the Changing environment in Wujiang *Advances in Water Science* **24**(3) 392-401

[7]   Xiong L H, Jiang C, Du T, *et al.* 2015 Review on Nonstationary Hydrological frequency analysis under Changing environments *Journal of Water Resources Research* **4**(4) 310-19

[8]   Liang Z M, Hu Y M and Wang J 2011 Advances in hydrological frequency analysis of non-stationary time series *Advances in Water Science* **22**(6) 864-71

[9]   Yan H and Moradkhani H 2015 A regional Bayesian hierarchical model for flood frequency analysis *Stochastic Environmental Research and Risk Assessment* **29**(3) 1019-36

[10]   Yan H and Moradkhani H 2016 Toward more robust extreme flood prediction by Bayesian hierarchical and multimodeling *Natural Hazards* **81**(1) 203-25

[11]   Du T, Xiong L H and Jiang C 2014 Nonstationary frequency analysis of rainfall time series in Weihe River Basin *Arid Land Geography* **37**(3) 468-79

[12]   Lang M, Ouarda T B M J and Bobée B 1999 Towards operational guidelines for over- threshold modeling *Journal of Hydrology* **225**(3-4) 103-17

[13]   Ngongondo C, Li L, Gong L, *et al.* 2013 Flood frequency under Changing climate in the upper Kafue River Basin, Southern Africa: A large scale hydrological model *Stochastic Environmental Risk Assessment* **27**(8) 1883-98

[14]   Zhang L J, Chen X H, Ye C Q, *et al.* 2013 Pot flood frequency analysis with historical floods in Wujiang River *Shuili Xue Bao* **3**(44) 268-75

[15]   Wang J F and Song S B 2010 Application of generalized pareto distribution in flood series

frequency analysis *Journal of Northwest A&F University(Nat&Sci Ed)* **38**(2) 191-6

[16] Bermudez P Z and Samuel K 2010 Parameter estimation of the generalized pareto distribution part-II *Journal of Statistical Planning and Inference* **140**(6) 1374-88

[17] Hosking J R M and Wallis J R 1987 Parameter and quantile estimation for the generalized Pareto distribution *Technometrics* **29**(3) 339-49

[18] Fang B, Guo S L, Chai X L, *et al*. 2005 Flood frequency analysis based on FPOT method *Water Power* **31**(2) 9-12

[19] Begueria S 2005 Uncertainties in partial duration series modeling of extreme related to the threshold value *Journal of Hydrology* **303**(1-4) 215-30

[20] Cunnane C 1973 A particular comparison of annual maxima and partial duration series methods of flood frequency prediction *Journal of Hydrology* **18**(3-4) 257-71

[21] Chen L, Xu W S, He X H, *et al*. 2006 Lower water level descending trend of Yichang section after water storing of Three Gorges and its solution *Yangtze River* (5) 59-64, 73