IOP Conf. Series: Journal of Physics: Conf. Series **890** (2017) 012154 doi:10.1088/1742-6596/890/1/012154

Extraction method of extreme rainfall data

Roslinazairimah Zakaria[1], Noor Fadhilah Ahmad Radi[1,2] and Siti Zanariah Satari[1]

[1]Faculty of Industrial Sciences and Technology, Universiti Malaysia Pahang, Lebuhraya Tun Razak, 26300 Gambang, Kuantan, Pahang, Malaysia
[2]Institute of Engineering Mathematics, Universiti Malaysia Perlis, Taman Bukit Kubu Jaya, Jalan Seraw, 02000 Kuala Perlis, Perlis, Malaysia

E-mail: roslinazairimah@ump.edu.my

Abstract. This study is aimed to describe step by step procedure in extracting extreme rainfall data series. Basically, the extraction of extreme rainfall data can be achieved using two methods, block maxima (BM) and peak over threshold (POT) methods. The BM method considers extracting the extreme rainfall data recorded each year during a specific duration, meanwhile the POT method is extracting all extreme rainfall data above a predefined threshold. Using the BM method, the regional pooling of 1-, 3-, 5- and 10-day are used and the maximum rainfall data are chosen among the pooled day within each year. For POT method, two methods are presented. Method 1 of POT method determines a threshold based on 95% percentile while Method 2 determines the threshold graphically using mean residual life plot and threshold stability plot. Based on the selection of the threshold value, a simulation study is conducted to identify the range of appropriate quantile estimate for a proper selection of the threshold value. For illustration of the methodology, daily rainfall data from the rainfall station at Klinik Chalok Barat, Terengganu is chosen. Both methods used are able to identify the extreme rainfall series. This study is important as it helps in identifying the good set of extreme rainfall series for further use such as in extreme rainfall modelling.

1. Introduction

The estimation of extreme rainfall data can be achieved using two methods, namely block maxima (BM) and peak over threshold (POT) methods. The BM method considers blocking the data for a specific duration and choosing the maximum rainfall amount that occurred in that particular duration [1]. The choice of block size in BM method can be critical, the blocks that are too small is likely to be poor, leading to bias in estimation and extrapolation and large blocks generate few block maxima, leading to large estimation of variance [1]. Hence, there must be a balance between the bias and the size of variances. In POT method, it identifies a certain threshold and extracting the extreme rainfall series above the chosen threshold from a continuous record, [1].

Recently, most procedures for extreme rainfall estimation are based on the BM series [2–6]. This is due to the simplicity of this method which suggests blocking the data accordingly and choosing the maximum rainfall amount within the study period. The BM series based on moving average is employed as suggested by Fowler and Kilsby [7] especially to derive a dimensionless number of regional quantile in order to determine the return period of the exceedances events. In this study, both methods, BM and POT are used to identify the extreme rainfall series.

The BM method considers one value of extreme for each blocks, while the POT method characterises extreme rainfall data if it exceeds a certain threshold. In the literatures, both BM and POT methods have been widely used to identify the extreme rainfall data. However, there is no standard procedure to choose the best method or to clarify which method is the best. For illustration of the methodology, meteorology station at Klinik Chalok Barat, Terengganu (5.4111°S, 102.8236°E) is chosen.

2. Methodology

The first step in modelling of extreme rainfall, requires the preparation of the extreme data series. Extracting the correct set of extreme data will ensure the extreme rainfall modelling is correct and valid for further use. The term extreme or maximum/ minimum rainfall refers to the amount of rainfall receive at a time more/ less than normal amount. The challenge is how to choose and define the suitable extreme value. This section presents two methods used for extracting extreme rainfall data, namely BM and POT methods.

2.1. Block Maxima

The BM method consists of grouping the data into blocks with equal length and fitting the data to the set of block maxima. Suppose that $X_1, X_2, \ldots, X_n$ is a sequence of independent and identically distributed (IID) having a common distribution function $F(x)$. The distribution of the maximum order statistic is given by

$$M_n = \max\{X_1, X_2, \ldots, X_n\} \tag{1}$$

The data in block maxima is usually modelled using the distribution from the Generalised Extreme Value (GEV) families such as Gumbel, Weibull and Fréchet, refer to [1] for details. Steps to determine extreme data series based on BM (moving average) method are as follows.

Step 1 Decide the block size.

Step 2 For 1-day BM (BM1), consider the daily rainfall data for a year and identify the maximum value. For example, if there are 12 years of data, there will be 12 maximum data. For 3-day BM (BM3), find the moving average for every 3 days and find one maximum value within that year. Similarly, these steps are repeated for 5-day BM (BM5) and 10-day BM (BM10) where the moving average of 5 and 10-day are calculated, the yearly maximum rainfall amount is identified and the maximum rainfall series is constructed.

The *R* codes for BM method are as follows.

```
library(extRemes)

## 1-day Block Maxima ##
bmone <- blockmaxxer(dt, blocks = dt$Year, which="Amount")

## 3-day Block Maxima ##
bmthree <- lapply(split(dt, dt$Year), function(x) max(sapply(1:(nrow(x)-2),
                    function(i) with(x, mean(Amount[i:(i+2)],na.rm=TRUE)))))

## 5-day Block Maxima ##
bmfive <- lapply(split(dt, dt$Year), function(x) max(sapply(1:(nrow(x)-4),
                     function(i) with(x, mean(Amount[i:(i+4)],na.rm=TRUE)))))

## 10-day Block Maxima ##
```

```
bmten <- lapply(split(dt, dt$Year), function(x) max(sapply(1:(nrow(x)-9),
                    function(i) with(x, mean(Amount[i:(i+9)],na.rm=TRUE)))))
```

Since the BM series only consider one value of extreme event each year, this will decrease the sample size of rainfall extreme events that is being considered. Hence, some of extreme events may not be included and thus extreme analysis may be misleading. Therefore, POT method is also studied in the following section.

2.2. Peak Over Threshold

The extreme series extracted using POT method can be modelled by the Generalised Pareto distribution (GPD) families. The details of GPD are discussed in [1]. According to [7], the extreme value theory (EVT) recommends the POT method over the BM method since the POT method is able to include all maximum events within the period of study. Some applications using POT series can be found from [8–11]. [12] highlights that the POT method improves the sampling number of extreme events since it may have more than one extreme value for each year. The POT method considers the distribution of the exceedances over a certain threshold, u. Consider the same sequences of $X_1, X_2, ..., X_n$ of IID random variables with distribution function $F(x)$ and maximum $M_n = \max\{X_1, X_2, ..., X_n\}$. The conditional probability that the value of X exceeds u by at most an amount y given that X exceeds u can be written as

$$P(X > u + y|X > u) = \frac{1 - F(u + y)}{1 - F(u)}. \tag{2}$$

Then the distribution function of M_n is

$$P(M_n \leq x) = P(X_1 \leq x, \ldots, X_n \leq x) = P(X_1 \leq x) \cdots P(X_n \leq x) = F^n(x) \tag{3}$$

In order to explore the probability of quantile estimate to detect a theoretical threshold, a simulation study is conducted. The purpose is to determine the theoretical quantile estimate that can be used to identify a suitably high threshold for extreme modelling using GPD. If the GPD is a valid model for the exceedances of u_0, then based on the threshold stability property, it will also be valid for all thresholds $u > u_0$. The mean of the threshold excess is given by

$$E[X - u|X > u] = \frac{\sigma_{u_0} + \xi u}{1 - \xi}$$

where σ_{u_0} and ξ are the scale and shape parameters of GPD respectively, for the exceedances of u_0. Choosing the correct threshold for POT method is subjective and always a difficult task. The best choice of the threshold depends on the issues of precision and biasness. If the threshold is too high that there are limited sample, then the analysis is unlikely to give good results. However, if the threshold is too low that most data are exceedances, then the analysis is likely to violate the asymptotic bias of the model, leading to bias [1] and will only be valid if the data can be fitted to the GPD. Hence, the optimum possible threshold needs to be identified carefully subjected to the constraint that extreme value model will provide a reasonable fit to exceedances of the chosen threshold.

In this study, two methods based on POT method are discussed. The first method (Method 1) is to identify the quantile position of the observed data and the data at this position will be considered as a threshold. For simplicity, 95% quantile is chosen as it is widely used in the literatures [11, 13–15]. Then, any value above the threshold is considered as extreme data. For the second method (Method 2), the theoretical quantile estimate is used to determine an optimal threshold through simulation process. Then the extreme data series obtained from Method 1 is fitted to GPD and the parameters are estimated. Using the estimated parameters, a set of

synthetic extreme data is generated. Next, the mean residual life plot and threshold stability plot (also known as parameter stability plot) are used to determine the optimal threshold graphically. Method 2 also will reconfirm that the optimal quantile estimate is used in Method 1. Steps to construct extreme data series based on the two POT methods are as follows.

Method 1

Step 1 Arrange the rainfall amount in ascending series

Step 2 Identify the rainfall amount at 95% percentile and let it be the threshold

Step 3 Extract all the rainfall amount above that threshold (extreme data)

Step 4 Fit the extreme data to GPD and estimate the parameters.

Method 2

Step 1 Generate a random data based on GPD estimated from Method 1

Step 2 Draw the mean residual life plot and threshold stability plot

Step 3 Identify the threshold and the associated quantile estimate in Step 2, compare with Method 1

Step 4 Extract all the rainfall amount above the identified threshold in Step 3

The *R* codes for POT method are as follows.

```
## Method 1 (Use percentile to determine threshold)
V30.dt <- dt[dt$Year %in% 1971:2012,]
dt_rank <- sort(V30.dt$Amount)
threshold_1 <- quantile(dt_rank,.95); threshold_1

## Extract extreme series above threshold
threshold <- threshold_1
dt_pot <- V30.dt[V30.dt$Amount > threshold, ]

## Method 2 (Visual inspection to determine threshold)
library(extrafont); library(fExtremes); library(extRemes)

## Mean Residual Life Plot ##
set.seed(1120)

## generate random number for Generalised Pareto Distribution
r <- rgpd(n = 10000, xi = 1/4, beta = 1)
mrlplot(r,cex.axis=0.8)

## Threshold Stability Plot ##
threshrange.plot(r, r = c(0, 20), nint=20)
```

3. Results and discussion

We demonstrate the BM and POT methods to extract extreme rainfall data. For BM method, the regional pooling of 1-, 3-, 5- and 10-day are used and the maximum rainfall amount are chosen among the pooled day within each year. For BM1, the yearly maximum rainfall amount in 1971 is 102.2 mm, 74.6 mm in 1972, 412.5 mm in 1989 and so on. The process is repeated for each year until 2012. Hence, there are 42 samples of extreme rainfall series ready to be used for extreme rainfall modelling. For BM3, the moving average rainfall for every three days are

Table 1. Sample of 3-day block maxima method.

Month	Day	Rainfall amount (mm)	Moving averages of 3-day
	1	55.81	
	2	49.02	
January	3	38.08	$(55.81 + 49.02 + 38.08)/3 = 47.64$
	4	17.77	$(49.02 + 38.08 + 17.77)/3 = 34.96$
⋮	⋮	⋮	⋮

Table 2. Sample of extreme series for BM1, BM3, BM5 and BM10.

Year	BM1	BM3	BM5	BM10
1971	102.18	70.82	50.03	36.93
1972	74.62	68.95	58.81	43.16
1973	110.66	102.72	85.24	65.37
⋮	⋮	⋮	⋮	⋮
2010	199.00	99.33	70.50	38.40
2011	219.50	162.83	136.30	81.20
2012	375.00	247.67	187.50	114.80

calculated as shown in table 1 and compared with all the 3-day moving average for each year and the maximum value is chosen. The same procedure is used to construct extreme series of BM5 and BM10. Samples of extreme series constructed are shown in table 2. These extreme series will be used in extreme rainfall modelling which is not in the scope of this study.

For Method 1 of POT method, figure 1 shows how the 95% threshold is identified at 47 mm. It is found that 762 data are above this threshold and is considered as extreme rainfall. For illustration of the simulation process for Method 2 of POT, a random sample of size $N = 10,000$ are simulated using GPD with parameters value are set arbitrarily to be $\xi = 0.25$ (shape) and $\sigma_{u_0} = 1.00$ (scale). Then, the mean residual life plot and threshold stability plot are drawn from the simulated data. Figure 2 shows an example of the mean residual life plot of 95% confidence interval of the simulated data. The interpretation of the residual life plot is not simple in practice, but it is helpful. From $u = 0$ to $u = 10$ mm, there is an increasing trend. However, the graph appears to be linear from $u_0 = 10$ mm and beyond approximately $u = 25$ mm it becomes unreliable due to high variability since only limited data is above such high thresholds and it decays sharply. Thus, the suitable threshold is $u_0 = 10$ mm for the data set. To further support our decision on the chosen threshold, the threshold stability plot is used which is based on the estimated GPD distribution of a range of thresholds, refer to figure 3. As pointed by Coles [1], for data above a level u_0 at which the GPD distribution is a valid model, the estimates of the shape $\widehat{\xi}$ and scale $\widehat{\sigma}_{u_0}$ parameters are approximated as constants.

4. Conclusion

In this study, we have identified the extreme rainfall series using block maxima (BM) and peak over threshold (POT) methods. For BM method, the extreme series is constructed based on moving average of 1-day, 3-day, 5-day and 10-day. For Method 1 of POT method, the threshold is determined based on 95% percentile. The data above that threshold is extracted. For Method

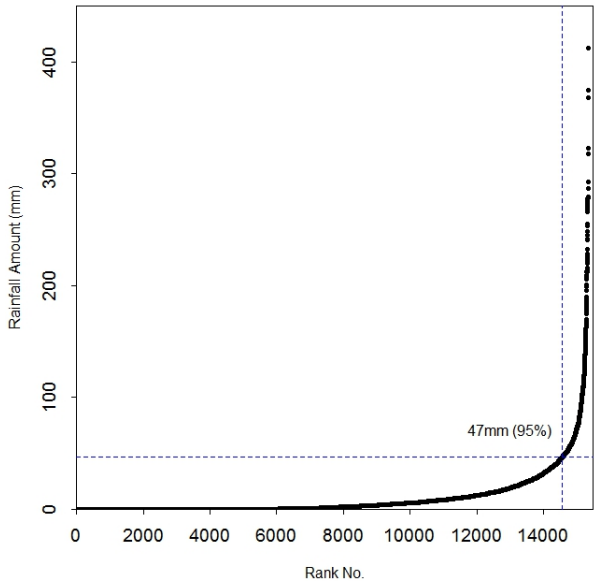

Figure 1. The 95% threshold of extreme rainfall using Method 1 of POT method.

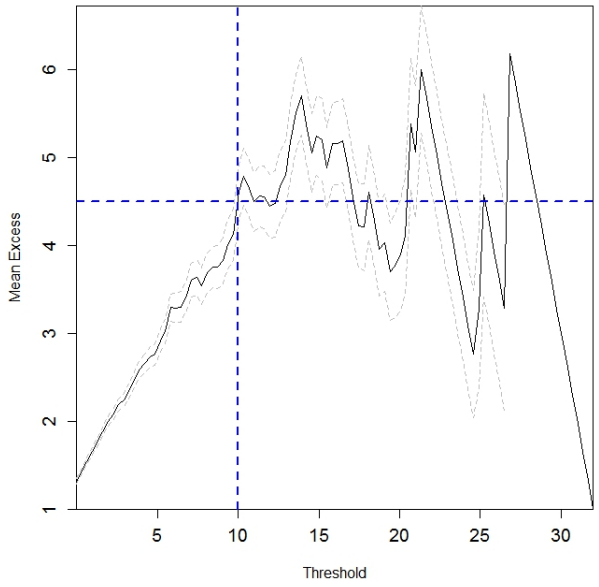

Figure 2. The mean residual life plot of the simulated data.

2 of POT method, the optimal threshold is determined graphically using mean residual life plot and the threshold stability plot to obtain the series of exceedances. The optimal threshold can be justified or readjusted based on Method 1 of POT method. Choosing the suitably high threshold is important to ensure the modelling of extreme data later is valid.

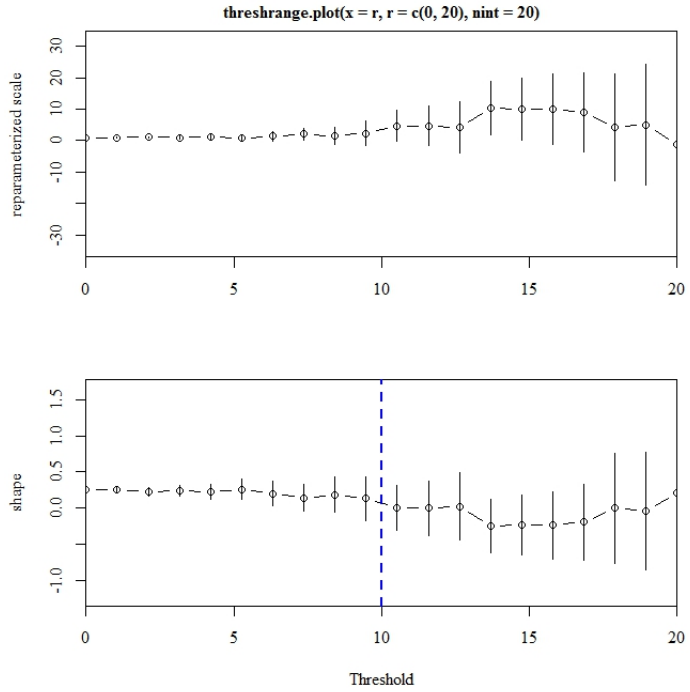

Figure 3. The threshold stability plot of the simulated data.

Acknowledgments
This study was supported by Universiti Malaysia Pahang (RDU120101 and RDU16117-FRGS). (http://www.ump.edu.my/)

References

[1] Coles S 2001 *An introduction to statistical modeling of extreme values* (London: Springer)
[2] Koutsoyiannis D and Baloutsos G 2000 *Natural Hazards* **22** 29–48
[3] Zalina M, Desa M, Nguyen V and Kassim A 2002 *Water Science & Technology* **45** 63–68
[4] Deka S, Borah M and Kakaty S 2009 *European Water* **27** 3–14
[5] de Carvalho J R P, Assad E D, de Oliveira A F and Pinto H S 2014 *Weather and Climate Extremes* **5** 7–15
[6] Mayooran T and Laheetharan A 2014 *Sri Lankan Journal of Applied Statistics* **15** 107–130
[7] Fowler H and Kilsby C 2003 *International Journal of Climatology* **23** 1313–1334
[8] Li Y, Cai W and Campbell E 2005 *Journal of Climate* **18** 852–863
[9] Wadsworth J L and Tawn J A 2012 *Biometrika* **99** 253–272
[10] Davison A C and Gholamrezaee M M 2012 *Proc. R. Soc. A* vol 468 pp 581–608
[11] Thibaud E, Mutzner R and Davison A C *Water Resources Research* **49**
[12] Mailhot A, Lachance-Cloutier S, Talbot G and Favre A C 2013 *Journal of Hydrology* **476** 188–199
[13] Beguera S, Angulo-Martnez M, Vicente-Serrano S M, Lpez-Moreno J I and El-Kenawy A *International Journal of Climatology* **31**
[14] Anagnostopoulou C and Tolika K 2012 *Theoretical and Applied Climatology* **107** 479–489
[15] Saidi H, Ciampittiello M, Dresti C and Ghiglieri G 2013 *Hydrology and Earth System Sciences Discussions* **10** 6049–6079