proceedings

MDPI

Type of the Paper (Abstract, Meeting Report, Preface, Proceedings, etc.)

The r-largest four parameter kappa distribution†

Yire Shin [1], Piyapatr Busababodhin [2] and Jeong-Soo Park [1,*]

[1] Department of Mathematics & Statistics, Chonnam National University, South Korea; shinyire@hanmail.net
[2] Department of Mathematics, Mahasarakham University, Thailand;
* Correspondence: jspark@jnu.ac.kr;

Abstract: The generalized extreme value distribution (GEVD) has been widely used to model the extreme events in many areas. It is however limited to using only block maxima, which motivated to model the GEVD dealing with r-largest order statistics (rGEVD). The rGEVD which uses more than one extreme per block can significantly improves the performance of the GEVD. The four parameter kappa distribution (K4D) is a generalization of some three-parameter distributions including the GEVD. It can be useful in fitting data when three parameters in the GEVD are not sufficient to capture the variability of the extreme observations. The K4D still uses only block maxima. In this study, we thus extend the K4D to deal with r-largest order statistics as analogy as the GEVD is extended to the rGEVD. The new distribution is called the r-largest four parameter kappa distribution (rK4D). We derive a joint probability density function (PDF) of the rK4D, and the marginal and conditional cumulative distribution functions and PDFs. The maximum likelihood method is considered to estimate parameters. The usefulness and some practical concerns of the rK4D are illustrated by applying it to Venice sea-level data. This example study shows that the rK4D gives better fit but larger variances of the parameter estimates than the rGEVD. Some new r-largest distributions are derived as special cases of the rK4D, such as the r-largest logistic (rLD), generalized logistic (rGLD), and generalized Gumbel distributions (rGGD).

Keywords: r-largest order statistics; Hydrology; Annual maximum sea level

Citation: Lastname, F.; Lastname, F.; Lastname, F. Title. *Proceedings* **2021**, 65, x. https://doi.org/10.3390/xxxxx

Received: date
Accepted: date
Published: date

1. Introduction

The generalized extreme value distribution (GEVD) has been widely used to analyse univariate extreme values (Coles 2001). The GEVD encompasses all three possible asymptotic extreme value distributions predicted by large sample theory. The cumulative distribution function (cdf) of the GEVD is as follows (Hosking and Wallis 1997):

$$F_3(x) = exp\left\{-(1-k\frac{x-\mu}{\sigma})^{1/k}\right\}$$

When $1-k(x-\mu)/\sigma > 0$ and $\sigma > 0$, where μ, σ, k are the location, scale, and shape parameters, respectively. The particular case for k = 0 in (1) is the Gumbel distribution. Note that the sign of k is changed from the book of Coles (2001).

One difficulty of applying the GEVD is using the limited amount of data for model estimation. Since extreme values are scarce, making effective use of the available information is important in extremes. This issue has motivated the search for a model to use more data other than just block maxima. The inclusion of more data up to r-th order statistics in each block other than just maxima will improve precision of model estimation, but the interpretation of parameters is unaltered from the univariate GEVD for block maxima. The above univariate result was extended to the r-largest order statistics model, which gives the joint density function of the limit distribution (Coles 2001);

$$f_3\left(\underline{x}^{(r)}\right) = \exp\left\{-\omega(x^{(r)})^{1/k}\right\} \times \prod_{s=1}^{r} \sigma^{-1}\omega(x^{(s)})^{\frac{1}{k}-1}$$

$$where\ x^{(1)} \geq x^{(2)} \geq \cdots \geq x^{(r)},\ and\ w\left(x^{(s)}\right) = 1 - k\frac{x^{(s)} - \mu}{\sigma} > 0\ for\ s = 1,2,\cdots,\ r$$

The rGEVD was encouraged to use by Zhang (2004), and has been employed in some real applications (Soares and Scotto 2004; An and Pandey 2007; Wang and Zhang 2008; Feng and Jiang 2015; Naseef and Kumar 2017). The number r comprises a bias-variance trade-off: small values of r generate few data leading to high variance; large values of r are likely to violate the asymptotic support for the model, leading to bias (Coles 2001). Bader et al.(2017) developed automated methods of selecting r from the rGEVD.

The inclusion of more data up to r-th order statistics in each block other than just maxima will improve precision of model estimation, but the interpretation of parameters is unaltered from the univariate GEVD for block maxima. For small to moderate sample sizes, the GEVD sometimes yields inadequate results. It may be because the GEVD is derived by a large sample theory for the extremes of independent sequences.

As a generalization of some common three-parameter distributions including the GEVD, the four parameter kappa distribution (K4D) was introduced by Hosking (1994). It can be useful in fitting data when three parameter distributions including the GEVD are not sufficient to capture the variability of observations. Some researchers studied on the K4D (Dupuis 1997; Dupuis and Winchester 2001; Singh and Deng 2003; Park and Kim 2007; Murshed et al. 2014).

The probability density function (pdf) of K4D is,

$$f_4(x) = \sigma^{-1}\omega(x)^{(1/k)-1}F_4(x)^{1-h}$$

$where\ w(x) = 1 - k\frac{x-\mu}{\sigma},\ \mathrm{F_4(x)} = \left\{1 - \mathrm{h}\omega(\mathrm{x})^{1/\mathrm{k}}\right\}^{1/\mathrm{h}}$ is the cdf of the K4D.

The K4D includes many distributions as special cases, as shown in Figure1 the generalized Pareto distribution for h=1, the GEVD for h=0, the generalized logistic distribution for h=-1, the generalized Gumbel distribution for k=0, the Gumbel distribution for h=0, k=0. The K4D is flexible and widely applicable to the data including not only extreme values but also skewed data. It has been used in many fields, particularly in hydrology and atmospheric sciences, for fitting extreme values or skewed data (e.g., Parida 1999; Park and Jung 2002; Seo et al. 2015; Kjeldsen et al. 2017; Brunner et al. 2019; Jung and Schindler 2019). Hosking and Wallis (1997) employed the K4D in regional frequency analysis as a parent distribution from which the samples are drawn. Blum et al.(2017) found that the K4D provides a very good representation of daily streamflow across most physiographic regions in the conterminous United States.

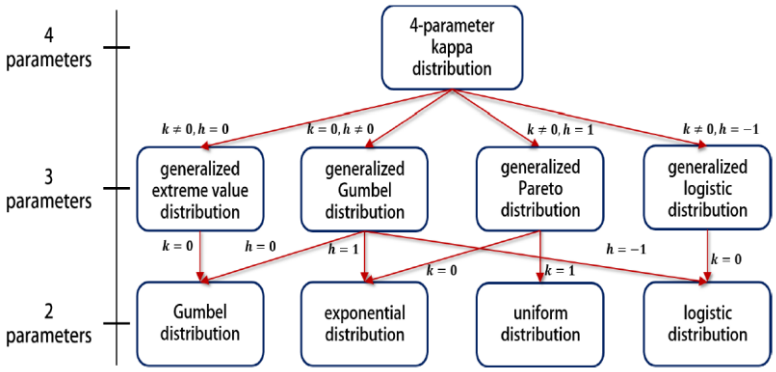

Figure 1. Relationship of the four parameter kappa distribution (K4D) to other distributions, which indicates a wide coverage of K4D.

In analyzing extreme values, the K4D has the same limitation of using only the block maxima as the GEVD has Like as the GEVD was extended to rGEVD, an extension of the K4D to r-largest order statistic model may be very useful to address this limitation. The inclusion of more observations up to r-th order statistics other than just maxima will improve precision of model estimation. The extension in the K4D is not published yet.

In this study, we thus developed an r-largest order statistics model as an extension of the K4D as well as of the rGEVD. It is referred to the rK4D. Figure2 illustrates our motivic schema. The remainder of this paper is organized as follows. Section 2 includes the definition of the rK4D. Section 3 some practical concerns of the rK4D by applying it to Venice sea-level data. Section 4 concludes with discussion.

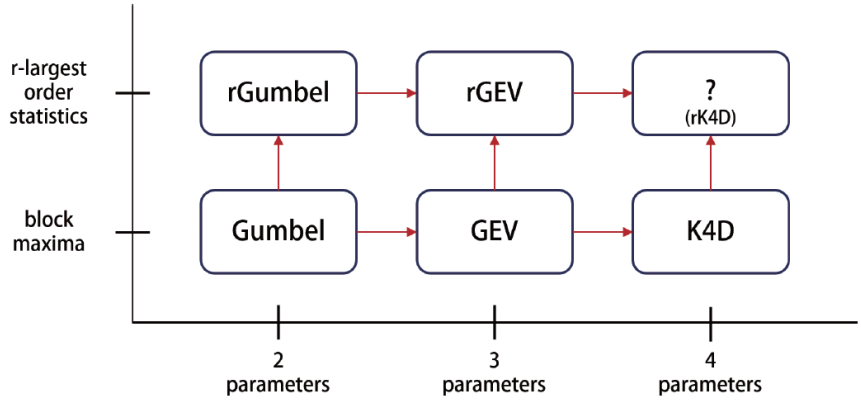

Figure 2. A motivic schema on generalizations from 2 parameters to 4 parameters, and extensions from the block maxima models to the r-largest order statistic models, which leads to the r-largest four parameter kappa distribution (rK4D).

2. r-largest four parameter kappa distribution

2.1. Definition of the rK4D

The r-largest four parameter kappa distribution (rK4D) is not the result from any theoretical derivation but just an analogous extension from the K4D and the rGEVD. To define the joint probability density function (pdf) of the rK4D, we considered and followed the generalization processes from the GEVD to the K4D and to the rGEVD.

We define the joint pdf of the Rk4d; under k ≠0, h ≠0,

$$f_4\left(\underline{x}^{(r)}\right) = \rho^{-r} C_r \times g(\underline{x}^r) \times F_4(x^{(r)})^{1-rh}$$

$$C_r = \begin{cases} \prod_{i=1}^{r-1} [1-(r-i)h] & if\ r \geq 2 \\ 1 & if\ r = 1 \end{cases}$$

$$g\left(\underline{x}^r\right) = \prod_{s=1}^{r} \omega(x^{(s)})^{\frac{1}{k}-1}$$

The supports of this pdf are $x^{(1)} \geq x^{(2)} \geq \cdots \geq x^{(r)}$, $\sigma > 0$, $w\left(x^{(s)}\right) > 0\ for\ s = 1,2,\cdots, r$, $C_r > 0$, $and\ 1 - h \times w\left(x^{(r)}\right)^{1/k} > 0$, When $r = 1$, this pdf is same as the pdf of the K4D in (3), when $h \to 0$, this pdf goes to the pdf of the rGEVD.

3. Real application : Venice sea-level data

These data consist of the 10 largest sea-levels in Venice over the period 1931-1981, except for the year 1935 (Coles 2001). The rK4D model is fitted to the values for r=1,2,…,10. The MLE of parameters and the 20-year return levels with standard errors in the parenthesis for several values of r are given in Table 1. For comparison, similar results from the fitted rGEVD are also presented. The upper table is for the rGEVD and the lower one is for the rK4D. In Table 1, the standard errors of parameter estimates decrease with increasing values of r for the rGEVD. That is not obvious in the rK4D but generally shows a decreasing trend. These non-monotonic decreasing cases may be because of the trouble in numerical optimization with 4 parameters in the rK4D or the intrinsic property of the rK4D. The SEs of $\hat{\mu}$, $\hat{\sigma}$, and $\hat{k}$ in the rK4D are generally bigger than those in the rGEVD. The SEs of h estimates in the rK4D are much larger compared to those of the other parameter estimates.

Table 1. The estimates of parameters and 20-year return level (r20) with standard errors (se) of the estimates in parenthesis which are obtained from the r-largest order statistic models fitted to Venice sea-level data with different values of r. Upper table is for the rGEVD and lower one is for the rK4D. `nllh' stands for the negative log-likelihood function value.

r	nllh	$\hat{\mu}$ (se)	$\hat{\sigma}$ (se)	$\hat{k}$ (se)	rGEV r_{20}(se)
1	222.7	111.1 (2.6)	17.2 (1.8)	-0.077(0.074)	156.7 (6.2)
2	379.5	114.5 (1.9)	15.0 (1.2)	-0.056(0.057)	155.6 (5.6)
3	515.4	117.3 (1.8)	14.8 (0.9)	-0.097(0.040)	155.6 (4.4)
4	632.2	118.3 (1.7)	14.3 (0.8)	-0.099(0.035)	155.0 (4.1)
5	732.0	118.6 (1.6)	13.7 (0.8)	-0.088(0.033)	154.3 (4.0)
6	829.6	118.8 (1.5)	13.4 (0.7)	-0.086(0.031)	154.0 (3.9)
7	916.5	119.1 (1.5)	13.2 (0.7)	-0.090(0.029)	153.6 (3.7)
8	995.7	119.6 (1.4)	13.1 (0.7)	-0.097(0.025)	153.3 (3.4)
9	1064.3	119.8 (1.4)	12.9 (0.6)	-0.098(0.024)	153.0 (3.3)
10	1139.1	120.5 (1.4)	12.8 (0.6)	-0.113(0.020)	152.8 (2.9)

r	nllh	$\hat{\mu}$ (se)	$\hat{\sigma}$ (se)	$\hat{k}$ (se)	$\hat{h}$	rK4Dr_{20}(se)
1	221.8	120.0 (5.2)	9.0 (2.4)	-0.16 (0.057)	-1.67 (1.34)	156.7 (6.2)
2	372.6	116.9 (2.4)	10.2 (1.3)	-0.23(0.064)	-1.31 (0.58)	155.6 (5.6)
3	499.8	118.0 (2.1)	10.4 (1.1)	-0.10 (0.051)	-1.03 (0.32)	155.6 (4.4)
4	610.6	117.2 (1.9)	10.9 (1.0)	-0.10(0.048)	-0.83 (0.24)	155.0 (4.1)
5	705.4	116.9 (2.0)	11.5 (1.1)	-0.13(0.050)	-0.77 (0.21)	154.3 (4.0)
6	803.8	117.0 (1.9)	12.0 (1.1)	-0.10(0.052)	-0.61 (0.17)	154.0 (3.9)
7	889.4	116.9 (1.8)	12.2 (1.0)	-0.08(0.048)	-0.49 (0.14)	153.6 (3.7)
8	961.9	117.1 (1.8)	11.9 (0.9)	-0.06(0.042)	-0.49 (0.13)	153.3 (3.4)
9	1023.0	117.2 (1.8)	11.8 (0.9)	-0.06(0.039)	-0.52 (0.13)	153.0 (3.3)
10	1089.1	117.2 (1.7)	11.4 (0.8)	-0.03(0.033)	-0.49 (0.12)	152.8 (2.9)

The 20-year return levels and its standard errors (SE) decrease with r in rGEVD, whereas those values for rK4D do not show a monotonic decrease. This phenomenon for the return levels of the rK4D is probably explained by that the return level and its SE are obtained for the annual maximum while the rK4D is fitted to the r-largest order statistics. Because the parameter estimates of the rK4D are obtained to take account into all data up to the r-largest observations, it may not work good for the annual maximum only.

This phenomenon may be more serious for the rK4D than the rGEVD because the standard errors of the return levels of the rK4D are greater than those of the rGEVD. This

is a re-confirmation of the general rule that the model with more parameters usually results in bigger variance (and less bias) than the model with fewer parameters (James et al. 2013).

Table 2. The Akaike information criteria (AIC), the Bayesian information criteria (BIC), and the trace and the log determinant of the covariance matrix (V) of the parameter estimates which are obtained from the r-largest order statistic models (the rGEVD and the rK4D) fitted to Venice sea-level data with different values of r.

r	rGEVD				rK4D			
	AIC	BIC	tr(V)	log|V|	AIC	BIC	tr(V)	log|V|
1	451.4	457.2	10.16	-2.31	451.7	459.4	34.72	-3.56
2	764.9	770.7	5.12	-4.55	753.2	761.0	7.92	-6.14
3	1036.9	1042.6	4.16	-6.01	1007.5	1015.2	5.49	-7.73
4	1270.5	1276.3	3.49	-6.98	1229.1	1236.9	4.84	-8.83
5	1469.9	1475.7	3.06	-7.63	1418.7	1426.4	5.01	-9.31
6	1665.3	1671.1	2.86	-8.09	1615.5	1623.2	4.85	-9.84
7	1839.0	1844.8.	2.67	-8.60	1786.7	1794.5	4.41	-10.68
8	1997.4	2003.2	2.48	-9.10	1931.7	1939.4	4.05	-11.22
9	2134.6	2140.4	2.34	-9.48	2054.0	2061.7	4.00	-11.48
10	2284.2	2292.0	2.16	-10.06	2186.2	2194.0	3.41	-12.25

Table 2 provides the Akaike information criteria (AIC), the Bayesian information criteria (BIC), and the trace and the determinant of the covariance matrix (V) of the parameter estimates. The AIC and the BIC are defined as

$$\mathrm{AIC(p)} = -2l(\hat{\theta}) + 2p, \ \mathrm{BIC(p)} = -2l(\hat{\theta}) + pln(m)$$

where $l(\hat{\theta})$ is the log-likelihood function evaluated at the parameter estimates $\hat{\theta}$, m is the sample size, and p is the number of parameters. In Venice sea-level data, m=50. These criteria are employed to select a preferred model by the rule that smaller is better. The trace tr(V) is the sum of variances, and the determinant $|V|$ is interpreted as the volume of V occupied by the probability dispersion it describes. The $|V|$ is thus sometimes called the generalized variance. It increases as the variances of parameter estimates increase; but also decreases as the correlations among the parameter estimates increase (Wilks 2011).

In Table 2, the AIC and the BIC are smaller in the rK4D for each r than the corresponding values in the rGEVD, except for the case r=1. The rK4D is preferable to the rGEVD for every r except for r=1. The tr(V)s in the rK4D for each r are greater than those in the rGEVD, whereas the log$|V|$s in the rK4D are smaller than those in the rGEVD. This means that there are more correlations among parameter estimates in the rK4D than in the rGEVD. The tr(V) and the log $|V|$ in the rK4D (and in the rGEVD) decrease monotonically as r increases. The biggest decreases in these values occur at the change from r=1 to r=2. That is, the variance decreases relatively a lot while the bias is not much increase, as r changes from 1 to 2. This observation leads to the interpretation that the biggest benefit of employing the rK4D over the K4D is obtained at r=2, for this data.

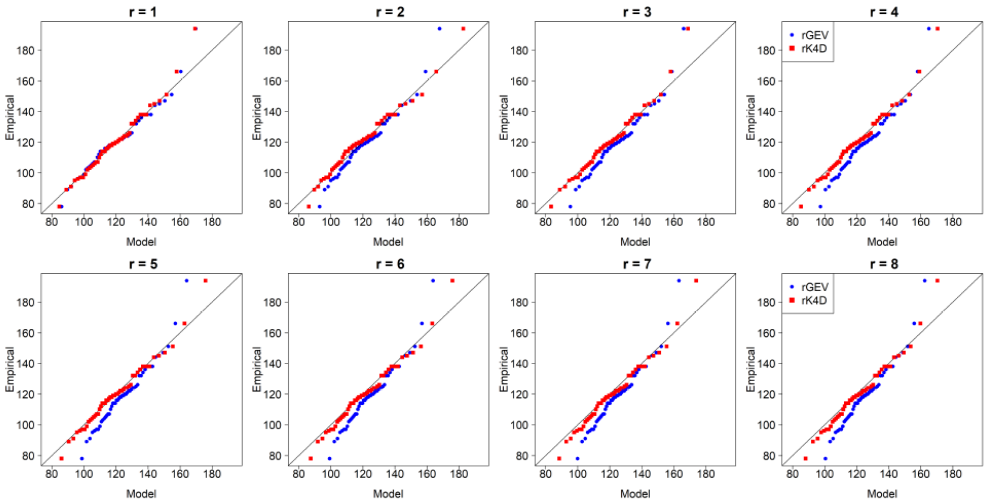

Figure 3. Quantile-per-quantile plots obtained from the largest order statistics for the rK4D fit and for the rGEV fit to Venice sea-level data with several values of r.

Figure 3 shows quantile-per-quantile plots obtained from the largest (s=1) order statistics for the rK4D fit (red points) and for the rGEV fit (blue points) to Venice sea-level data with several values of r. In this figure, one can see that the rK4D fits the data better than the rGEVD. We thus infer the rK4D provides less biased predictions than the rGEVD, because the rK4D with 4 parameters is more flexible than the rGEVD with 3 parameters.

4. Conclusion and discussion

In this study, we introduced the r-largest four parameter kappa distribution (rK4D). Application to Venice sea-level data is presented with comparison to the r-largest GEVD. This study illustrates that the rK4D gives better fitting or less biases but larger variances of the parameter estimates than the rGEVD. The pdf definition of the rk4d may not be unique, because it is not a result from any theoretical derivation but just an analogous extension from the K4D and the rGEVD. A point process approach for extremes (Smith 1989; Coles 2001) may provide a theoretical insight. The rK4D, as an extension of the rGEVD, can serve to model the r-largest observations flexibly with less bias than the rGEVD, specially when three parameters in the rGEVD are not enough to capture the variability of observations well. Even though there are defects such as larger estimation variance in the rK4D compared to the rGEVD, the introduction of the rK4D will enrich and improve our modelling methodology for extreme events.

Acknowledgments: This work was supported by the BK21 FOUR funded by the Ministry of Education, Korea (No. 5120200913674) and the National Research Foundation of Korea (NRF) grant funded by the Korean government (No.2020R1I1A3069260, No. 2021R1A6A3A13044162).

Conflicts of Interest: Declare conflicts of interest or state “The authors declare no conflict of interest.” Authors must identify and declare any personal circumstances or interest that may be perceived as inappropriately influencing the representation or interpretation of reported research results. Any role of the funders in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript, or in the decision to publish the results must be declared in this section. If there is no role, please state “The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results”.

References

1. Ahmad MI, Sinclair CD, Werritty A (1988) Log-logistic flood frequency analysis. Journal of Hydrology 98:215-224
2. An Y, Pandey MD (2007) The r largest order statistics model for extreme wind speed estimation. Journal of Wind Engineering and Industrial Aerodynamics 95:165-182. https://doi.org/10.1016/j.jweia.2006.05.008
3. Bader B, Yan J, Zhang XB (2017) Automated selection of r for the r largest order statistics approach with adjustment for sequential testing. Statistics and Computing 27:1435-1451. https://doi.org/10.1007/s11222-016-9697-3
4. Blum AG, Archfield SA, Vogel RM (2017) On the probability distribution of daily streamflow in the United States. Hydrology and Earth System Sciences 21:3093-3103. https://doi.org/10.5194/hess-21-3093-2017
5. Coles S (2001) An introduction to statistical modeling of extreme values. Springer, London, pp 224.
6. Dupuis DJ (1997) Extreme value theory based on the r largest annual events: a robust approach. Journal of Hydrology 200:295-306. https://doi.org/10.1016/s0022-1694(97)00022-x
7. Dupuis DJ, Winchester C (2001) More on the four-parameter kappa distribution. Journal of Statistical Computation and Simulation 71:99-113. https://doi.org/10.1080/00949650108812137
8. Feng JL, Jiang WS (2015) Extreme water level analysis at three stations on the coast of the Northwestern Pacific Ocean. Ocean Dynamics 65(11):1383-1397 DOI:10.1007/s10236-015-0881-3
9. Hosking JRM (1994) The four-parameter kappa distribution. IBM Journal of Research and Development 38:251-258.
10. Hosking JRM, Wallis JR (1997) Regional Frequency Analysis: An Approach Based on L-Moments. Cambridge University Press, Cambridge. pp 244. https://doi.org/10.1017/CBO9780511529443
11. Jung C, Schindler D (2019) Wind speed distribution selection - A review of recent development and progress. Renewable & Sustainable Energy Reviews 114:UNSP109290 DOI:10.1016/j.rser.2019.109290
12. Park JS, Kim TY (2007) Fisher information matrix for a four-parameter kappa distribution. Statistics \& Probability Letters 77(13):1459–1466.
13. Murshed MS, Seo YA, Park JS (2014) LH-moment estimation of a four parameter kappa distribution with hydrologic applications. Stochastic Environmental Research and Risk Assessment 28:253-262. https://doi.org/10.1007/s00477-013-0746-6
14. Naseef TM, Kumar VS (2017) Variations return value estimate of ocean surface waves - a study based on measured buoy data and ERA-Interim reanalysis data. Natural Hazards and Earth System Science 17(10):1763-1778 DOI:10.5194/nhess-17-1763-2017
15. Parida BP (1999) Modelling of Indian summer monsoon rainfall using a four-parameter kappa distribution. International Journal of Climatology 19:1389-1398. https://doi.org/10.1002/(sici)1097-0088(199910)19:12<1389::Aid-joc435>3.0.Co;2-t
16. Seo YA, Lee Y, Park JS, Kim MK, Cho C, Baek HJ (2015) Assessing changes in observed and future projected precipitation extremes in South Korea. International Journal of Climatology 35:1069-1078. https://doi.org/10.1002/joc.4039
17. Singh VP, Deng ZQ (2003) Entropy-based parameter estimation for kappa distribution. Journal of Hydrologic Engineering 8:81-92. https://doi.org/10.1061/(asce)1084-0699(2003)8:2(81)
18. Soares CG, Scotto MG (2004) Application of the r largest-order statistics for long-term predictions of significant wave height. Coastal Engineering 51:387-394. https://doi.org/10.1016/j.coastaleng.2004.04.003
19. Wang JF, Zhang XB (2008) Downscaling and projection of winter extreme daily precipitation over North America. Journal of Climate 21(5):923-937 DOI: 10.1175/2007JCLI1671.1
20. Zhang XB, Zwiers FW, Li GL (2004) Monte Carlo experiments on the detection of trends in extreme values. Journal of Climate 17:1945-1952.