The 4TH International Electronic Conference on Atmospheric Sciences

16–31 JULY 2021 | ONLINE

The r-largest four parameter kappa distribution

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

[1]Department of Mathematics & Statistics, Chonnam National University, South Korea, [2]Department of Mathematics, Mahasarakham University, Thailand

*Corresponding author, E-mail : jspark@jnu.ac.kr

1. Introduction GEVD, r-largest GEVD, K4D

1.1 GEVD

The generalized extreme value distribution (GEVD) has been widely used to analyse univariate extreme values (Coles 2001). The cumulative distribution function(cdf) of the GEVD is as follows (Hosking and Wallis 1997).

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

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.

1.2 r-largest GEVD

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). This model is referred to the rGEVD.

$$f_3\left(\underline{x}^{(r)}\right) = \exp\left\{-\omega(x^{(r)})^{1/k}\right\} \times \prod_{s=1}^{r} \sigma^{-1}\omega(x^{(s)})^{\frac{1}{k}-1} \qquad (2)$$

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

1.3 K4D

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.

The probability density function (pdf) of K4D is, for $k \neq 0, h \neq 0, \sigma > 0$

$$f_4(x) = \sigma^{-1}\omega(x)^{(1/k)-1}F_4(x)^{1-h} \qquad (3)$$

$$where\ w(x) = 1-k\frac{x-\mu}{\sigma} \qquad F_4(x) = \left\{1-h\omega(x)^{1/k}\right\}^{1/h} \qquad (4)$$

is the cdf of the K4D. Note that a new shape parameter h is added from the GEVD.

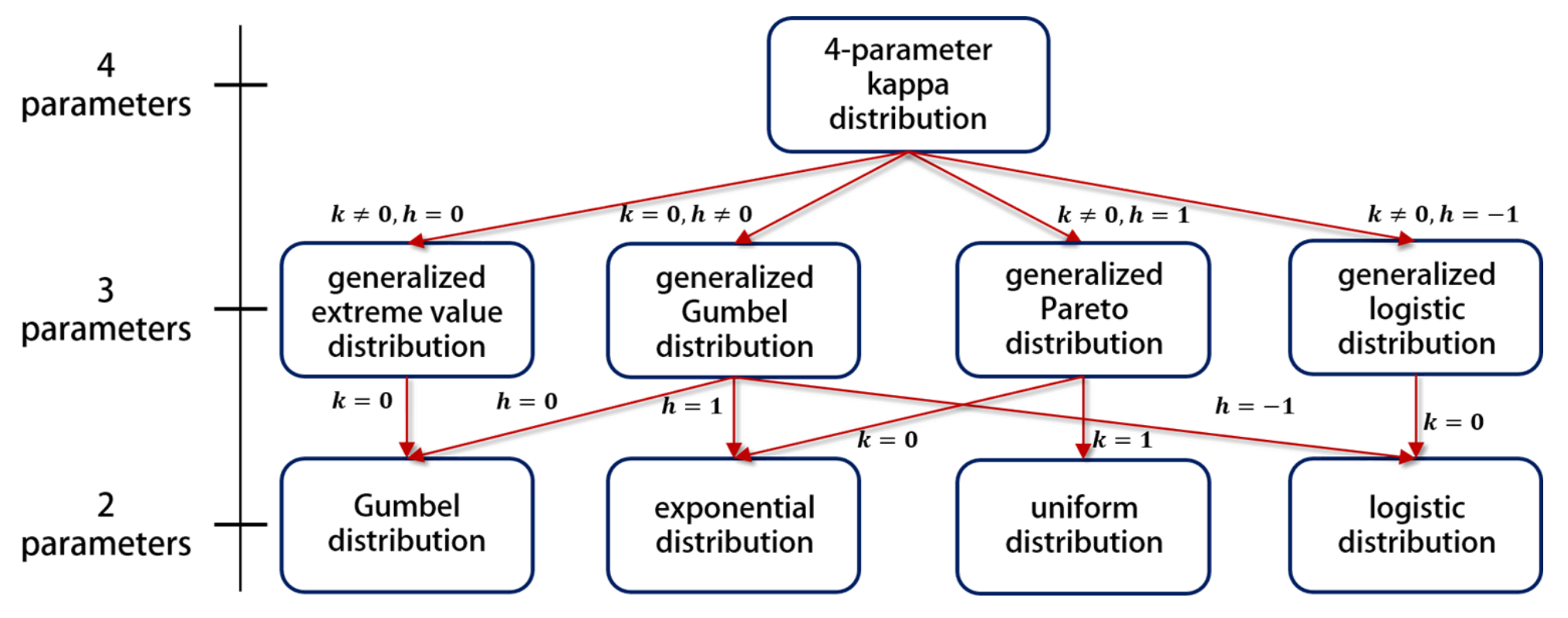

Figure1. Relationship of the four parameter kappa distribution (K4D) to order distributions, which indicates a wide coverage of K4D

1.4 Extension of the 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.

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. Figure 2 illustrates our motivic schema.

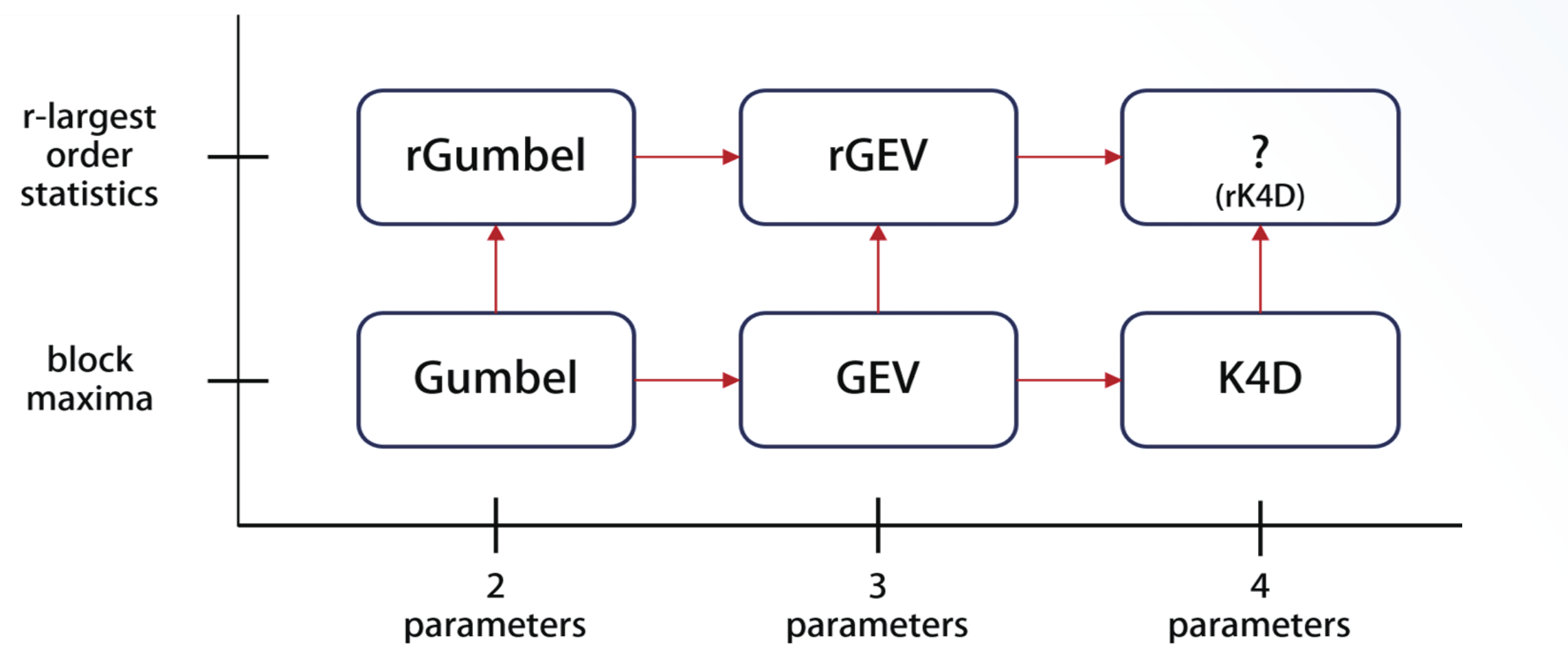

Figure2. A motivic schema on generalizations from 2 parameters to 4 parameters, and extensions from the block maxima models to the r-largest order statistic models, which leads to the r-largest four parameter kappa distribution (rK4D)

2. Definition r-largest four parameter kappa distribution

2.1 r-largest K4D

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 rGEVD. To define the joint probability density function 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 $k \neq 0, h \neq 0, \sigma > 0$

$$f_4\left(\underline{x}^{(r)}\right) = \rho^{-r}C_r \times g(\underline{x}^r) \times F_4(x^{(r)})^{1-rh} \qquad (5)$$

$$where \quad C_r = \begin{cases}\prod_{i=1}^{r-1}[1-(r-i)h] & if\ r \geq 2\\ 1 & if\ r\ = 1\end{cases} \quad (6) \qquad g\left(\underline{x}^r\right) = \prod_{s=1}^{r}\omega(x^{(s)})^{\frac{1}{k}-1} \quad (7)$$

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 in (2).

2.2 Marginal pdf for the rK4D

The marginal pdf of s-th order statistic from the rK4D is derived to the following by consecutive integrals of $f_4(\underline{x}^s)$ with respect to $(x^{(1)}, \cdots, x^{(s-1)})$

$$f_4\left(x^{(s)}\right) = \int_{x^{(s)}}^{\infty}\cdots\int_{x^{(2)}}^{\infty} f_4(\underline{x}^s)\,dx^{(1)},\ldots,dx^{(s-1)}$$

$$where\ w(x) = 1-k\frac{x-\mu}{\sigma} \qquad = \frac{C_{s-1}}{(s-1)!} \times \omega(x^{(s)})^{\frac{s-1}{k}} \times F_4(t)^{1-(s-1)h} \qquad (8)$$

For $2 \leq s \leq r$, where $g(\underline{x})$ is defined as in (7), $w(x)$ is defined as in (3), and F_4 is the cdf of K4D as in (4). We can see, as $h \to 0$, that the above marginal pdf (8) goes to the corresponding marginal pdf of the rGEVD,

$$f_3\left(x^{(s)}\right) = \sigma^{-1}\frac{1}{(s-1)!}w(x^s)^{\frac{s}{k}-1} \times \exp[-w(x^{(s)})^{\frac{1}{k}}] \qquad (9)$$

2.3 Marginal cdf for the rK4D

The marginal cdf of s-th order statistic from the rK4D is obtained by integrating $f_4(x^{(s)})$ as follows

$$H_4\left(x^{(s)}\right) = \int_{-\infty}^{t} f_4\left(x^{(s)}\right)dx^{(s)}$$

$$= \frac{C_{s-1}}{(s-1)!} \times \omega(x^{(s)})^{\frac{s-1}{k}} \times F_4(t)^{1-(s-1)h} \qquad (10)$$

When $h \to 0$, this marginal cdf goes to the marginal cdf of the rGEVD in (11) as provided in Coles (2001, p.67). The quantiles from this marginal cdf is obtained by solving the equation $H_4\left(z_p\right) = 1-p$ numerically, because (10) is not analytically inverted. Nonetheless, this is straightforward using standard numerical techniques.

$$H_3\left(x^{(s)}\right) = \exp[-w(x^{(s)})^{\frac{1}{k}}]\sum_{i=1}^{s-1}\frac{w(x^{(s)})^{\frac{1}{k}}]^i}{i!} \qquad (11)$$

3. Estimation MLE, Quantiles, Delta method

3.1 MLE

The likelihood function of μ, σ, h, k is as follows, for $k \neq 0, h \neq 0$ under constraints.

$$L\left(\mu, \sigma, h, k|\underline{x}^r\right) = \prod_{i=1}^{m}[\sigma^{-r}C_rF_4(x_i^{(r)})^{1-rh} \times \prod_{j=1}^{r}\left(1-k\frac{(x_i^{(j)}-\mu)}{\sigma}\right)^{\frac{1}{k}-1} \qquad (12)$$

We implemented a numerical algorithm using the 'optim' package in R program by consulting the 'ismev' package (Coles 2001).

The standard errors of the maximum likelihood estimates are obtained approximately by the squared root of the diagonal terms of the inverse of the observed Fisher information matrix.

3.2 Quantiles of the block maxima

z_p is known as the return level associated with the return period 1/p, since the level z_p is expected to be exceeded on average once every 1/p years (Coles 2001). For example, a 20-year (50-year) return level is computed as the 95th (98th) quantile of the fitted K4D.

The quantiles of the K4D are

$$z_p = \mu + \frac{\sigma}{k}\left\{1-\left(\frac{1-(1-p)^h}{h}\right)^k\right\} \qquad (13)$$

$$where\ F_4\left(z_p\right) = 1-p$$

3.3 Delta method for variance estimation

The variance estimation of the 1/p years return level (z_p) can be calculated by the delta method (Coles 2001). We present the details of the procedure including the derivatives of z_p with respect to each parameter as follows;

$$Var(\hat{z}_p) \approx \nabla z_p^t V \nabla z_p \qquad (14)$$

where V is the covariance matrix of parameter estimates which is approximated by the inverse of the observed Fisher information matrix

$$\nabla z_p^t = [\frac{\partial z_p}{\partial \mu}, \frac{\partial z_p}{\partial \sigma}, \frac{\partial z_p}{\partial k}, \frac{\partial z_p}{\partial h}] \qquad (15)$$

$$\frac{\partial z_p}{\partial \mu} = 1, \quad \frac{\partial z_p}{\partial \sigma} = \frac{1-y_p^k}{k}, \quad \frac{\partial z_p}{\partial k} = -\frac{\sigma\ln(y_p) \times y_p^k}{k} - \frac{\sigma(1-y_p^k)}{k^2},$$

$$\frac{\partial z_p}{\partial h} = \frac{\sigma\{(1-p)^h hln(1-p) + 1-(1-p)^h\}}{h^2} \times y_p^{k-1}$$

$$where\ y_p = 1-\frac{(1-p)^h}{h}, \text{ evaluated at } (\hat{\mu}, \hat{\sigma}, \hat{k}, \hat{h})$$

4. Application Venice sea-level data

4.1 Venice sea-level data

These data consist of the 8 largest sea-levels in Venice over the 1931-1981, except for the year 1935 (Coles 2001). In the 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.

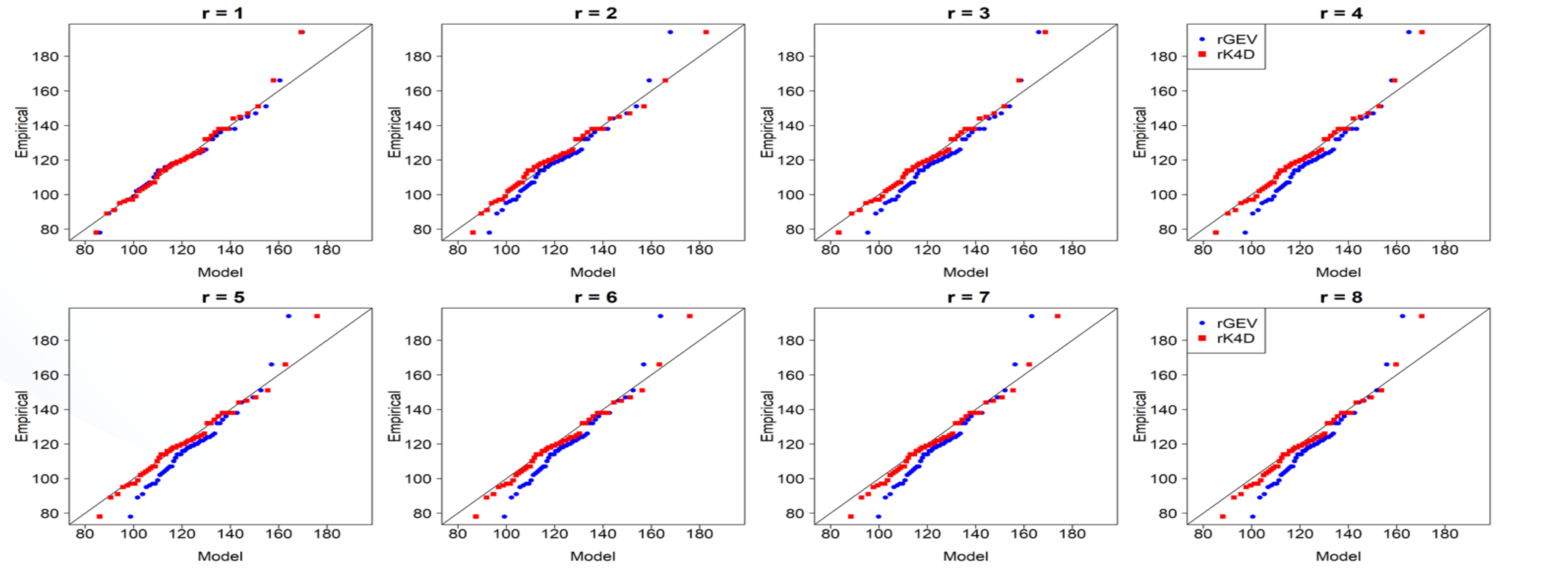

Figure3. Quantile-per-quantile plots obtained from the largest order statistics for the rK4D fit(red points) and for the rGEV fit(blue points) to Venice sea-level data with several values of r

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 h estimates in the rK4D are much larger compared to those of the other parameter estimates. 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 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)

$rGVED$					
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 (1.2)	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	118.6 (1.7)	14.3 (0.8)	-0.099(0.035)	155.0 (4.1)
5	732.0	118.8 (1.6)	13.7 (0.8)	-0.088(0.033)	154.3 (4.0)
6	829.6	119.1 (1.5)	13.4 (0.7)	-0.086(0.031)	154.0 (3.9)
7	916.5	119.8 (1.5)	13.2 (0.7)	-0.090(0.029)	153.6 (3.7)
8	995.7	120.5 (1.4)	13.1 (0.7)	-0.097(0.025)	153.3 (3.4)

$rK4D$						
r	$nllh$	$\hat{\mu}\ (se)$	$\hat{\sigma}\ (se)$	$\hat{k}\ (se)$	$\hat{h}\ (se)$	rk4d $r_{20}(se)$
1	221.8	120.0 (5.2)	9.0 (2.4)	-0.16 (0.057)	-1.67 (1.34)	153.6 (6.2)
2	372.6	116.9 (2.4)	10.2 (1.3)	-0.23 (0.064)	-1.31 (0.58)	159.5 (5.6)
3	499.8	118.0 (2.1)	10.4 (1.1)	-0.10 (0.051)	-1.03 (0.32)	153.8 (6.3)
4	610.6	117.2 (1.9)	10.9 (1.0)	-0.10 (0.048)	-0.83 (0.24)	154.8 (6.5)
5	705.4	116.9 (2.0)	11.5 (1.1)	-0.13 (0.050)	-0.77 (0.21)	157.9 (7.5)
6	803.8	117.0 (1.9)	12.0 (1.1)	-0.10 (0.052)	-0.61 (0.17)	158.4 (7.6)
7	889.4	116.9 (1.8)	12.2 (1.0)	-0.08 (0.048)	-0.49 (0.14)	157.5 (7.0)
8	961.9	117.1 (1.8)	11.9 (0.9)	-0.06 (0.042)	-0.49 (0.13)	154.5 (6.2)

$rGEVD$				
r	AIC	BIC	$tr(V)$	log $\vert V\vert$
1	451.4	457.2	10.2	-2.31
2	764.9	770.7	5.12	-4.55
3	1036.8	1042.6	4.16	-6.01
4	1270.5	1276.3	3.49	-6.98
5	1469.9	1475.7	3.06	-7.63
6	1665.3	1671.1	2.86	-7.63
7	1839.0	1844.8	2.67	-8.09
8	1997.4	2003.2	2.48	-8.60

$rk4d$			
AIC	BIC	$tr(V)$	log $\vert V\vert$
451.7	459.4	34.72	-3.56
753.2	761.0	7.92	-6.17
1007.5	1015.2	5.49	-7.73
1229.1	1236.9	4.84	-8.83
1418.7	1426.4	5.01	-9.31
1615.5	1623.2	4.85	-9.84
1786.7	1794.5	4.41	-10.68
1931.7	1939.4	4.05	-11.22

Table1. The estimate 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. **first table** is for the RGEVD and **second table** is for the Rk4d. 'nllh' stands for the negative log-likelihood function value. 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 (**third table)** and the Rk4d **(fourth table)**) fitted to Venice sea-level data with different value of r.

In table2, 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.

5. Discussion Reduce variance, Selection of r

5.1 Reduce variance

The standard error (SE) of the MLE of the rK4D parameters decrease in general as r increase, but the SE of the return level does not show a monotonic reduction trend. Moreover, the variances of the return levels in the rK4D are greater than those in the rGEVD. Some techniques to reduce the variance of the return level of the rK4D are anticipated in the future work. Since the rK4D generally will result in less bias than the rGEVD, one can consider the mean squared error (MSE) criterion for selecting better model between the rGEVD and the rK4D.

5.2 Selection of r

The use of the r-largest values as extremes enhances the power of estimation for moderate values of r, but the use of larger values of r may lead to bias in the estimation(Zhang et al. 2004). The selection of r is thus important in the rGEVD or in the rK4D.

Coles S (2001) An introduction to statistical modeling of extreme values. Springer, London, pp224.

Hosking JRM (1994) The four-parameter kappa distribution. IBM Journal of Research and Development 38:251-258.

Hosking JRM, Wallis JR (1997) Regional Frequency Analysis: An Approach Based on L-Moments.342 Cambridge University Press, Cambridge. pp 244.

James G, Witten D, Hastie T, Tibshirani R (2013) An Introduction to Statistical Learning with Applications in R. Springer, pp 426.

Zhang XB, Zwiers FW, Li GL (2004) Monte Carlo experiments on the detection of trends in extreme values. Journal of Climate 17:1945-1952.

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) (No. 2020R1I1A3069260, No. 2021R1A6A3A13044162).