Vol. 34: 185–194, 2007
doi: 10.3354/cr00701

CLIMATE RESEARCH
Clim Res

Published September 18

Estimation of a large quantile of the distribution of multi-day seasonal maximum rainfall: the value of stochastic simulation of long-duration sequences

T. A. Buishand*

Royal Netherlands Meteorological Institute (KNMI), PO Box 201, 3730 AE De Bilt, The Netherlands

ABSTRACT: The usefulness of time-series simulation of daily rainfall for estimating large quantiles of the distribution of 10 d seasonal maximum rainfall is questioned. The emphasis is on rare 10 d events having a mean recurrence time much longer than the length of the historical record. Time-series simulation uses nonparametric resampling. With a theoretical example assuming no temporal dependence, it is shown that simulation of a long-duration sequence by resampling from the original historical record using the standard bootstrap method yields a much better estimate of a large quantile of the 10 d seasonal maximum distribution than fitting a Gumbel or Generalized Extreme Value (GEV) distribution to the 10 d seasonal maxima from the original data. This is because a sample of daily values contains more information on the distribution of the 10 d maxima than the individual 10 d seasonal maxima from the sample. Using observed daily rainfall data from Stuttgart (Germany), it is demonstrated that the tail of the distribution of the daily rainfall amounts and temporal dependence strongly influence the distribution of extreme 10 d rainfalls. The incorporation of temporal dependence into the simulated data using nearest-neighbour resampling is considered. Using a first- and second-order resampling model, it is demonstrated that misspecification of the order of dependence may lead to a substantial bias in the quantiles of the 10 d seasonal maximum distribution. It is shown that the underestimation of large quantiles of the distribution of the 10 d maxima, resulting from the inability to generate larger daily values than those observed, is small. Despite these biases, it is expected that, in the case of temporal dependence of daily rainfall data, nearest-neighbour resampling is able to provide reliable estimates of large quantiles of the distribution of multi-day rainfall amounts, as in the example of no dependence.

KEY WORDS: Rainfall simulation · Nearest-neighbour resampling · Exponential distribution · Gamma distribution · Extreme-value distributions

Resale or republication not permitted without written consent of the publisher

1. INTRODUCTION

There is currently a large literature on the stochastic simulation of daily rainfall sequences. This Monte Carlo approach is sometimes viewed as a 'magic tool' for producing much-wanted data. For instance, a time series of sufficient length can be generated if the historical rainfall record is too short for the desired application. However, a simulation model does not contain more information than is available in the existing data. Moreover, the generated rainfall time series may not have the same statistical properties as the observed rainfall data. This brings into question the benefit of generating long synthetic sequences. The author was faced with this problem after the presentation of preliminary results from a rainfall generator for the Rhine river basin.

Despite the large number of time-series simulation studies in the climatological and hydrological literature, the question of whether or not a stochastic simulation can result in better estimates of a design variable than a statistical analyses of the historical data alone

*Email: adri.buishand@knmi.nl

© Inter-Research 2007 · www.int-res.com

has seldom been addressed. One study (Vogel & Stedinger 1988) examined the value of stochastic streamflow models for the design of storage reservoirs. They concluded that a simple log-normal first-order autoregressive model generally lead to improved estimates of the design capacity in terms of root mean square error than when only the historical flow data were considered, even in situations of model misspecification.

In the present study, the estimation of a large quantile of the distribution of 10 d seasonal maximum rainfall is considered. The motivation for choosing 10 d rainfall is that prolonged heavy rainfall during the winter is the major cause of extreme flows of the Rhine river in the Netherlands. Non-parametric resampling techniques are used to generate synthetic sequences. Section 2 considers a simple daily rainfall model for which the distribution of the 10 d seasonal rainfalls is known. Quantile estimates from resampled sequences using the bootstrap method are compared with those based on extrapolation of distributions fitted to the 10 d seasonal maximum rainfall amounts from the original data. The sensitivity of the 10 d seasonal maximum distribution to model assumptions is discussed in Section 3. Nearest-neighbour resampling is introduced in Section 4, as an extension of the bootstrap method for autocorrelated data, to demonstrate some aspects of deficiencies in the model. Section 5 closes the paper with a discussion and conclusions.

2. EXTREME 10 d RAINFALL IN A STOCHASTIC MODEL

Here, we assume that the probability of rain p_{wet} is not dependent on rainfall occurrence on previous days and on the time of the year. For the wet days, it is assumed that the rainfall amounts are mutually independent and identically distributed. The positively skewed distribution of these wet-day rainfall amounts is taken to be exponential with scale parameter α:

$$\Pr(X \le x) = 1 - \mathrm{e}^{-x/\alpha}, \quad x \ge 0 \tag{1}$$

Let S_D be the total rainfall amount over some period of D days, and N_D the number of wet days in that period. The distribution of S_D follows through conditioning on N_D:

$$\begin{aligned} F_D(x) &= \Pr(S_D \le x) \\ &= \sum_{n=0}^{D} \Pr(S_D \le x \mid N_D = n) \times \Pr(N_D = n), \quad x \ge 0 \end{aligned} \tag{2}$$

Since the sum of n independent values from the exponential distribution (Eq. 1) has a gamma distribution with shape parameter n and scale parameter α, the conditional distribution of S_D on the right-hand side of Eq. (2) is given by

$$\Pr(S_D \le x \mid N_D = n) = 1 - \mathrm{e}^{-x/\alpha} \sum_{i=0}^{n-1} \frac{(x/\alpha)^i}{i!}, \quad n \ge 1 \tag{3}$$

and $\Pr(S_D \le x \mid N_D = n) = 1$ for $n = 0$. This equation is the relation that links the distribution of the gamma variable with an integral shape parameter to the Poisson distribution (e.g. Abramowitz & Stegun 1964 their Section 26.4, Pearson & Hartley 1976 their Section 3.3). The number N_D of wet days follows a binomial distribution:

$$p_n = \Pr(N_D = n) = \binom{D}{n} p_{\text{wet}}^n (1 - p_{\text{wet}})^{D-n}, \quad n = 0, \ldots, D \tag{4}$$

Eq. (2) can now be rewritten as:

$$F_D(x) = 1 - \mathrm{e}^{-x/\alpha} \sum_{n=1}^{D} p_n \sum_{i=0}^{n-1} \frac{(x/\alpha)^i}{i!}, \quad x \ge 0 \tag{5}$$

Assume that there are I non-overlapping periods of D days in a season. Let the rainfall amounts in these periods be $S_D(1),\ldots,S_D(I)$, respectively. Then, for the distribution of the maximum $S_{D,\max} = \max[S_D(1),\ldots,S_D(I)]$ we can write

$$F_{D,\max}(x) = [F_D(x)]^I \tag{6}$$

because $S_{D,\max}$ can only be $\le x$ if all $S_D(i) \le x$.

The distribution of $S_{D,\max}$ can easily be calculated with Eqs. (4), (5) & (6). We now consider the estimation of a large quantile of this distribution through resampling from a short data record from the underlying stochastic rainfall model. We take $p_{\text{wet}} = 0.5$, $D = 10$ and $I = 20$. The season of interest has thus a length of 200 d (e.g. a winter or summer half-year). Because the performance criteria used are insensitive to the scale parameter α of the exponential distribution, we set α = 1.

A large number of sequences of $J = 20$ yr were generated. For each sequence, the seasonal maxima $S_{10,\max}$ were determined. The $S_{10,\max}$ for the first 3 simulation runs are presented as a Gumbel plot in Fig. 1. For short and moderate return periods, there is little difference between the 3 simulations. The ordered maxima almost fall on the line representing the true distribution from Eq. (6). The largest values in the 3 simulations, however, greatly differ. This shows that it is not possible to get reliable estimates of large quantiles of the distribution of $S_{10,\max}$ from the observed 10 d maxima in a short record of $J = 20$ yr, e.g. the 10 d rainfall amount that is exceeded on average once in $T = 100$ yr. The question is whether or not time-series simulation can provide better estimates of such quantiles. This is explored here with the bootstrap method.

A 20 yr simulation contains 20 × 200 = 4000 values. For each simulation, a new series of $J^* = 499$ yr was generated by random sampling with replacement from these 4000 values. Thus, no new daily values were generated, and a specific daily value was generally sampled several times. The largest 10 d rainfalls in the

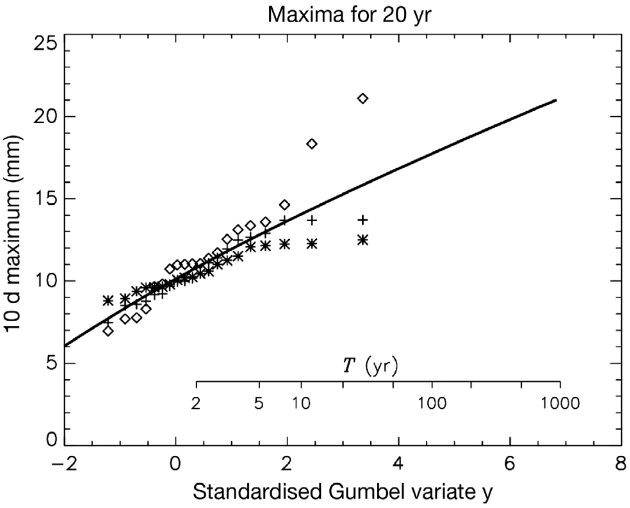

Fig. 1. Gumbel plots of the 10 d seasonal maximum rainfall amounts in three 20 yr simulations (◇, ✳, +) with a simple stochastic rainfall model. The solid curve represents the true 10 d maximum distribution for the underlying rainfall model

Table 1. Relative bias, SE and RMSE of the estimated 10 d rainfall with 100 yr return period for sequences of $J = 20$ yr from a simple stochastic rainfall model and for resampled sequences of $J^* = 499$ and $J^* = 1999$ yr using the data in the 20 yr sequences. Bias, SE and RMSE: percentage relative to the true 100 yr event. MAX_r^*: the rth largest seasonal maximum of the resampled data. Parentheses: the asymptotic SE of the rth largest seasonal maximum in a model simulation of $J = J^*$ years, as given by Eq. (7)

Method	Bias	SE	RMSE
$J = 20$			
Gumbel	3.4	11.0	11.6
GEV	1.6	16.3	16.3
$J^* = 499$			
Gumbel	3.0	3.7	4.7
GEV	−0.6	4.4	4.4
MAX_r^*	0.5	5.0	5.1
		(3.8)	
$J^* = 1999$			
Gumbel	3.1	3.4	4.5
GEV	−0.4	3.9	3.9
MAX_r^*	0.0	4.0	4.0
		(1.9)	

new 499 yr series differed, however, from those in the original 20 yr simulation because the temporal sequence of the daily values had changed. The ordered $S_{10,\max}$ values for the first 3 simulations are shown in Fig. 2. In contrast to Fig. 1, each 499 yr simulation describes the upper tail of the distribution quite well, showing that a 20 yr record of daily rainfall amounts contains much more information about the distribution of $S_{10,\max}$ than the 20 individual $S_{10,\max}$ alone.

For the quantile $S_{D,\max}(100)$ that is exceeded on average once in 100 yr (the '100 yr event'), Table 1 compares the estimates based on extrapolation of distributions fitted to the 10 d seasonal maximum rainfall amounts in the original record of $J = 20$ yr with those obtained from resampled sequences of $J^* = 499$ and $J^* = 1999$ yr. Both the Gumbel and Generalized Extreme Value (GEV) distributions were fitted to the $S_{10,\max}$ distributions using probability weighted moments (Landwehr et al. 1979, Hosking et al. 1985). For the resampled data, the 100 yr event was also estimated as the rth largest seasonal maximum MAX_r^*, with $r = (J^* + 1)/100$. The bias, standard error (SE) and root mean square error (RMSE) of the various quantile estimates were derived from 399 simulated 20 yr sequences using the stochastic daily rainfall model. In Table 1, these performance measures are divided by the true value (17.75) of $S_{D,\max}(100)$.

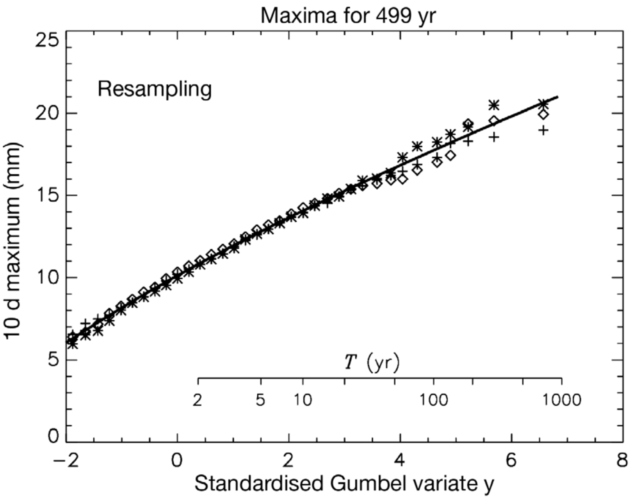

Fig. 2. As for Fig. 1, but resampled sequences of 499 yr using the data of the 20 yr simulations. For ease of readability not all ordered 10 d maxima have been plotted

The Gumbel distribution overestimates $S_{D,\max}(100)$ by about 3% in all cases. This departure from the true distribution can be seen in Figs. 1 & 2 where the true distribution appears as a slightly concave curve rather than a straight line. The more flexible GEV distribution is able to describe such a concave plot. This higher flexibility is, however, paid for by an increase in the SE of the 100 yr event, especially for the original 20 yr series. As a result, for small samples, the estimate from the GEV distribution is not better than that from the Gumbel distribution in terms of RMSE. There is also some bias in the GEV estimate for $J = 20$ which is, unlike the bias in the Gumbel estimate, mainly due to the small sample size. This bias of the GEV estimate almost disappears in the longer resampled series. More important than the differences in bias are the smaller SEs of the estimates from the resampled sequences resulting from the better use of the available information about extreme 10 d rainfall events in the data. The RMSEs of these estimates are, therefore, less than half those of the best estimate from the seasonal maximum rainfalls in the original 20 yr series.

From the theory of order statistics, a simple approximation can be given to the variance of the rth largest value MAX_r in a sequence of J years (David 1981, his Section 4.6):

$$\text{var}(MAX_r) \approx \frac{1}{J+2}\,\frac{1/T(1-1/T)}{\{f_{D,\max}[S_{D,\max}(T)]\}^2} \tag{7}$$

where T is the return period, $r = (J + 1)/T$ and $f_{D,\max}(.)$ is the probability density of $S_{D,\max}$

$$f_{D,\max}(x) = I f_D(x)\{F_D(x)\}^{I-1} \tag{8}$$

where $F_D(x)$ is given by Eq. (5) and

$$f_D(x) = F'_D(x) = \mathrm{e}^{-x/\alpha}\sum_{n=1}^{D} p_n \frac{\alpha^{-n}x^{n-1}}{(n-1)!},\;\; x > 0 \tag{9}$$

The SE based on Eq. (7) is given in parentheses in Table 1. It is smaller than the corresponding SE for the resampled data. The extra variation of MAX_r^* is due to the limited length of the records from which the data are resampled. For a finite record length of J years, var(MAX_r^*) tends to a non-zero limit with growing J^*. The variance can then only be further reduced by resampling from records longer than J years. This also applies to other estimates of $S_{D,\max}(T)$. For the estimate of the 100 yr event in Table 1, the decrease in SE and RMSE is small if J^* is increased from 499 to 1999, in particular for the estimates from the Gumbel and GEV distributions.

3. SENSITIVITY OF THE 10 d MAXIMA TO MODEL ASSUMPTIONS

Some extensions to the simple stochastic rainfall model in the previous section are possible that allow for easy calculation of the $S_{10,\max}$ distribution. First, the exponential distribution for the wet-day rainfall amounts can be replaced by the more general gamma distribution. The conditional distribution of S_D given the number N_D of wet days then becomes:

$$\Pr(S_D \le x \mid N_D = n) = \frac{1}{\Gamma(n\nu)}\int_0^{x/\alpha} t^{n\nu-1}\mathrm{e}^{-t}\,\mathrm{d}t \tag{10}$$

where $\Gamma(.)$ is the gamma function, α is the scale parameter and ν the shape parameter of the gamma distribution for the wet-day rainfall amounts. For $\nu = 1$ (exponential distribution), Eq. (10) reduces to Eq. (3).

A second extension is to vary the wet-day probability p_{wet} between the D-day periods according to a beta distribution. This leads to the beta-binomial distribution for the number N_D of wet days (e.g. Prentice 1986):

$$\Pr(N_D = n) = \tag{11}$$
$$\binom{D}{n}\prod_{i=0}^{n-1}(\pi_{\text{wet}}+\gamma i)\prod_{i=0}^{D-n-1}(1-\pi_{\text{wet}}+\gamma i)\Big/\prod_{i=0}^{D-1}(1+\gamma i),\quad n = 1,\ldots,D-1$$

with π_{wet} the mean of p_{wet} (the unconditional probability of a day being wet). The probabilities that $N_D = 0$ and $N_D = D$ follow from Eq. (11) by adopting the convention that $\prod_{i=0}^{-1}(\pi_{\text{wet}}+\gamma i) = \prod_{i=0}^{-1}(1-\pi_{\text{wet}}+\gamma i) = 1$. The beta-binomial distribution is an example of a compound distribution, i.e. a distribution that arises by assigning a distribution to one or more parameters of a particular family of distributions. The randomness of p_{wet} may be interpreted as a persistence effect, wet days tend to be clustered in D-day periods with relatively high values of p_{wet}. The degree of clustering increases with increasing γ. For $\gamma = 0$, the beta-binomial distribution reduces to the binomial distribution, given by Eq. (4). The use of the beta distribution is similar to that in Hansen & Mavromatis (2001) and Wan et al. (2005) who perturbed the probability of rain in each month in their stochastic models to get an increase in the interannual variability of the frequency of wet days.

A third extension is to introduce persistence in the wet-day rainfall amounts as well by varying the scale parameter α of the gamma distribution in Eq. (10) between the D-day periods. The most tractable way to do so is to assume that $1/\alpha$ has a gamma distribution with shape parameter τ and scale parameter β (e.g. Johnson et al. 1994 Chapter 17, Koutsoyiannis 2004). The conditional distribution of S_D then becomes a beta distribution of the second kind:

$$\Pr(S_D \le x \mid N_D = n) = \frac{1}{\mathrm{B}(n\nu,\tau)}\int_0^{\beta x/(\beta x+1)} t^{n\nu-1}(1-t)^{\tau}\,\mathrm{d}t \tag{12}$$

with B(.,.) the beta function.

The different models are now compared using the 10 d maximum rainfall amounts at Stuttgart (Germany) in the 180 d periods 2 October to 30 March (29 March in leap years) of the 34 winters 1961/62,…,1994/95. In order to reduce the effect of the annual cycle, the 180 d period was divided into three 60 d seasons. For each season, the wet-day probabilities p_{wet} in Eq. (4) and π_{wet} in Eq. (11) were set equal to the observed fraction of wet days (days with 0.1 mm rainfall or more). The parameters of the exponential and gamma distributions of the wet-day rainfall amounts were estimated by the method of moments. The method of moments was preferred here to avoid bias in the standard deviation (SD) of the 10 d rainfall amounts due to parameter estimation. Such a bias may occur if the gamma distribution is fitted by the method of maximum likelihood because of underestimation of the SD of the wet-day rainfall amounts (Wilks 1999). An estimate of the parameter γ of the beta-binomial distribution was obtained for each season by equating the theoretical variance of N_{10} to the corresponding sample variance. Fig. 3 compares frequency distributions of N_{10} from the fitted distributions with the observed distribution of N_{10}. It can be seen that the binomial distribution is too concentrated about its mean; it underestimates the frequency of both 10 d periods with a large number of wet days and those with a large number of dry days. The beta-binomial distribution fits quite well. The estimation of the parameters ν, β and τ of the compound gamma distribution for wet-day rainfall is given in Appendix 1.

Eqs. (2) & (6) were applied to obtain the distribution $F_{10,\max,i}(x)$ of the maximum 10 d rainfalls for each of the three 60 d seasons ($i = 1, 2, 3$), from which the distribution of the maximum for the whole 180 d winter period follows as

$$F_{10,\max}(x) = \prod_{i=1}^{3} F_{10,\max,i}(x) \qquad (13)$$

Fig. 4 shows that the simple rainfall model in Section 2 with no persistence in daily rainfall occurrence and exponentially distributed rainfall amounts grossly underestimates the quantiles of the 10 d winter maximum distribution. The underestimation is roughly halved if the gamma distribution is fitted to the wet-day rainfall amounts. The latter fits the upper tail of the distribution of the wet-day rainfall amounts much better than the exponential distribution. The underestimation (20 to 25%) of the quantiles of the distribution of the 10 d winter maxima is mainly due to the absence of persistence in the daily rainfall amounts. Bootstrapping the historical daily values results in an underestimation of similar magnitude (Brandsma & Buishand 1999). Due to the absence of persistence, the SD of the 10 d rainfall amounts S_{10} is underestimated by about 25%, resulting in too few 10 d periods with extreme rainfall. The use of the beta-binomial model for rainfall occurrence reduces the underestimation of the quantiles of the distribution of the 10 d winter maxima, but these quantiles are still not satisfactorily reproduced. This is in line with results elsewhere that incorporation of persistence in the occurrence of wet and dry days is generally not sufficient to preserve the standard deviation of monthly rainfall totals (e.g. Katz & Parlange 1998). The beta-binomial model with independent wet-day rainfall amounts from the gamma distribution underestimates the SD of S_{10} by 13%. Introducing randomness of the scale parameter of the gamma distribution leads to a further increase of the quantiles of the 10 d winter maximum distribution. There is still some underestimation of the SD of S_{10} (about 7%) with this extension. The reproduction of the 10 d winter maximum distribution looks, however, reasonable.

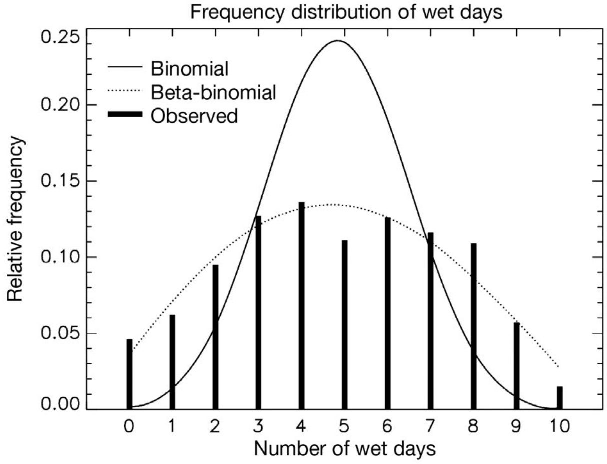

Fig. 3. Frequency distribution of the number of wet days in a 10 d period at Stuttgart for the winter half-years in the period 1961–1995

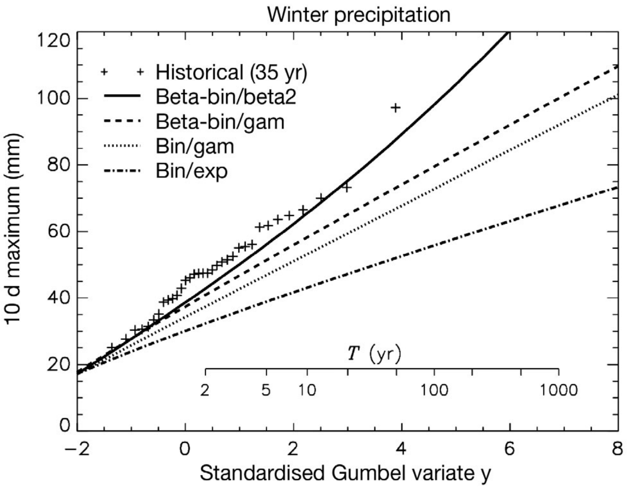

Fig. 4. Gumbel plot of the 10 d winter maximum rainfalls at Stuttgart (1961–1995) compared with the 10 d maximum distributions for 4 stochastic rainfall models: number of wet days binomial with exponentially (Bin/exp) or gamma (Bin/gam) distributed wet-day rainfall amounts; and number of wet days beta-binomial with a gamma (Beta-bin/gam) or a compound gamma (Beta-bin/beta2) distribution for the wet-day rainfall amounts

4. NEAREST-NEIGHBOUR RESAMPLING

From Section 3 it is clear that we need an extension of the standard bootstrap method to build persistence into the simulated daily rainfall sequences. Nearest-neighbour resampling is such an extension. Lall & Sharma (1996) discussed the connection between nearest-neighbour resampling and Markov chain modelling and explored the utility of the method for resampling hydrologic time series. Young (1994) and Rajagopalan & Lall (1999) used nearest-neighbour resampling for generating daily weather data. Different resampling models can be defined which may lead to different extreme-value distributions of the generated multi-day rainfall amounts. In Section 4.1, a first-order resampling model is compared with a second-order model.

The fact that the highest generated value cannot be larger than the highest historical daily value may limit the use of re-sampling techniques. This limitation is explored in Section 4.2 using the 10 d winter maximum rainfalls at Stuttgart.

4.1. First- and second-order resampling models

In the first-order model, resampling is conditioned on the last generated value x_t^*. This is achieved by finding the k nearest neighbours of x_t^* in the historical record in terms of Euclidean distance:

$$\delta_s = |x_s - x_t^*| \qquad (14)$$

where x_s is the sth historical value. For the Stuttgart data, the search for nearest neighbours was restricted to the historical values for the season of interest. One of the k nearest neighbours was selected at random and the historical value subsequent to that nearest neighbour was adapted as the simulated value for the next day t + 1. As in Lall & Sharma (1996), the jth closest neighbour was selected with probability

$$p_j = \frac{1/j}{\sum_{i=1}^{k}(1/i)}, \qquad j = 1,\ldots,k \qquad (15)$$

If there were several potential j closest neighbours because of an equal distance from x_t^*, then one of these days was sampled at random. This occurred e.g. if $x_t^* = 0$, in which case one of the N_{dry} dry days in the 60 d season of interest was selected with probability $1/N_{\text{dry}}$. For each 180 d winter period, the value for the first day (2 October) was randomly sampled from the historical values in the first 60 d season.

A second-order resampling model is obtained by including the value simulated for day $t - 1$ in the search for nearest neighbours, using the metric:

$$\delta_s = \sqrt{w_1(x_s - x_t^*)^2 + w_2(x_{s-1} - x_{t-1}^*)^2} \qquad (16)$$

For simplicity, the weights w_1 and w_2 were taken to be 1. The optimization of weights is discussed by Mehrotra & Sharma (2006). Their algorithm considers the minimization of the error of predicting a new value rather than statistics of extremes. The simulation was started by sampling 2 successive historical values in the first 60 d season.

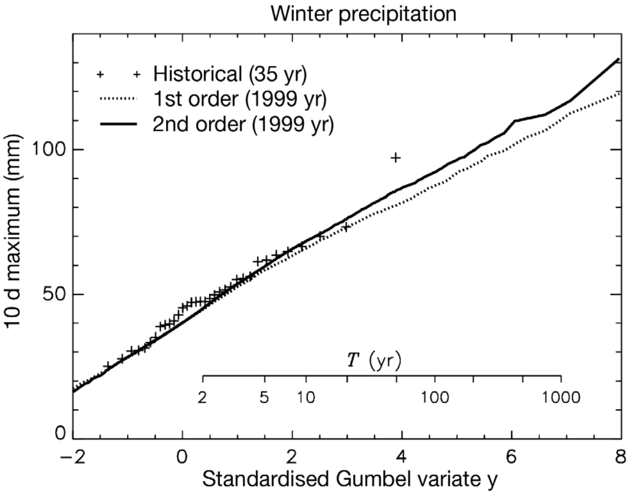

Fig. 5. Gumbel plots of the 10 d winter maximum rainfalls in the historical record of Stuttgart (1961–1995) and of the 10 d maxima in resampled sequences. Plots for the resampled data are based on 5 simulation runs of 1999 yr with a first-order or second-order nearest-neighbour model

The models defined above were used to generate 1999 yr sequences for Stuttgart. Fig. 5 compares the distribution of the 10 d winter maxima of the simulated data with that of the observed 10 d winter maxima. The parameter k was taken to be 5 in these simulations. The values for the simulated data in Fig. 5 refer to the average ordered 10 d winter maximum amounts in 5 runs of 1999 yr each. Both the first- and second-order model reproduce the observed winter maximum distribution well, in particular the second-order model. For long return periods, the quantiles from the simulations with the second-order model exceed those from the first-order model. This was also found in simulations with $k = 10$ and $k = 50$.

The first- and second-order models underestimate the SD of the 10 d rainfall amounts S_{10} by 5 and 3%, respectively. In practice, it is often difficult to decide which is the correct model. Misspecification of the order of dependence may lead to a substantial bias in the large quantiles of the 10 d winter maximum distribution. For the simulations in Fig. 5, the quantiles from the second-order model differ about 5% from those from the first-order model if $T > 20$ yr.

4.2. Significance of larger daily amounts than observed

In Section 2, long daily rainfall sequences were generated by resampling from a relatively short series of 20 yr. Despite the fact that the highest simulated value could not be larger than the highest daily value in the 20 yr series, a large quantile of the 10 d seasonal maximum distribution was adequately reproduced. The extreme 10 d rainfalls S_{10} in that example mainly refer to 10 d periods with moderately large rainfall over a number of days rather than to 10 d periods with very large rainfall on a single day. A daily value exceeding $S_{10,\max}(100)$ occurs, for instance, only once in 500 000 yr. The situation can, however, be quite different for heavy-tailed distributions. Several classes of heavy-tailed distributions can be defined (Embrechts et al. 1997, their Sections 1.3 and 1.4). One is the class of subexponential distributions. Let S_D be the sum of D realizations from a certain distribution $F(x)$, and M_D the largest value of these realizations. Then $F(x)$ is subexponential if

$$\lim_{x\to\infty} \frac{\Pr(S_D > x)}{\Pr(M_D > x)} = 1 \qquad (17)$$

for all $D \geq 2$. For this class of distributions, large values of S_D are caused by a single extreme realization. It is therefore interesting to study what would happen with the 10 d winter maximum distribution if we replace the largest daily rainfall amounts in the 1999 yr simulations for Stuttgart by values drawn from the tail of a subexponential distribution. Here, the generalized Pareto distribution (GPD) is used for this purpose.

The generalized Pareto variable has distribution function:

$$G(x) = \Pr(X \leq x) = 1 - [1 - \theta(x-u)/\alpha]^{1/\theta}, \qquad x \geq u \qquad (18)$$

The parameter θ determines the shape of the distribution. The distribution is subexponential if $\theta < 0$. For $\theta = 0$, the distribution should be interpreted as the limit of Eq. (18) as $\theta \to 0$, giving

$$G(x) = \mathrm{e}^{-(x-u)/\alpha}, \qquad x \geq u \qquad (19)$$

This distribution arises if we truncate the distribution in Eq. (1) at the threshold u and is also known as the exponential distribution. If $\theta > 0$, the GPD has an upper bound of $u + \alpha/\theta$.

Unfortunately, it is not possible to get an accurate estimate of the parameter θ from a series of 35 yr. Quite often θ can be assumed to be constant over a certain geographical area and a reasonable estimate of θ can then be derived from a large number of records for that area. For daily maximum rainfalls in the winter half-year (October to March), such regional analyses have been performed by Buishand (1983) for the Netherlands, Brandsma & Buishand (1999) for the German part of the Rhine basin, and Gellens (2002) for Belgium. These analyses suggest that $\theta < 0$. Here, we consider the GPD with the shape parameter taking values $\theta = -0.2$, -0.1 and 0. The parameter α was estimated as $(1 + \theta)(\bar{x}_u - u)$ with $\bar{x}_u$ the mean of the daily rainfall amounts exceeding u. The threshold u was set equal to 20 mm (17 exceedances), 17 mm (14 exceedances) and 14 mm (16 exceedances) for the first, second and third 60 d seasons, respectively.

Table 2. Estimated quantiles of the 10 d winter maximum distribution from the original simulations for Stuttgart with the second-order resampling model (Fig. 5), and from simulations perturbed with random values from the generalized Pareto distribution with shape parameter θ. The quantile estimates are averages of MAX_r^* over 5 simulation runs of 1999 yr

Return period	r	Estimated quantile (mm)			
T (yr)		Original	θ = –0.2	θ = –0.1	θ = 0
10	200	68.4	68.4	67.9	67.4
25	80	78.4	79.7	78.6	77.7
100	20	92.3	95.7	92.8	90.3
250	8	102.3	107.1	99.7	98.1

Whenever the largest or second largest historical daily rainfall amount of a 60 d season occurred in the 1999 yr simulated sequence, it was replaced by a value from the GPD. This value was conditioned to exceed the third largest historical value. Details are given in Appendix 2. For the second-order resampling model, Table 2 compares some estimated quantiles (MAX_r^*) of the 10 d winter maximum distribution from the original simulations with those from the perturbed simulations. The values of MAX_r^* in the table are averages over 5 simulation runs of 1999 yr. The effect of the perturbation is small for return periods shorter than the length of the historical series. For longer return periods, the quantiles of the simulations that were perturbed with values from the tail of the exponential distribution are lower than the corresponding quantiles of the original simulations. The opposite is found if the perturbations are based on the GPD with $\theta = -0.2$. The relative differences are about 5% for $T = 100$ and $T = 250$ yr. Similar differences were obtained by replacing the 4 largest historical daily rainfall amounts of each 60 d season in the simulated sequences by random numbers from the tail of the GPD. The earlier mentioned regional analyses of daily maximum rainfalls suggest that the upper tail of the distribution of the daily rainfall amounts is longer than that of the exponential distribution, but not as long as that of the GPD with $\theta = -0.2$. It can therefore be assumed that the relative differences between the perturbed simulations in Table 2 are larger than the potential bias caused by the inability of the resampling procedure to generate larger values than the highest historical daily rainfall.

5. DISCUSSION AND CONCLUSIONS

Resampling of daily rainfall data was used to estimate a quantile of the distribution of the 10 d seasonal maximum rainfall amounts. The emphasis was on quantiles with a mean recurrence time exceeding the length of the historical record.

For synthetic data generated with a simple stochastic rainfall model assuming no temporal dependence in the probability and amount of rain, it was demonstrated that resampling by the standard bootstrap method could provide a much better estimate of a large quantile of the 10 d seasonal maximum distribution than the classical method of fitting a Gumbel or GEV distribution to the 10 d seasonal maximum rainfall amounts. The longer the record from which the data are resampled, the more accurate the quantile estimate will be that is obtained with the bootstrap method.

For daily rainfall data from Stuttgart it was shown that both the persistence in the occurrence of rain and the wet-day rainfall amounts have to be taken into account if the distribution of extreme 10 d rainfalls has to be preserved, which is not the case for the standard bootstrap method. Nearest-neighbour resampling was employed to reproduce the temporal dependence structure of these data. Resampling with a second-order model then resulted in somewhat higher quantiles of the 10 d winter maximum distribution than resampling with a first-order model. Particular attention was given to the potential bias in large quantiles of the distribution due to the inability of a resampling procedure to generate larger daily rainfall amounts than the highest observed daily value. Replacing the largest simulated daily rainfall amounts by random values from the tail of the GPD suggests that this bias is small.

This study was restricted to daily rainfall simulation at a single site. The joint simulation of several weather variables at a single site was discussed in early papers examining nearest-neighbour resampling (Young 1994, Rajagopalan & Lall 1999). Buishand & Brandsma (2001) and Leander et al. (2005) used the method for multi-site simulation of daily rainfall and temperature. In contrast to the simulations presented in this paper, Sharma & Lall (1999) applied nearest-neighbour methods to generate first the wet and dry spells, and then the daily rainfall amounts for each wet spell. Harrold et al. (2003a) formulated a nearest-neighbour model for generating daily rainfall occurrence that was able to reproduce long-term variability. The wet-day rainfall amounts were generated from a kernel-based nonparametric estimate of the probability density (Harrold et al. 2003b). The advantage of the latter is that larger values than the highest observed daily rainfall amounts can be generated. It is uncertain, however, whether or not the method properly describes the extreme upper tail of the daily rainfall distribution.

To summarize: stochastic simulation offers the opportunity to obtain reliable estimates of large quantiles of the distribution of multi-day rainfall amounts. However, one should realise that the SE of a quantile estimate cannot be made arbitrarily small by generating very long records and one should be aware of the consequences of model misspecification.

Acknowledgements. This article is based on a lecture presented by the author to the 8th International Meeting on Statistical Climatology, Lüneburg, Germany, 12 to 16 March 2001. The author is most grateful to the Programme Committee of this meeting for inviting him to present a paper on extreme-value statistics. He further wishes to thank J. J. Beersma for discussions about the presentation of the material and for producing the figures. The daily precipitation data for Stuttgart were made available by the Deutscher Wetterdienst via the International Commission for the Hydrology of the Rhine basin (CHR/KHR). The work was performed in cooperation with the Institute for Inland Water Management and Waste Water Treatment (RIZA), Lelystad.

LITERATURE CITED

Abramowitz M, Stegun IA (1964) Handbook of mathematical functions. Dover, New York

Brandsma T, Buishand TA (1999) Rainfall generator for the Rhine basin: multi-site generation of weather variables by nearest-neighbour resampling. Publication 186-II. KNMI, De Bilt

Buishand TA (1983) Extremely high rainfall amounts and the theory of extreme values. Cultuurtechnisch Tijdschrift 23:9–20, corrigendum 23:81 (in Dutch)

Buishand TA (1989) Statistics of extremes in climatology. Statistica Neerlandica 43:1–30, corrigendum 43:i

Buishand TA, Brandsma T (2001) Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbor resampling. Water Resour Res 37:2761–2776

David HA (1981) Order statistics, 2nd edn. Wiley, New York

DuMouchel WH (1983) Estimating the stable index α in order to measure tail thickness: a critique. Ann Stat 11:1019–1031

Embrechts P, Klüppelberg C, Mikosch T (1997) Modelling extremal events for insurance and finance. Springer-Verlag, Berlin

Gellens D (2002) Combining regional approach and data extension procedure for assessing GEV distribution of extreme precipitation in Belgium. J Hydrol 268:113–126

Hansen JW, Mavromatis T (2001) Correcting low-frequency variability bias in stochastic weather generators. Agric For Meteorol 109:297–310

Harrold TI, Sharma A, Sheather SJ (2003a) A nonparametric model for stochastic generation of daily rainfall occurrence. Water Resour Res 39:1300, doi:10.1029/2003WR002182

Harrold TI, Sharma A, Sheather SJ (2003b) A nonparametric model for stochastic generation of daily rainfall amounts. Water Resour Res 39:1343, doi:10.1029/2003WR002570

Hosking JRM, Wallis JR, Wood EF (1985) Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. Technometrics 27:251–261

Johnson NL, Kotz S, Balakrishnan N (1994) Continuous univariate distributions, Vol 1, 2nd edn. Wiley, New York

Katz RW, Parlange MB (1998) Overdispersion phenomenon in stochastic modeling of precipitation. J Clim 11:591–601

Koutsoyiannis D (2004) Statistics of extremes and estimation of extreme rainfall. I. Theoretical investigation. Hydrol Sci J 49:575–590

Lall U, Sharma A (1996) A nearest neighbor bootstrap for resampling hydrologic time series. Water Resour Res 32: 679–693

Landwehr JM, Matalas NC, Wallis JR (1979) Probability weighted moments compared with some traditional techniques in estimating Gumbel parameters and quantiles. Water Resour Res 15:1055–1064

Leander R, Buishand A, Aalders P, de Wit M (2005) Estimation of extreme floods of the river Meuse using a stochastic weather generator and a rainfall-runoff model. Hydrol Sci J 50:1089–1103

Mehrotra R, Sharma A (2006) Conditional resampling of hydrologic time series using multiple predictor variables: a *K*-nearest neighbour approach. Adv Water Resour 29:987–999

Pearson ES, Hartley HO (1976) Biometrika tables for statisticians, Vol 1, 3rd edn. Biometrika Trust, London

Prentice RL (1986) Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. J Am Stat Assoc 81:321–327

Rajagopalan B, Lall U (1999) A *k*-nearest-neighbor simulator for daily precipitation and other weather variables. Water Resour Res 35:3089–3101

Sharma A, Lall U (1999) A nonparametric approach for daily rainfall simulation. Math Comput Simul 48:361–371

Vogel RM, Stedinger JR (1988) The value of stochastic streamflow models in overyear reservoir design applications. Water Resour Res 24:1483–1490

Wan H, Zhang X, Barrow EM (2005) Stochastic modelling of daily precipitation for Canada. Atmos Ocean 43:23–32

Wang QJ (1991) The POT model described by the generalized Pareto distribution with Poisson arrival rate. J Hydrol 129: 263–280

Wilks DM (1999) Interannual variability and extreme-value characteristics of several stochastic daily precipitation models. Agric For Meteorol 93:153–169

Young KC (1994) A multivariate chain model for simulating climatic parameters from daily data. J Appl Meteorol 33: 661–671

Appendix 1. Estimation of the parameters of the compound gamma distribution for wet-day rainfall

In Section 3, a compound gamma distribution for wet-day rainfall was defined by varying its scale parameter α between the *D*-day periods. Let $K \leq IJ$ be the number of wet *D*-day periods in the season of interest, i.e. *D*-day periods for which $N_D > 0$. Then the wet-day rainfall amounts x_{ik} can be represented as

$$x_{ik} = \alpha_k g_{ik},\ i = 1,\ldots,n_k;\ k = 1,\ldots,K \tag{A1}$$

with n_k the number of wet days in the *k*th wet *D*-day period, g_{ik} a gamma variable with scale parameter 1 and shape parameter ν, and α_k a random scale parameter. g_{ik} and α_k are assumed to be independent. It is further assumed that $1/\alpha_k$ has a gamma distribution with scale parameter β and shape parameter τ.

For the derivation below we need the first 2 moments of α_k, which are given by

$$\mathrm{E}(\alpha_k) = \frac{1}{\beta(\tau-1)}, \quad \tau > 1 \tag{A2}$$

$$\mathrm{E}(\alpha_k^2) = \frac{1}{\beta^2(\tau-1)(\tau-2)}, \quad \tau > 2 \tag{A3}$$

The model for the wet-day rainfall amounts has 3 unknown parameters: ν, β and τ. In this study, these parameters were estimated from the mean, the within *D*-day periods sums of squares SS_w and the between *D*-day periods sums of squares SS_b of wet-day rainfall. The latter 2 quantities are defined by:

$$SS_\mathrm{w} = \sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_{ik} - \bar{x}_{\cdot k})^2 \tag{A4}$$

$$SS_\mathrm{b} = \sum_{k=1}^{K} n_k(\bar{x}_{\cdot k} - \bar{x})^2 \tag{A5}$$

with *n* the total number of wet days,

$$n = \sum_{k=1}^{K} n_k \tag{A6}$$

$\bar{x}_{\cdot k}$ the mean wet-day rainfall amount in the *k*th wet *D*-day period,

$$\bar{x}_{\cdot k} = \sum_{i=1}^{n_k} x_{ik} / n_k \tag{A7}$$

and $\bar{x}$ the overall mean wet-day rainfall amount,

$$\bar{x} = \sum_{k=1}^{K} n_k \bar{x}_{\cdot k} / n \tag{A8}$$

For the mean wet-day rainfall amount, it follows from Eqs. (A1) & (A2):

$$\mu_\mathrm{wet} = \mathrm{E}(x_{ik}) = \mathrm{E}(\alpha_k)\mathrm{E}(g_{ik}) = \frac{\nu}{\beta(\tau-1)}, \quad \tau > 1 \tag{A9}$$

Substitution of Eq. (A1) into Eq. (A4) results in:

$$SS_\mathrm{w} = \sum_{k=1}^{K}\sum_{i=1}^{n_k}(\alpha_k g_{ik} - \alpha_k \bar{g}_{\cdot k})^2 = \sum_{k=1}^{K}\alpha_k^2\sum_{i=1}^{n_k}(g_{ik} - \bar{g}_{\cdot k})^2 \tag{A10}$$

with $\bar{g}_{\cdot k} = \sum_{i=1}^{n_k} g_{ik}/n_k$. For the mean of SS_w it then follows:

$$\begin{aligned}\mathrm{E}(SS_\mathrm{w}) &= \sum_{k=1}^{K}\mathrm{E}(\alpha_k^2)\mathrm{E}\left[\sum_{i=1}^{n_k}(g_{ik} - \bar{g}_{\cdot k})^2\right] \\ &= \sum_{k=1}^{K}(n_k - 1)\mathrm{E}(\alpha_k^2)\mathrm{var}(g_{ik}) \\ &= \frac{(n-K)\nu}{\beta^2(\tau-1)(\tau-2)}, \quad \tau > 2\end{aligned} \tag{A11}$$

using Eq. (A3).

For SS_b we can write:

$$SS_\mathrm{b} = \sum_{k=1}^{K} n_k\left[(1 - n_k/n)\bar{x}_{\cdot k} - \sum_{j \neq k}^{K} n_j \bar{x}_{\cdot j}/n\right]^2 \tag{A12}$$

Since the mean of the term within square brackets equals zero, the mean of SS_b is given by:

$$\begin{aligned}\mathrm{E}(SS_\mathrm{b}) &= \sum_{k=1}^{K} n_k \ \mathrm{var}\left[(1 - n_k/n)\bar{x}_{\cdot k} - \sum_{j \neq k}^{K} n_j \bar{x}_{\cdot j}/n\right] \\ &= \sum_{k=1}^{K} n_k(1 - n_k/n)^2 \ \mathrm{var}\,\bar{x}_{\cdot k} + \sum_{k=1}^{K} n_k \sum_{j \neq k}^{K}(n_j/n)^2 \ \mathrm{var}\,\bar{x}_{\cdot j}\end{aligned} \tag{A13}$$

because the $\bar{x}_{\cdot j}$'s are independent. For the variances of the $\bar{x}_{\cdot j}$'s in the right-hand side of Eq. (A13) we can write:

$$\begin{aligned}\mathrm{var}(\bar{x}_{\cdot j}) &= \mathrm{var}(\alpha_j \bar{g}_{\cdot j}) = \mathrm{E}(\alpha_j^2)\mathrm{E}(\bar{g}_{\cdot j}^2) - [\mathrm{E}(\alpha_j)]^2[\mathrm{E}(\bar{g}_{\cdot j})]^2 \\ &= \frac{1}{\beta^2(\tau-1)(\tau-2)}\left(\frac{\nu}{n_j} + \nu^2\right) - \frac{\nu^2}{\beta^2(\tau-1)^2} \\ &= \frac{\nu}{n_j\beta^2(\tau-1)(\tau-2)} + \frac{\nu^2}{\beta^2(\tau-1)^2(\tau-2)}, \quad \tau > 2\end{aligned} \tag{A14}$$

Substitution of Eq. (A14) into Eq. (A13) gives, after some algebra:

$$\mathrm{E}(SS_\mathrm{b}) = \frac{\nu(K-1)}{\beta^2(\tau-1)(\tau-2)} + \frac{\nu^2\left(n - \sum_{k=1}^{K} n_k^2/n\right)}{\beta^2(\tau-1)^2(\tau-2)}, \quad \tau > 2 \tag{A15}$$

The estimates of ν, β and τ are readily obtained by replacing μ_wet, $\mathrm{E}(SS_\mathrm{w})$ and $\mathrm{E}(SS_\mathrm{b})$ in Eqs. (A9), (A11) & (A15) by $\bar{x}$, SS_w and SS_b, respectively. The fit to SS_w and SS_b ensures that the total variance of wet-day rainfall is preserved as in the model with fixed parameters for the gamma distribution.

Appendix 2. Conditional simulation of large values from the generalized Pareto distribution

Eq. (18) defined the generalized Pareto distribution for daily values exceeding a threshold u. Here, the joint density of the m largest values from this distribution is derived under the condition that the (m + 1)th largest value is known. This condition ensures that the simulated values in Table 2 exceed the third largest value for the season of interest in the historical record. The ordered (m + 1) largest values are denoted as $Z_1 \geq Z_2 \geq \ldots \geq Z_{m+1}$.

For large u it may be assumed that the exceedance times form a Poisson process. The largest value Z_1 has then a GEV distribution (e.g. Buishand 1989, Wang 1991):

$$H(z) = \Pr(Z_1 \leq z) = \mathrm{e}^{-\lambda\{1-\theta(z-u)/\alpha\}^{1/\theta}}, \quad z \geq u \qquad \text{(B1)}$$

where λ is the expected number of exceedances of the threshold u. For $\theta \to 0$, this distribution reduces to the Gumbel distribution:

$$H(z) = \mathrm{e}^{-\lambda \mathrm{e}^{-(z-u)/\alpha}}, \quad z \geq u \qquad \text{(B2)}$$

The joint density of $Z_1, \ldots, Z_m$ conditionally on $Z_{m+1} = z_{m+1}$ is defined as:

$$h_{1,\ldots,m\,|\,m+1}(z_1,\ldots,z_m \mid Z_{m+1} = z_{m+1}) = \frac{h_{1,\ldots,m+1}(z_1,\ldots,z_{m+1})}{h_{m+1}(z_{m+1})} \qquad \text{(B3)}$$

where $h_{1,\ldots,m+1}(z_1,\ldots,z_{m+1})$ is the joint density of $Z_1, \ldots, Z_{m+1}$, and $h_{m+1}(z_{m+1})$ is the marginal density of $Z_{m+1.}$ For these densities, the following expressions can be derived (e.g. Buishand 1989, Embrechts et al. 1997 their Section 4.2):

$$h_{1,\ldots,m+1}(z_1,\ldots,z_{m+1}) = h(z_{m+1})\prod_{j=1}^{m}\frac{h(z_j)}{H(z_j)} \qquad \text{(B4)}$$

$$h_{m+1}(z_{m+1}) = h(z_{m+1})\frac{\{-\ln H(z_{m+1})\}^m}{m!} \qquad \text{(B5)}$$

where $h(z) = H'(z)$ is the density of Z_1. Using these expressions in Eq. (B3) gives:

$$h_{1,\ldots,m\,|\,m+1}(z_1,\ldots,z_m \mid Z_{m+1} = z_{m+1}) = m!\prod_{j=1}^{m}\frac{h(z_j)}{\{-H(z_j)\ln H(z_{m+1})\}} \qquad \text{(B6)}$$

The right-hand side of Eq. (B6) represents the joint density of an ordered sample of size m from the density:

$$k(x) = \frac{h(x)}{\{-H(x)\ln H(z_{m+1})\}}, \quad x \geq z_{m+1} \qquad \text{(B7)}$$

Thus the conditional simulation of the m largest values reduces to the generation of m independent values from the density $k(x)$ or the distribution function:

$$K(x) = \int_{z_{m+1}}^{x} k(z)\mathrm{d}z = 1 - \frac{\ln H(x)}{\ln H(z_{m+1})}, \quad x \geq z_{m+1} \qquad \text{(B8)}$$

Substitution of Eqs. (B1) & (B2) into (B8) results in:

$$K(x) = \begin{cases} 1-\{1-\theta(x-z_{m+1})/\alpha^*\}^{1/\theta}, & \text{if } \theta \neq 0 \\ 1-\mathrm{e}^{-(x-z_{m+1})/\alpha^*}, & \text{if } \theta = 0 \end{cases} \qquad \text{(B9)}$$

for $x \geq z_{m+1}$, and where $\alpha^* = \alpha - \theta\,(z_{m+1} - u)$. Eq. (B9) represents a GPD with the same shape parameter θ as in Eq. (18), but with a changed scale parameter if $\theta \neq 0$. This result is similar to that in DuMouchel (1983) and Wang (1991) for the truncation of the GPD at a fixed value > u.

Editorial responsibility: Bryson Bates,
Wembley, Western Australia, Australia

Submitted: March 23, 2006; Accepted: July 10, 2007
Proofs received from author(s): September 7, 2007