Application of the partial duration series approach in the analysis of extreme rainfalls

H. MADSEN & D. ROSBJERG
Institute of Hydrodynamics and Hydraulic Engineering, Technical University of Denmark, DK-2800 Lyngby, Denmark

P. HARREMOÉS
Department of Environmental Engineering, Technical University of Denmark, DK-2800 Lyngby, Denmark

Abstract Since 1979 a country-wide system of recording raingauges has been operated in Denmark in order to obtain a better basis for design and analysis of sewer systems. As an alternative to the traditional non-parametric methods the Partial Duration Series method is employed in the analysis of the recorded series. The extreme value modelling is applied to maximum rain intensity of specified duration during individual storms. In the at-site analysis the classical model assumptions comprising Poisson-distributed annual number of threshold exceedances and independent exponentially distributed exceedances are critically examined. In case of non-exponential exceedances the generalized Pareto distribution is adopted. On the basis of the at-site modelling a regional analysis of the distribution parameters is carried out. Finally the regional and the local rain information is weighted using Bayesian theory.

INTRODUCTION

In the design and analysis of sewer systems the effects of floods and acute pollution have to be evaluated on the basis of statistics for extreme rainfalls. In Denmark only very few rainfall records have been available as basis for country-wide design. In order to obtain a better design basis a new system of recording raingauges was introduced in 1979. At present rain measurements from 41 stations with a recording period of 10 years or more are available. The recorded series consist of measurements of cumulative precipitation depth within each storm with a time resolution of one minute. The separation of individual storms is defined as periods exceeding one hour without precipitation. In this paper the maximum rain intensity of the duration 10 minutes during individual storms is analyzed.

As an alternative to the traditional non-parametric methods using intensity-duration-frequency-curves and historical rain series, respectively, the Partial Duration Series (PDS) method (also denoted the Peak Over Threshold method) is employed. The basic PDS model introduced in hydrology by Shane & Lynn (1964) and Todorovic & Zelenhasic (1970) comprises the assumptions of Poisson-distributed annual number of threshold exceedances and independent exponentially distributed exceedances. In the at-site analysis the classical model assumptions are examined in detail. In addition, the trade-off between this parameter-parsimonious model and a more accurate description of the rainfall process taking non-exponential exceedances into consideration is carried out.

On basis of the at-site modelling a regional analysis is performed in order to
improve at-site estimates and to make inferences at non-monitored sites. This analysis is based on the Bayesian approach combining site-specific and regional information using Bayes’ theorem. By means of the regional analysis empirical regional distributions of the PDS parameters are used as prior information.

THE PARTIAL DURATION SERIES MODEL

By introducing a threshold level $I_0$ in the hydrological series of 10 minutes maximum rain intensity $I_i$ (denoted rain intensity in the following) and considering only peaks above this level, i.e. $X_i = I_i - I_0$, a PDS is obtained. The basic model assumes that the occurrence of peaks above the threshold level is described by a Poisson process with constant or one-year periodic intensity. Hence, the number of exceedances $N$ in time $t$ is Poisson-distributed:

$$P \{ N = n \} = \frac{(\lambda t)^n}{n!} \exp(-\lambda t)$$ (1)

where $\lambda$ equals the expected number of exceedances per year.

The exceedance magnitudes are assumed to be independent and identically distributed (i.i.d.) following the exponential distribution (ED). The ED with the mean $\alpha$ has the probability density function $f(x)$ and cumulative distribution function $F(x)$:

$$f(x) = \frac{1}{\alpha} \exp \left( -\frac{x}{\alpha} \right) ; \quad F(x) = 1 - \exp \left( -\frac{x}{\alpha} \right)$$ (2)

The T-year event $x_T$, i.e. the level which on the average is exceeded once in T years, is in a PDS context defined as the $(1-1/\lambda T)$ quantile in the distribution of the exceedances (see e.g. Rosbjerg, 1985):

$$x_T = \alpha \ln(\lambda T)$$ (3)

Replacing $\alpha$ and $\lambda$ in (3) with the maximum likelihood estimators:

$$\hat{\alpha} = \frac{1}{N} \sum_{i=1}^{N} x_i \quad ; \quad \hat{\lambda} = \frac{N}{t}$$ (4)

the T-year event estimator $\hat{x}_T$ is obtained. By a Taylor series expansion of $\hat{x}_T$ the mean and the variance become (Rosbjerg et al., 1991):

$$E \{ \hat{x}_T \} = \alpha \ln(\lambda T) - \frac{1}{2} \frac{\alpha}{\lambda t} \quad ; \quad Var \{ \hat{x}_T \} = \frac{\alpha^2}{\lambda t} \left[ 1 + K(\ln \lambda T)^2 \right]$$ (5)

where the correction factor $K$ is (Rosbjerg, 1985):

$$K = \frac{1}{e^{\lambda t} - 1} \sum_{n=1}^{\infty} \frac{(\lambda t)^{n+1}}{n n!}$$ (6)

Although the basic PDS model has shown excellent performance in many applications, it may in some cases be difficult to justify the ED assumption unless a very high threshold is chosen. In case of exceedances with a heavier tail than the exponential due to possible presence of outliers, the generalized Pareto distribution
Partial duration series approach in the analysis of extreme rainfalls

(GPD) with negative shape parameter is a natural choice. The GPD is introduced in PDS modelling by Van Montfort & Witter (1986), and it contains the ED as a special case. For the shape parameter \( \kappa \neq 0 \) and scale parameter \( \alpha \) the probability density function and the cumulative distribution function read:

\[
f(x) = \frac{1}{\alpha} \left( 1 - \frac{x}{\alpha} \right)^{1-\kappa} \quad ; \quad F(x) = 1 - \left( 1 - \frac{x}{\alpha} \right)^{1-\kappa}
\]  

(7)

with mean and variance:

\[
E\{X\} = \frac{\alpha}{1+\kappa} \quad ; \quad Var\{X\} = \frac{\alpha^2}{(1+\kappa)^2(1+2\kappa)}
\]  

(8)

The T-year event is easily obtained from (7):

\[
X_T = \frac{\alpha}{\kappa} \left[ 1 - \left( \frac{1}{\lambda T} \right)^{\kappa} \right]
\]  

(9)

An evaluation of the properties of the T-year event estimator is performed by Rosbjerg et al. (1992), and they recommended the classical method of moments (MOM) estimation. Inserting the MOM estimators of the GPD parameters in (9) the T-year event estimator is obtained.

AT-SITE ANALYSIS

The first step in the at-site analysis is the extraction of peaks from the historical records. If no direct physical interpretation is possible in the determination of the threshold level, other criteria have to be used. Different methods for threshold determination are discussed in Rosbjerg & Madsen (1992). An objective formulation of the threshold level is

\[
\theta_0 = \frac{\lambda}{\kappa} \left[ E\{I\} + k \cdot S\{I\} \right]
\]  

(10)

where \( k \) is the standard frequency factor, and \( E\{I\} \) and \( S\{I\} \) is the mean and the standard deviation, respectively, of the rain intensity.

A fundamental assumption made above in the formulation of the PDS model is the assumption of i.i.d. exceedances. Dependence between successive peaks should be sought among exceedances separated by short time intervals. For a relatively low threshold level corresponding to \( k=1 \) scatter diagrams of exceedances \( x_i \) against \( x_{i+1} \) separated with a time difference smaller than one day show no significant trend. Hence it is concluded that successive events in the PDS are independent. Whether the exceedances are identically distributed may be examined by considering the seasonal variation of exceedances. It is well known that the most intense rainfalls occur more frequently in the summer as shown in Fig. 1a. The magnitudes, however, show no significant seasonal variation (see Fig. 1b), and therefore it is concluded that the exceedances are identically distributed.

To check the goodness-of-fit of the basic model assumptions adequate test statistics are employed. A test of the Poisson assumption using the Fisher dispersion test
the statistic is described by Cunnane (1979). The test statistic R reads:

\[ R = \sum_{i=1}^{t} \frac{(n_i - \hat{\lambda})^2}{\hat{\lambda}} \]  

where \( n_i \) is the number of threshold exceedances in year \( i \) of a total record of \( t \) years. Under the \( H_0 \)-hypothesis (Poisson-distributed annual number of exceedances) \( R \) is distributed as \( \chi^2 \) with \( t-1 \) degrees of freedom. Hence, the Poisson hypothesis is accepted at confidence level \( \alpha \) if \( \chi^2_{\alpha/2}(t-1) < R < \chi^2_{1-\alpha/2}(t-1) \). A test of the overall performance of the ED is conducted by use of the Kolmogorov-Smirnov test. Given an ordered sample \( x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(N)} \) the test statistic is:

\[ D = \text{MAX}_i \left| S(x_{(i)}) - \hat{F}(x_{(i)}) \right| \]  

where \( S(\bullet) \) is the sample-cumulative distribution function and \( \hat{F}(\bullet) \) is the cumulative ED function with estimated parameter \( \hat{\lambda} \). When the parameter is estimated from the sample, modified critical values of the test statistic must be used. A table of critical values is given in Lilliefors (1969). For high quantile estimation the correct tailing-off of the right tail may be crucial. A test for detecting outliers in exponential data can be conducted by considering the ED as a special case of the GPD (Van Montfort & Witter, 1985). By testing \( \kappa=0 \) (ED) against \( \kappa<0 \) (GPD) the test statistic is:

\[ Z = \frac{\text{Maximum}}{\text{Median}} = \begin{cases} 
\frac{x_{(N)}}{x_{((N+1)/2)}} & ; N \text{ odd} \\
\frac{2x_{(N)}}{x_{(N/2)} + x_{(N/2+1)}} & ; N \text{ even}
\end{cases} \]  

where \( x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(N)} \). Critical values of \( Z \) are given by Van Montfort & Witter (1985).
Results of the described tests applied to the 41 records of rain intensity are summarized in Table 1. A very high threshold has to be chosen if the basic model assumptions should be strictly satisfied. This, however, significantly reduces the sample sizes. Thus, in order to both keep the sample size adequate for reliable prediction and maintain a reasonable degree of model compliance, it was decided to use a threshold level corresponding to $k=3.5$. At this level the Poisson distribution is generally accepted, whereas the ED assumption, at a significant number of the stations, is rejected due to outliers. The final choice of exceedance model must be made with respect to both the descriptive and the predictive abilities. In order to fully eliminate the model error a large number of parameters are usually required. This interferes with the principle of parsimony. Hence, the criteria of unbiasedness must be balanced against the sampling variance. The variance of the T-year event estimator is easily obtained using either approximate formulas or Monte Carlo simulations. A measure of the model error is more difficult to handle. It is, however, evident that the ED implies larger bias than the GPD for high quantile estimation at stations with outliers, whereas the ED and the GPD are almost similar for lower quantile estimation (see Fig. 2). In order to account for both the descriptive and the predictive abilities a criterion for selecting the exceedance model is formulated as a function of the return period $T$. If the GPD estimate of the $T$-year event, $\hat{x}_{T,GPD}$, fulfills:

$$\hat{x}_{T,GPD} < E \{\hat{x}_{T,ED}\} + S \{\hat{x}_{T,ED}\}$$

where $E\{\hat{x}_{T,ED}\}$ and $S \{\hat{x}_{T,ED}\}$ are the mean and the standard deviation, respectively, of the ED estimate, then the ED is chosen as exceedance model. Otherwise the GPD is adopted. By means of this criterion it is concluded that the ED can be used at all stations for $T<10$ years. For $T \geq 10$ years the GPD has to be used at stations with significant outliers.

<table>
<thead>
<tr>
<th>$k$</th>
<th>The Fisher dispersion test</th>
<th>The Kolmogorov-Smirnov test</th>
<th>Test for detecting outliers</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>9</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td>2.5</td>
<td>4</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>3.0</td>
<td>7</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>3.5</td>
<td>3</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>4.0</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>4.5</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

**REGIONAL BAYESIAN ANALYSIS**

In the Bayesian analysis parameters are treated as stochastic variables in order to account for the imperfect knowledge of their exact values. The prior information about a parameter, say $\theta$, is expressed by a prior density function $f(\theta)$. Combining the prior knowledge and the sample information using Bayes' theorem the posterior distribution of $\theta$ is obtained:
Fig. 2 The empirical distribution of exceedances at station no. 25171 compared to the ED and the GPD, respectively, at a threshold level corresponding to $k=3.5$.

$$f^t(\theta) = \frac{l(\theta)f(\theta)}{\int_\Theta l(\theta)f(\theta)\,d\theta}$$  \hspace{1cm} (15)$$

where $l(\theta)$ is the sample likelihood function.

The Bayesian approach in a PDS context where regional information is used as prior knowledge has previously been described by Rasmussen & Rosbjerg (1991). Using regression techniques prior information about the exponential parameter was made available. In order to obtain the posterior distribution of the $T$-year event they used a non-informative prior of the $\lambda$-parameter. In this analysis prior information is expressed by means of an empirical distribution of both the $\alpha$ and the $\lambda$-parameter. For mathematical convenience conjugate priors are used to describe the regional variation of the PDS parameters.

In case of exponentially distributed exceedances the conjugate prior is an inverse gamma distribution. The prior density and the sample likelihood function of $\alpha$ are:

$$f_\alpha(a) = \frac{1}{\theta \Gamma(\beta)} \left[ \frac{\theta}{a} \right]^{\beta+1} \exp \left( -\frac{\theta}{a} \right) ; \quad l_\alpha(a) = \left( \frac{1}{a} \right)^N \exp \left( -\frac{S}{a} \right) ; \quad S = \sum_{i=1}^{N} x_i$$ \hspace{1cm} (16)$$

where $\Gamma(\beta)$ is the gamma function, and $(\beta, \theta)$ are the prior parameters. The posterior distribution is an inverse gamma distribution with updated parameters:

$$\beta_t = \beta + N ; \quad \theta_t = \theta + S$$ \hspace{1cm} (17)$$

For the Poisson distribution the conjugate prior is a gamma distribution. Hence, the prior density and the sample likelihood function of $\lambda$ are:

$$f_\lambda(\ell) = \frac{\tau}{\Gamma(\nu)} (\ell \tau)^{\nu-1} \exp(-\ell \tau) ; \quad l_\lambda(\ell) = \frac{(\ell \tau)^N}{N!} \exp(-\ell \tau)$$ \hspace{1cm} (18)$$

where $(\nu, \tau)$ are the prior parameters. The updated parameters in the posterior gamma distribution become:

$$\nu_t = \nu + N ; \quad \tau_t = \tau + t$$ \hspace{1cm} (19)$$
On applying the regional Bayesian approach to the records of rain intensity prior information is expressed by means of regional distributions of the PDS parameters. Using MOM estimation the prior parameters become:

$$
\hat{\beta} = 2 + \frac{\hat{\mu}_\alpha^2}{\hat{\sigma}_\alpha^2}; \quad \hat{\theta} = \hat{\mu}_\lambda \left(1 + \frac{\hat{\mu}_\alpha^2}{\hat{\sigma}_\alpha^2}\right); \quad \hat{\nu} = \frac{\hat{\mu}_\lambda^2}{\hat{\sigma}_\lambda^2}; \quad \hat{\tau} = \frac{\hat{\mu}_\lambda}{\hat{\sigma}_\lambda^2}
$$

(20)

where $\hat{\mu}$ and $\hat{\sigma}^2$ are the estimated mean and variance, respectively, in the regional samples of $\alpha$ and $\lambda$ values. In Fig. 3 histograms of $\alpha$ and the $\lambda$-parameter are compared to the conjugate prior densities. The regional variation of the PDS parameters is fairly well explained by the conjugate priors. In fact, both the inverse gamma and the gamma distributions offer considerable flexibility in modeling the populations of $\alpha$ and $\lambda$. It should be noted that the application of Bayes' theorem demands prior knowledge and sample information to be independent. Hence, the prior information at the site in consideration is obtained from the 40 other stations in the region.

The posterior distribution of $x_T$ can now be deduced by change of variables (Rouselle & Hindie, 1976). Assuming $\alpha$ and $\lambda$ to be independent the posterior distribution is found to be:

$$
f_{x_T}(x) = \int_0^\infty f_{\alpha}^f(a) f_{\lambda}^f(\ell) \left| \frac{dg}{dx} \right|_{a=g(s)} d\ell
$$

(21)

where the transformation $a = g(x) = x/(\ln T)$ is obtained from (3). The integral in (20) has to be calculated numerically. If no sample information is available, i.e. $(t,N,S) = (0,0,0)$, a prior distribution of the T-year event may be calculated from (20). This distribution is then used to make inferences at non-monitored sites.
The effect of combining site-specific and regional information is illustrated at one of the stations in Fig. 4. The site-specific distribution of the T-year event is obtained by using non-informative priors of the PDS parameters, i.e. \((\beta, \theta, \tau, \nu) = (0, 0, 0, 0)\), in the calculation of the T-year event. It is seen from Fig. 4 that the inclusion of regional information increases the precision of the T-year event significantly.

![Figure 4](image)

**Fig. 4** Distributions of the 2-year event at station no. 23321. (1) Regional prior distribution. (2) Site-specific posterior distribution. (3) Weighted posterior distribution.

In order to assess the regional procedure in general a rainfall record of 33 years is analyzed, and a comparison of (i) at-site estimation, and (ii) estimation with inclusion of the regional information is carried out. A measure of the improvement in quantile estimation precision using regional information may be defined as:

\[
I_R = \frac{S \{\hat{x}_T\} - S^r \{x_T\}}{S \{\hat{x}_T\}} \cdot 100\% \tag{22}
\]

where \(S \{\hat{x}_T\}\) is the standard deviation obtained from (5) using the at-site estimation procedure, and \(S^r \{x_T\}\) is the standard deviation in the posterior distribution of the T-year event. It is evident that the improvement in precision is a function of the amount of available at-site data. Assuming that the sample size is 1, 2, .., years, respectively, \(I_R\) is calculated successively from (22). In Fig. 5a \(I_R\) is shown as a function of the recording period. A substantial improvement of the T-year event precision is obtained if regional information is included in the estimation procedure. The improvement in precision is especially pronounced if only few years of at-site data are available.

When no at-site data are available, a regional procedure of quantile estimation is based on the prior distribution of the T-year event. If a raingauge is placed at the site in consideration, the T-year event may be updated successively, say every year. The effect of the updating procedure at time \(t=1, 2, .., \) years, respectively, is shown in Fig. 5b. The improvement in precision is calculated from (22) in which \(S \{\hat{x}_T\}\) is replaced by the standard deviation in the prior distribution of the T-year event. It is seen from Fig. 5b that the successive inclusion of at-site data rapidly reduces the uncertainty of the T-year event. This feature is important in a decision-making process.
procedure where loss due to imperfect knowledge has to be balanced against loss due to data collection arrangements. If the quantile estimation is based solely on at-site data, a relatively long recording period, say 5-10 years, is necessary in order to make reliable predictions. Using the regional approach at-site data yield valuable information from the very first year of recording.

In case of quantile estimation for return periods $T \geq 10$ years the ED assumption is no longer adequate at all stations. The regional Bayesian approach must be reformulated to include information from stations with outliers. Based on 32 stations with ED exceedances and 9 stations with GPD exceedances, respectively, regional distributions of both ED and GPD parameters are obtained. By introducing a suitable procedure for weighting the two types of stations the updated distribution of the $T$-year event can be deduced.

CONCLUSIONS

The PDS model is employed in the analysis of extreme rainfalls from a country-wide system of raingauges in Denmark. In the at-site analysis the model assumptions are examined in detail. It is concluded that the classical PDS approach with Poisson-distributed annual number of threshold exceedances and independent ED exceedances is adequate for quantile estimation with return periods $T < 10$ years. For $T \geq 10$ years outliers become significant in the estimation of the $T$-year event. At stations with identified outliers the GPD is adopted. A regional estimation procedure using Bayesian theory is introduced and illustrated in the case of ED exceedances. Regional information is included by means of empirical distributions of the PDS parameters. Combining site-specific and regional information the posterior distribution of the $T$-year event is found by change of variables. A substantial improvement in precision of the $T$-year event estimate is obtained when the regional information is included. Even for at-site series of 30 years the improvement in precision by using the Bayesian approach is significant.
In addition the Bayesian theory yields an estimation procedure at non-monitored sites which is extremely important for a general use of the method. If at-site data later become available, the T-year event may be updated successively. Even for at-site series of a few years the uncertainty in the updated distribution is reduced significantly compared to the uncertainty in the prior distribution of the T-year event.

REFERENCES


