EXTREME VALUE ANALYSIS OF PRECIPITATION SERIES IN THE SOUTH OF IBERIAN PENINSULA

P. 233

Hidalgo-Muñoz*, J.M., D. Argüeso, D. Calandria-Hernández, S.R. Gámiz-Fortis, M.J. Esteban-Parra and Y. Castro-Díez

Universidad de Granada. Granada, Spain

1. INTRODUCTION

Extreme rainfall events and the resulting floods usually cause a lot of damage to life and properties of human society. Determination of frequencies and magnitudes of these events are very important for flood plain management and designs of hydraulic structures, civil protection plans, etc. However, length of available records is not enough large to define the risk of floods, heavy rainfall, low flows, droughts, etc. In these cases, magnitude-frequency analysis, fitting samples to a frequency distribution, permit the estimation of how often a specific event will occur or the frequency of events greater than those observed during the period records. The frequency analysis includes the following underlying assumptions:

- The extremes are a random variable, and thus can be described by a distribution of probability.
- This distribution does not change from sample to sample (homogeneity).
- The data are independent.

The aim of this work is to determine the magnitudes of events associated to a predetermined frequency, fitting the samples to a more common distribution used for extreme events analysis. For these propose, two approaches for sampling extreme series have been used (annual maximum and partial duration), and the results obtained using both approaches have been compared.

2. DATA

2.1. Dataset and Quality Control

The data used comprise a set of 85 daily precipitation series located in the South of the Iberian Peninsula (Figure 1), covering the period 1955-2006. The data series have overcome a quality control test:

- Unrealistic or negative values have been checked against neighbour stations and historical records.
- Just a maximum of 10 % of missing values was permitted.
- A homogeneity test (RHtestV2) was applied to ensure the homogeneity of series used (http://cccma.seos.uvic.ca/ETCCDMI/software.shtml)

The simplest and more commonly employed method for sampling original data for extreme analysis is Annual Maximum (AM), where only the greatest event for each year is considered. A drawback of this method is that because of the fact that as just one event per year is considered, useful information from the second, third or greatest events of that year are lost. Moreover, some included events could be not really extremes. For these reasons, another approach to create the sample series is considerer: Partial Duration (PD), where all events above an a priori determined threshold (x_0) are included. This approach permits include more cases and it adapts better to heavy-tailed distribution (Madsen et al. 1997).

The classical PD series model includes the assumption that the number of threshold exceedandes are distributed under a Poisson process and their magnitudes are described by an exponential distribution (Todorovic and Zelenhasic 1970). This implies that AMS follows the Gumbel (Extreme Values type I) distribution. Also, several papers have focused on three parameters distributions: Generalized Pareto (GP) (Pickands 1975) distribution for PDS (Rosbjerg et al. 1992) and Generalized Extreme Values (GEV) distribution for AMS.

2.2. Partial Duration Series Modelling

One of the most delicate tasks in the PDS modelling is the threshold selection, since it is directly related to the assumption of independence of arrival times and exceedances. A low threshold level permits include more information, increasing the sample size and reducing the uncertainty of the analysis. On the other hand, if the threshold is too low, a violation of the assumption of independence can happen if the events are too close in time. In this sense, a declustering process reduces the problem of serial correlation to a large extent, allowing threshold values as low as 90^{th} percentile (Begueria 2005). For this reason, a declustering process has been carried out in this study, and consecutive days with precipitation values above 1 mm were clustered and only the greatest event of each cluster was selected.

**Corresponding author address:* Jose M. Hidalgo Muñoz, Univ. of Granada, Dept. of Applied Physics, Granada, E-18071; email: jhidalgo@ugr.es

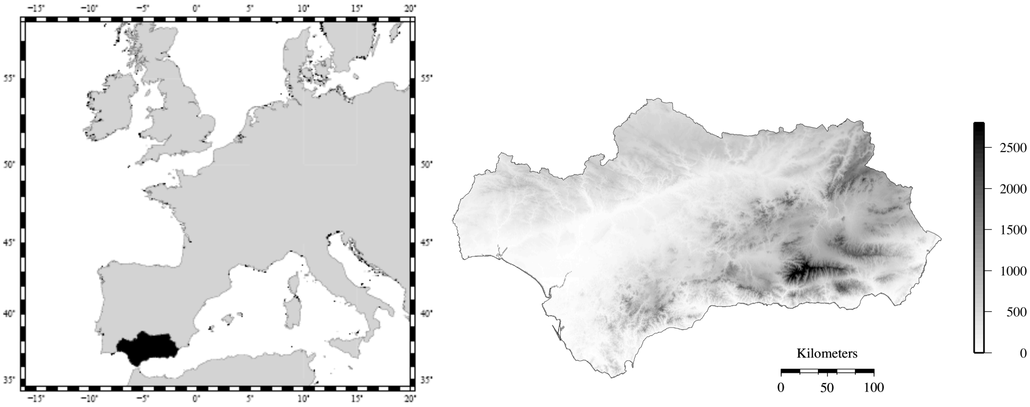

Figure 1. Location of the region of study inside western Europe (left) and topography, in meters (right).

Some systematic methods for the choice of the threshold value have been reviewed by Lang et al. (1999). In this work, the approach shown by Beguería (2005) for the choice of an adequate threshold was followed. PD series approach generally assumes that the number of exceedances follows a Poisson process (Cunnane 1979) The suitability of this assumption was tested by the Dispersion Index (DI) statistic (Cunnane 1979):

$$DI = \frac{s^2}{\hat{\lambda}} \qquad (1)$$

where s^2 is the annual number of exceedances estimated, and λ is the mean. If data follow a Poisson process DI should be close to 1. Confidence levels were calculated testing against a chi-squared distribution with M-1 degree of freedom, where M is the total number of years. Figure 2 shows that posteriorly to the declustering process, the number of exceedances is more suitable to follow a poisson process for different thresholds.

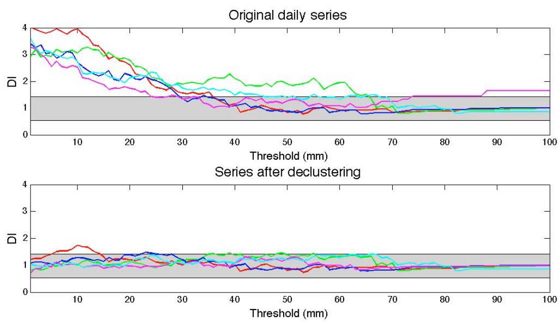

Figure 2. Example of DI statistic for different threshold values for 5 stations. In grey, confidence levels for DI at 95%, testing against chi-squared distribution. Original daily series, at the top, and series after declustering, at the bottom.

Mean Excess plot (ME plot) was used to determine the suitability of the GP model for the real data series at different threshold values (see Figure 3 as an example). In this plot, the average excess over a threshold against the value of this threshold is represented. A linear behaviour of ME plot appears if the variable follows a GP distribution over the threshold (Beguería 2005). Thus, a ME plot was represented for all stations. For most of them a linear behaviour appears until a value around 95-98th percentile. Therefore, a threshold value of 95th percentile, which favoured the series with as cases as possible, was selected to create PD series.

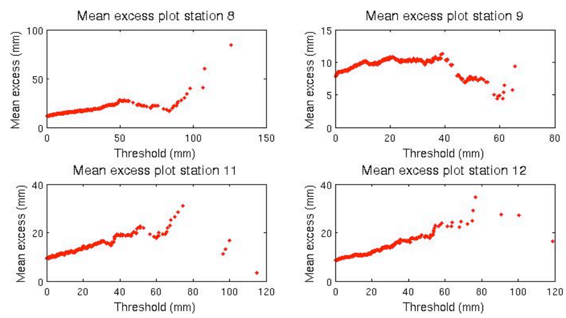

Figure 3. Mean excess plots for some stations.

3. METHODS.

The AM series are usually distributed by GEV (which includes ‘Gumbel‘ or ‘EV1’ distribution when k=0). The exceedandes magnitudes in the PD series are assumed to be GP (which includes Exponential distribution when k=0) distributed. For this reason, both distributions were proposed to be employed in this study. GEV and GP distributions have the following cumulative distribution function, F(x), and quantile function, x(F),:

Generalized extreme values (GEV)

Range of x: $-\propto < x \leq \xi + \alpha/\kappa$ if $\kappa > 0$; $-\propto < x < \propto$ if $\kappa = 0$; $\xi + \alpha/\kappa \leq x < \propto$ if $\kappa > 0$

$$F(x) = e^{-e^{y}} \quad (2)$$

$$y = \begin{cases} -\kappa^{-1}\ln\{1-\kappa(x-\xi)/\alpha\}, & \kappa \neq 0 \\ (x-\xi)/\alpha, & \kappa = 0 \end{cases}$$

$$x(F) = \begin{cases} \xi + \alpha\{1-(-\ln F)^{\kappa}\}/\kappa, & \kappa \neq 0 \\ \xi - \alpha \ln(-\ln F), & \kappa = 0 \end{cases} \quad (3)$$

Generalized Pareto (GP)

Range of x: $\xi < x \leq \xi + \alpha/\kappa$ if $\kappa > 0$; $\xi \leq x < \infty$ if $\kappa \leq 0$

$$F(x) = 1 - e^{-y} \quad (4)$$

$$y = \begin{cases} -\kappa^{-1}\ln\{1-\kappa(x-\xi)/\alpha\}, & \kappa \neq 0 \\ (x-\xi)/\alpha, & \kappa = 0 \end{cases}$$

$$x(F) = \begin{cases} \xi + \alpha\{1-(1-F)^{\kappa}\}/\kappa, & \kappa \neq 0 \\ \xi - \alpha \ln(1-F), & \kappa = 0 \end{cases} \quad (5)$$

Parameters: ξ (location), α (scale), κ (shape)

The parameters of the distributions have been estimated using L-moments. This approach, introduced by (Hosking 1990) is increasingly being used by hydrologist. L-Moments have the theoretical advantages over conventional moments of being able to characterize a wider range of distribution and, when estimated for a sample, being more robust to the presence of outliers in the data. More details and properties of L-moments can be found in Hosking and Wallis (1997). In summary, L-moments, λ_r, are linear combinations of Probability Weighted Moments (PWM) introduced by (Greenwood et al. 1979)The first L-Moments are the mean of the distribution (λ_1), a measure of the location, λ_2 is a measure of scale, λ_3 is a measure of skewness and λ_4 is a measure of kurtosis.

A useful dimensionless version of L-moments is defined by dividing the higher order L-moments by the scale measure. So, L-moments ratios are defined:

$$\tau_r = \frac{\lambda_r}{\lambda_2} \quad r=3,4,\ldots \quad (6)$$

L-moments ratios measure the shape of a distribution independently of its scale of measure. The L-moment ratios τ_3 and τ_4 are the L-skewness and L-kurtosis respectively.

The coefficient of L-variation is also defined as:

$$\tau = \frac{\lambda_2}{\lambda_1} \quad (7)$$

Although the L-Moments are theoretically defined for a probability distribution, in the practise, they are estimated for a given sample. Hence, the samples L-moments (l_r) and L–moments ratios (t_r) are defined. The samples L-moments l_r are an unbiased estimator of λ_r. In spite of the estimators of τ_r and τ are not unbiased, their bias are very small for moderate or large samples (Hosking and Wallis 1997)

4. RESULTS

4.1 Choice of a distribution

L-Moment ratio diagram

The L-moments ratio diagram is a graphical measure about if the data samples from different sites are consistent with the fitted probability distribution functions (see Figure 4). This shows the ratio between the L-kurtosis and L-skewness used to determined the goodness of fit for the selected distribution.

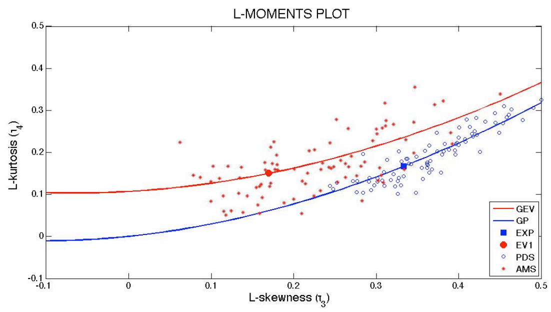

Figure 4. L-moments plot for samples and theoretic distributions used. The fitted distribution curves were drawn using polynomial approximations (Hosking 1990).

The 3 parameters distribution fit better to the samples due they are more flexible (because of the shape parameter inclusion). The cloud of points corresponding to PD series are closer to the GP distribution than that corresponding to AM series of GEV distribution. It can be due to PD series have more information that AM series.

Quantile-Quantile (QQ) plot

The QQ plots relate the quantiles derived of empirical probability distributions with the quantiles estimated by the fitted distributions. In this study, the empirical cumulative probability was assigned to the observed events via a plotting position formula:

$$CDF_{emp} = \frac{j - 0.35}{n} \quad (8)$$

where *j* is the *j*th observation sorted in ascending order and *n* is the total number of observations.

This plotting position formula was found to give good results for Wakeby GEV and GP distributions (Hosking and Wallis 1997).

The three parameters distributions show a better fit than the two parameters distributions. Both, GP and Exponential distribution fit better for smaller quantiles than GEV or EV1 distributions (see Figure 5).

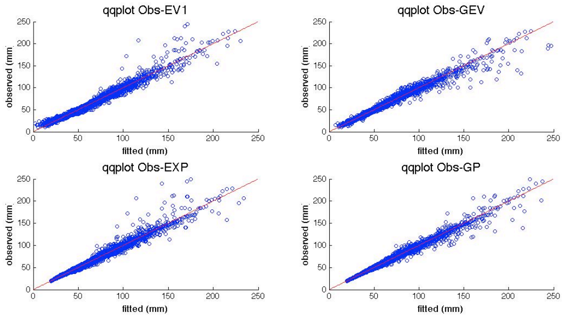

Figure 5. QQ plots for empirical and fitted distributions (EV1 - top left, GEV - top right, EXP- bottom left, GP - bottom left).

Uncertainty in the shape parameter

Although L-moment ratio diagrams as well as QQ plots show that the three parameters distributions fit better to the data, there is a large uncertainty involving the shape parameter in these distributions. Determining if the shape parameter differs significantly from zero is necessary. When the shape parameter is not significantly different from zero, two parameters distributions are advisable to use instead three parameters, due to their robustness. A test has been employed in this study to determine if the shape parameter is significantly different from zero, for GEV distribution (Hosking et al. 1985) and also for GP distribution (Rosbjerg et al. 1992). The results from both tests show that for most of stations (67% for GEV and 87% for GP) the shape parameter was not found significantly (at the 95 % confidence level) different from zero.

Summarizing the results obtained from L-Moment ration plots, QQ plots and the uncertainty in the shape parameter, in this work we have considered to use two parameters distributions (Exponential for PD series, and Gumbel or EV1 for AM series).

4.2. Estimation of the T-year return quantiles. Comparison between AMS-EV1 and PDS-EXP approaches.

The result of an extreme-values analysis is often simply a summary of quantiles corresponding to large cumulative probabilities. A p^{th} quantile, x_p, is the value with cumulative probability $F(x_p)=p$. Often these extreme probabilities are expressed as averaged return periods. A return period associated with a quantile x typically is interpreted to be the average time between occurrence of events of that magnitude or greater. The T-year event based on AM series is defined as the (1-1/T) quantile in the annual maximum distribution. In a PD series context, the T-year event is defined as the (1-1/λT) quantile in the distribution of exceedances. The quantile functions for both Gumbel and Exponential distributions are expressed in the Equations (3) and (5), respectively.

In this study, quantiles corresponding to four return periods (10, 25, 50 and 100 years) were calculated for both approach, AMS-EV1 and PDS-EXP.

The uncertainty in the estimation of quantiles for the different T return periods selected has been evaluated by a Montecarlo simulation. The simulation algorithm employed was similar to that used by (Madsen et al. 1997). For each case 10.000 samples were generated and the RMSE of the estimated T-year events were calculated. The relative RMSE for the different estimated T-year events are shown in the Figure 6.

Most of stations present a relative RMSE lower than 10%. Also, quantiles associated to higher return periods present higher relative RMSE. In general, lower relative RMSE were found with the PDS-EXP approach. An example of the spatial distribution of the estimated quantiles for 25 years return period, as well as its respective relative RMSE, is shown in the Figure 7. Note that the greatest relative RMSE appears in the southeast part of the region of study. In this area, the precipitation behaviour is different from the rest of the region, showing a lesser number of wet days, mainly consequence of the convective nature of the rainfall.

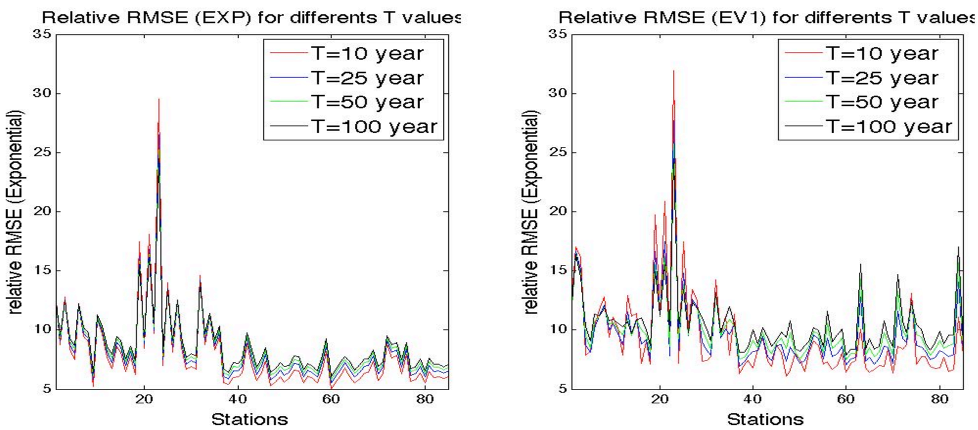

Figure 6. Relative RMSE in % for the four estimated quantiles associated to four return periods using the PDS-EXP approach (left), and the AMS-EV1 approach (right).

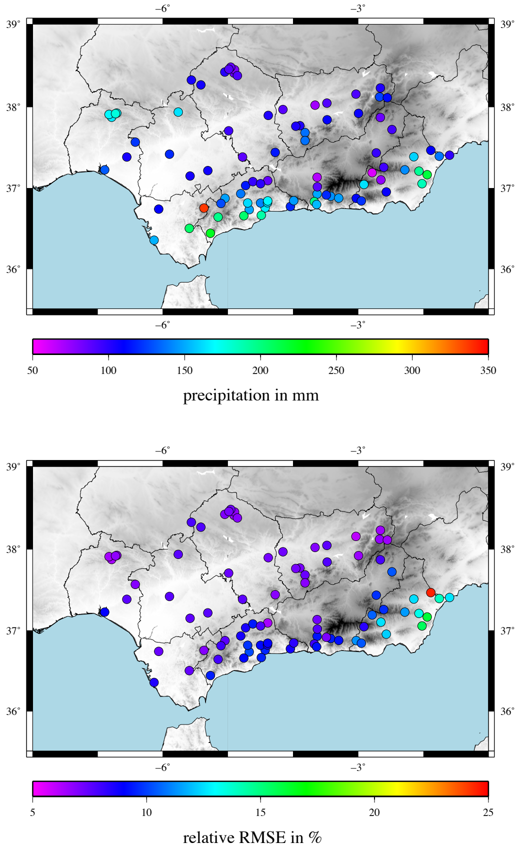

Figure 7. Estimated quantil values (in mm) for T=25 year using PDS-EXP approach (top), and their relative RMSE (in %) obtained through a Montecarlo simulation (bottom).

Finally, a comparison between the ratio of relative RMSE for both, PDS-EXP and AMS-EV1 approaches, and for all the quantiles estimated, against the length ratio of PD and AM series, for each station, was carried out (Figure 8).

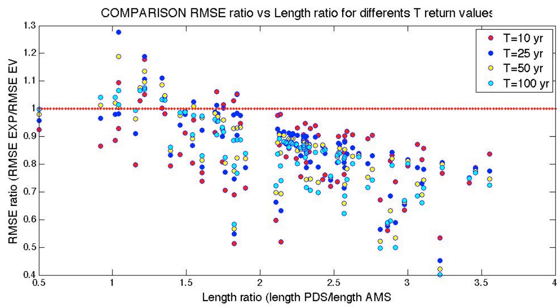

Figure 8. Comparison between the relative RMSEs for two approaches (PDS-EXP, AMS-EV1) in relation with the ratio between the length of PDS and AMS for all stations and return periods analyzed.

The relative RMSE for the PDS-EXP approach was found to be always lower for all the estimated quantiles than the relative RMSE for AMS-EV1 when the ratio between length of PD and AM series becomes greater than 2.

5. CONCLUSIONS

In this work we have calculated the occurrence probability of the precipitation extreme events in the South of the Iberian Peninsula using different approaches. Results can be summarized as follow:

- Due to the uncertainty that involves the shape parameter in the three parameters distributions (it was found non significant different from zero in most of the stations, for both, AMS and PDS approaches), two parameters distributions, Exponential and Gumbel, were selected to fit the PD series and AM series respectively.

- In general, the PDS-EXP fit shows lower relative RMSE for the estimated quantiles , for all the proposed return periods , than the AMS-EV1 fit, specially when the ratio between length of the PD series and AM series is greater than 2.

- In most of the stations, the relative RMSEs for estimated quantiles are below 10%, except in southern stations, where relative RMSEs around 10-20 % were found. For this latter case, to fit distributions to the series could be more difficult due to the convective nature of the precipitation in this region.

References

Begueria, S., 2005: Uncertainties in partial duration series modelling of extremes related to the choice of the threshold value. *J Hydrol*, **303,** 215-230.

Cunnane, C., 1979: A note on the Poisson assumption in partial duration series models. *Water Resour. Res.*, **15** (2), 257-271.

Greenwood, J., J. Landwehr, N. Matalas, and J. Wallis, 1979: Probability weighted moments: definition and relation to parameters of several distributions expressable in inverse form. *Water Resour. Res.*, **15** , 1049-1054.

Hosking, J., 1990: L-moments: analysis and estimation of distributions using linear combinations of order statistics. *J Roy Stat Soc B Met*, **52,** 105-124.

Hosking, J., J. Wallis, and E. Wood, 1985: Estimation of the Generalized Extreme Value distribution by the method of the Probability-Weighted moments. *Technometrics*, **27,** 251-261.

Hosking, J. R. M., and J. R. Wallis, 1997: *Regional Frequency Analysis: An Approach Based on L-Moments*: 244 pp.Cambridge University Press: Cambridge.

Madsen, H., P. Rasmussen, and D. Rosbjerg, 1997: Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events .1. At-site modeling. *Water Resour. Res.*, **33,** 747-757.

Pickands, J., 1975: Statistical inference using extreme order statistics. *Ann. Stat.*, **3**, 119-131.

Rosbjerg, D., H. Madsen, and P. Rasmussen, 1992: Prediction in partial duration series with generalized Pareto-distributed exceedances. *Water Resour. Res.*, **28,** 3001-3010.

Todorovic, P., and E. Zelenhasic, 1970: A stochastic model for flood analysis. *Water Resour. Res.*, **6**, 1641-1648.

ACKNOWLEDGMENTS

The Spanish Ministry of Science and Innovation, with additional support from the European Community Funds (FEDER), project CGL2007-61151/CLI, and the Regional Government of Andalusia, project P06-RNM-01622, have financed this study.