Log-Scaling rainfall data: effects on GPD Bayesian goodness of fit.

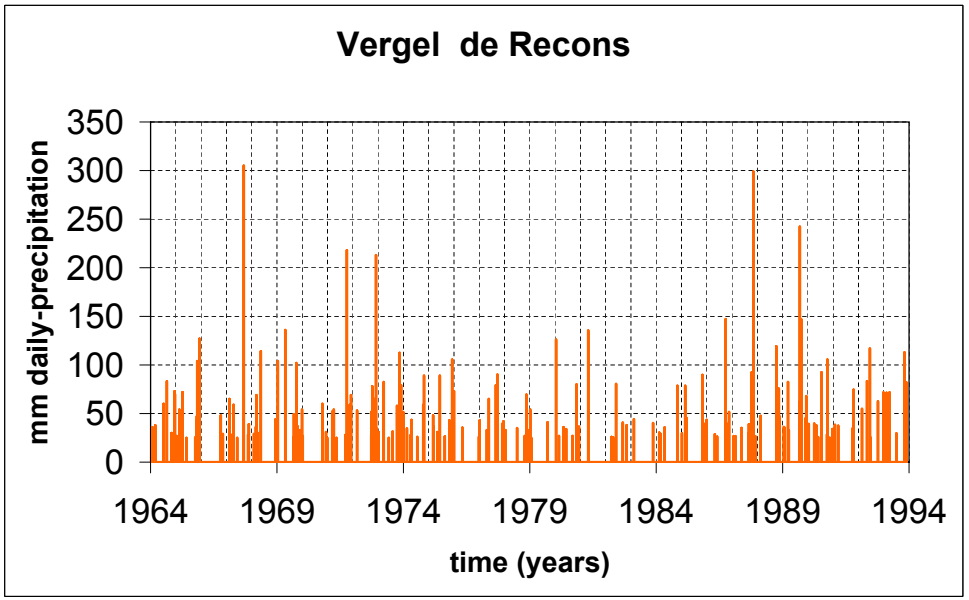
M.I. Ortego J.J. Egozcue

Departament de Matemàtica Aplicada III
E.T.S. Enginyeria Camins Canals Ports Barcelona (Civil Engineering)
Universitat Politècnica de Catalunya

4th conference on Extreme Value Analysis.
Gothenburg, 15-19 August 2005

Outline

Vergel de Racons data.



Main goals:

- Finding suitable model.
- Hazard analysis.
- Occurrence probabilities; return periods.

For reference, see Romero et al. (1998), [4]

The model

- Scale of the reference variable, precipitation:
  - ◇ is a positive variable:
    (0 mm rainfall is not rainfall!)
  - ◇ has a relative scale:
    50 mm is double than 25mm daily rainfall, but
    500mm and 525 mm daily rainfall is nearly the same!

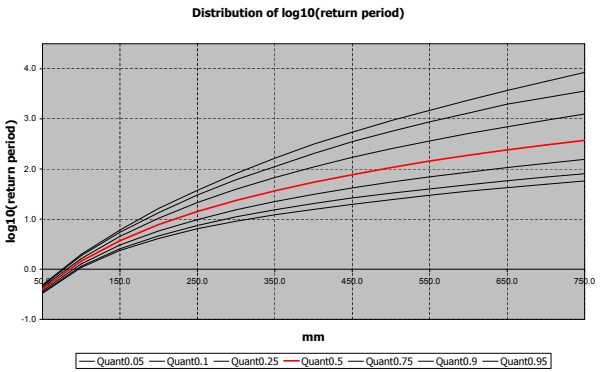
Logarithmic scale is needed!!!

The model

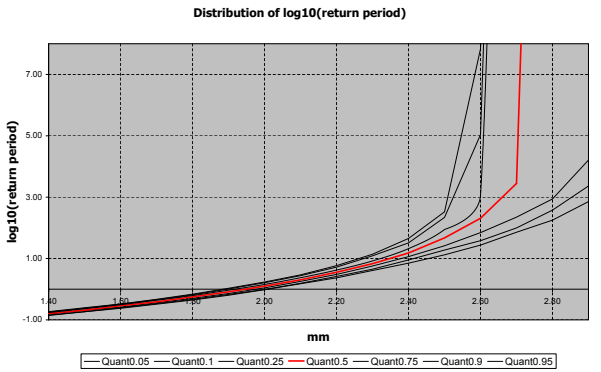
- Occurrence: Cramér-Lundberg model (Homogeneous Poisson process with intensity parameter λ).
- Magnitude: Excesses over threshold described by a Generalized Pareto Distribution (GPD).
- Bayesian parameter estimation.

Is it a suitable model?

- Hazard Estimates: At high levels, great uncertainty of estimates due to scarcity of data.
- Estimates of typical hazard parameters (e.g. Return period) vary *dramatically* depending on the selected model:

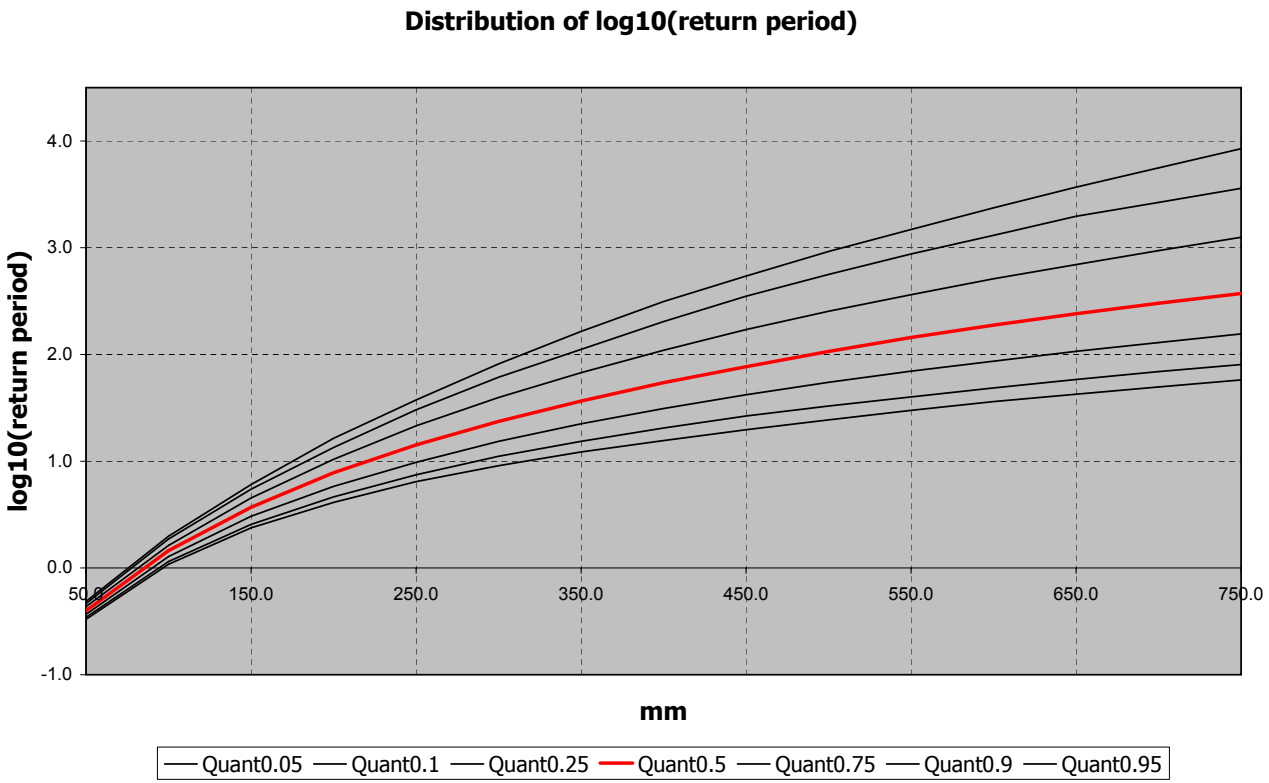


Return period of raw data

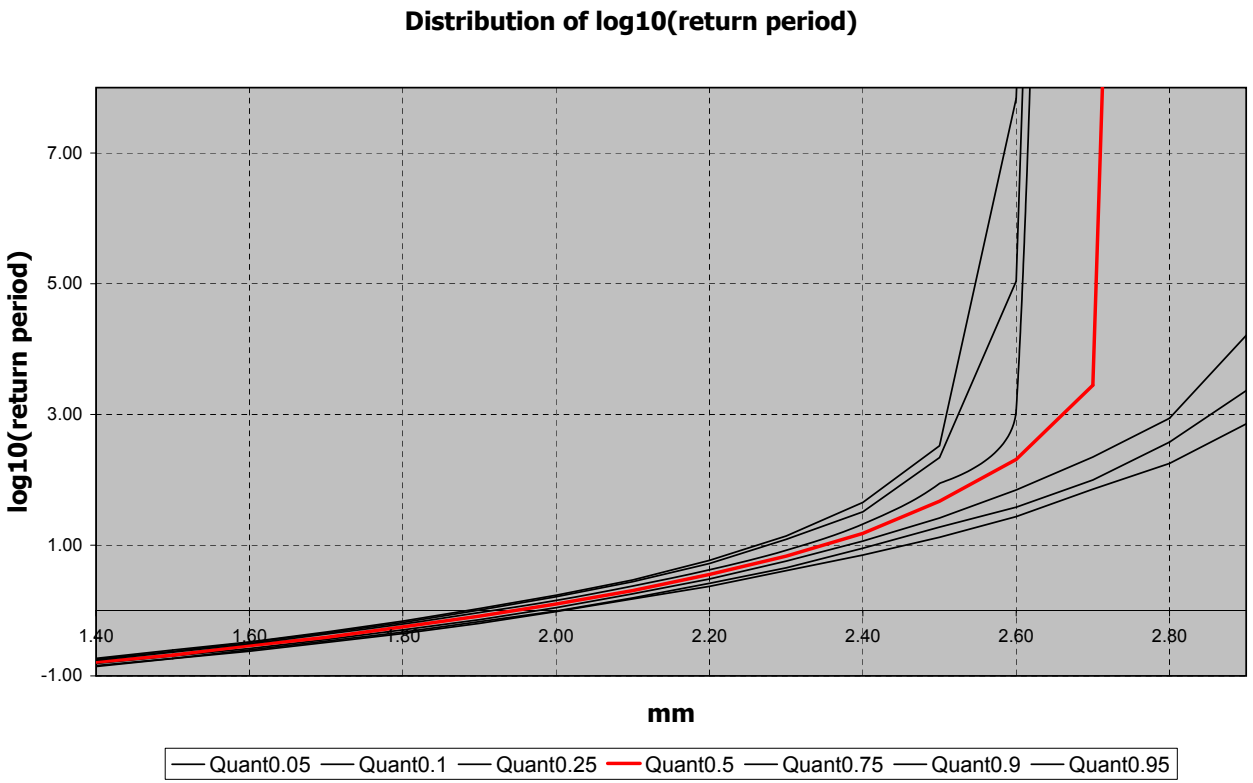


Return period of log data

Return period of raw data



Return period of log data



p-values

Whole model checking

(prior +likelihood +GPD) For reference, see Gelman et al. 1995, 1996, [6]
Several ways of checking it.

Goodness-of-fit checking

$GPD(\xi, \beta)$ goodness-of-fit assessing.
Several ways of checking it

p-values

A first approach:

plug-in-p-value (p_{plug})

$$p_{plug} = \mathrm{P}^{gpd(\cdot|\widehat{\xi},\widehat{\beta})}[t(\mathbf{X}) \geq t(\mathbf{x}_{obs})] \ ,$$

$gpd(\mathbf{x}|\xi,\beta)$ is replaced by $gpd(\cdot|\widehat{\xi},\widehat{\beta})$, where $\widehat{\xi},\widehat{\beta}$ is the maximum likelihood estimate of the parameters.

p-values

Bayesian *p*-values:

posterior predictive p-value (p_{post}) Guttman (1967) and Rubin (1984)

$$p_{post} = \mathrm{P}^{m_{post}(\cdot|\mathbf{x}_{obs})}[t(\mathbf{X}) \geq t(\mathbf{x}_{obs})] \ ,$$

where $m_{post}(\mathbf{x}|\mathbf{x}_{obs})$ is the *posterior predictive distribution*,

$$m_{post}(\mathbf{x}|\mathbf{x}_{obs}) = \int gpd(\mathbf{x}|\xi,\beta)\pi(\xi,\beta|\mathbf{x_{obs}})d(\xi,\beta) \ ,$$

and $\pi(\xi,\beta|\mathbf{x_{obs}})$ is the *posterior density* for ξ,β.

p-values

discrepancy p-value(p_{dis}) (Gelman et al., 1995)

The test statistic $t(\mathbf{X})$ is replaced by a discrepancy $t(\mathbf{X}, \xi, \beta)$

$$p_{dis} = \mathrm{P}^{m_{dis}(\cdot)}[t(\mathbf{X}, \xi, \beta) \geq t(\mathbf{x}_{obs}, \xi, \beta)] \ ,$$

where $m_{dis}(\mathbf{x}, \xi, \beta | \mathbf{x}_{obs})$ is ,

$$m_{dis}(\mathbf{x}, \xi, \beta | \mathbf{x}_{obs}) = gpd(\mathbf{x} | \xi, \beta) \pi(\xi, \beta | \mathbf{x_{obs}}) \ ,$$

and $\pi(\xi, \beta | \mathbf{x_{obs}})$ is the *posterior density* for ξ, β.

p-values: pros and cons

Desirable characteristics:

- Uniform distribution.
- Easy to compute.

Other useful characteristics:

- Known distribution of used statistic (even asymptotically).
- Easiness of interpretation.

Pros and cons:

- *plug-in p-value*: Easy to compute. Uncertainty ignored.
- *posterior predictive p-value* is not uniform. Easy to compute.
- *discrepancy p-value* is not uniform.

Bayesian GP Estimation (BGPE)

- Three parameters to estimate in the model:
  Poisson rate, λ, of $Poisson(\lambda)$ and ξ, β of the magnitude, modelled by $GPD(\xi, \beta)$:

$$GPD_X(x|\xi,\beta) = 1 - \left(1 + \frac{\xi}{\beta}x\right)^{\frac{-1}{\xi}}$$

- A suitable joint prior distribution for λ, ξ, β is set.
  Prior distributions for λ and ξ, β are independent $\rightarrow$ the joint prior factorizes:

$$\pi_{\lambda,\xi,\beta}(\lambda,\xi,\beta) = \pi_{\lambda}(\lambda) \cdot \pi_{\xi,\beta}(\xi,\beta)$$

- The joint likelihood of parameters, $L(\lambda, \xi, \beta | x_{obs})$, splits into two terms:

$$L(\lambda, \xi, \beta | x_{obs}) = L(\lambda | x_{obs}) \cdot L(\xi, \beta | x_{obs})$$

- Finally, the Posterior distribution of λ, ξ, β, $\pi_{\lambda,\xi,\beta}(\lambda, \xi, \beta | x_{obs})$, is obtained:

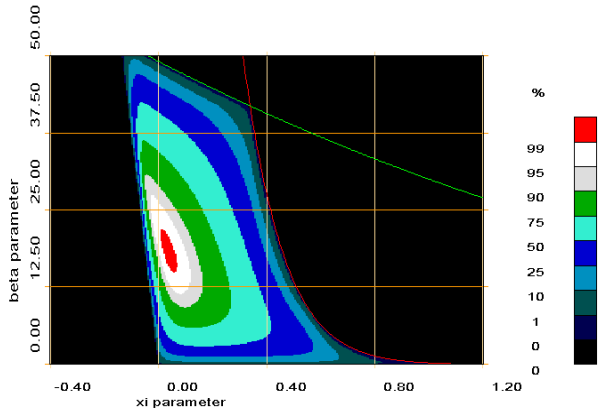
$$\pi_{\lambda,\xi,\beta}(\lambda, \xi, \beta | x_{obs}) = L(\lambda, \xi, \beta | x_{obs}) \cdot \pi_{\lambda}(\lambda) \cdot \pi_{\xi,\beta}(\xi, \beta)$$

Attention is set to marginal posterior distribution of ξ, β:

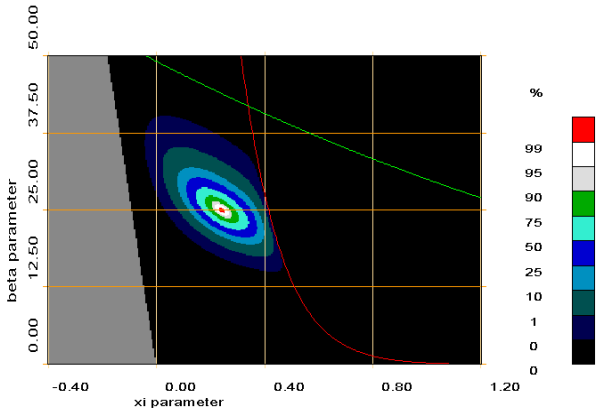
$$\pi_{\xi,\beta}(\xi, \beta | x_{obs}) = L(\xi, \beta | x_{obs}) \cdot \pi_{\xi,\beta}(\xi, \beta)$$

For reference, see Egozcue and Ramis (2001), [1], and Egozcue and Tolosana (2002), [2] .

Prior and posterior distributions : Raw data (I)



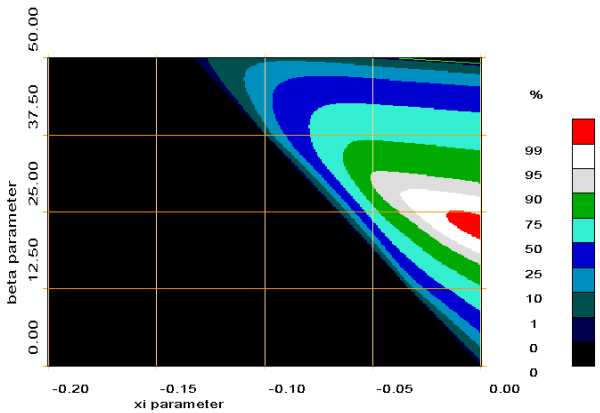
Prior density



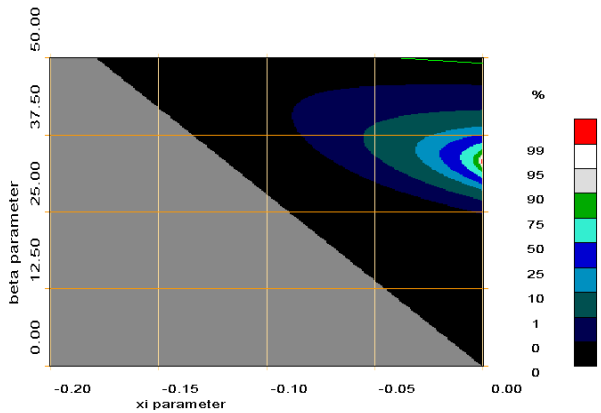
Posterior density

Something is lost!!

Prior and posterior distributions : Raw data (II)

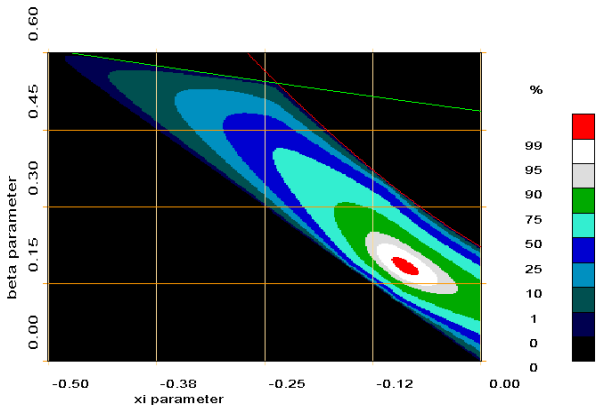


Prior density, $\xi < 0$

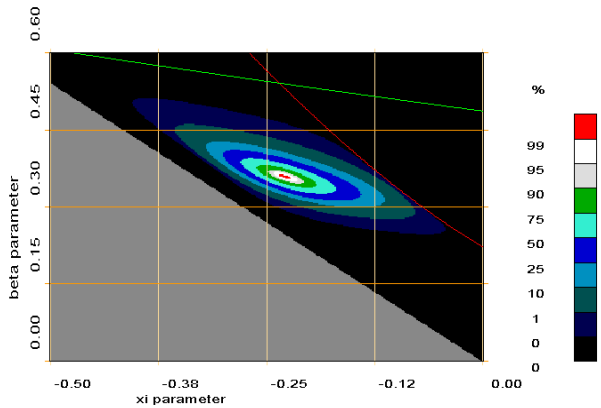


Posterior density, $\xi < 0$

Prior and posterior distributions: log data



Prior density



Posterior density

p-values: Our alternative

First approach

$$pp = \sum_i \psi_i p_i, \quad \textit{predictive KS } p - \textit{value} \text{ , where}$$

$p_i = KSGOF(\xi_i, \beta_i)$, for fixed (ξ_i, β_i) and $\psi_i = \pi(\xi_i, \beta_i | \mathbf{x_{obs}})$

Our alternative

$$pp = \Phi\left(\frac{\sum_{i=1}^{n} \psi_i \ \Phi^{-1}(p_i)}{\sqrt{\sum_i \psi_i^{\delta}}}\right) \ , \quad 1 \leq \delta \leq 2 \ , \delta \simeq 1 \ .$$

$p_i = KSGOF(\xi_i, \beta_i)$, for fixed (ξ_i, β_i); $\psi_i = \pi(\xi_i, \beta_i | \mathbf{x_{obs}})$

GPD Goodness-of-fit

	RAW DATA	LOG DATA
K-S posterior predictive Gof	$4.984 * 10^{-2}$	0.7028
K-S discrepancy *p*-value	$4.781 * 10^{-2}$	0.7549
Our approach	$\lesssim 2.075 * 10^{-2}$	$\gtrsim 0.9605$

(other statistics can be used)

Whole model assessing

Slope discrepancy

Estimation of the slope of the expected excesses regression line

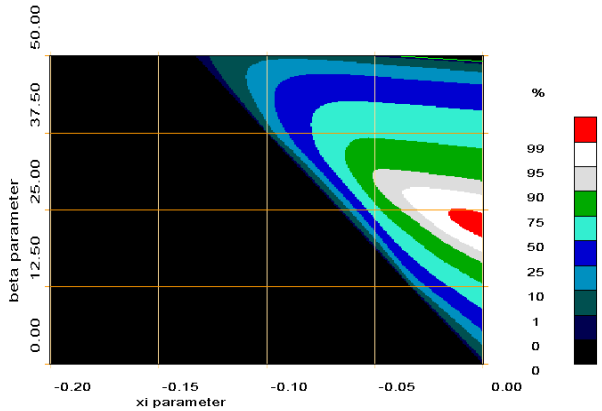
$$\mathrm{E}[X - u | X > u; \xi, \beta] = \frac{\beta + \xi u}{1 - \xi} \quad ,$$

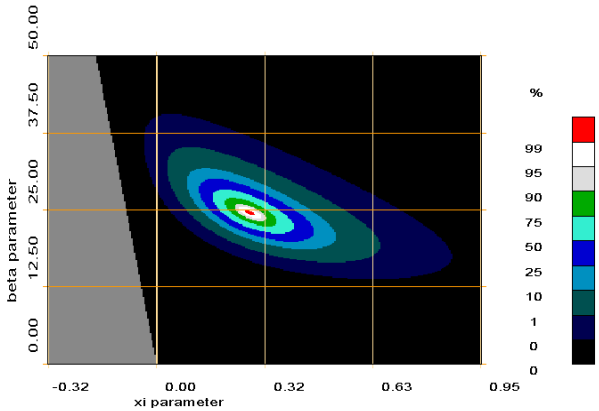
an estimator of ξ, $\xi < 1$.

Results:

	RAW DATA	LOG DATA
Slope discrepancy p-value	$8.765 * 10^{-2}$	0.4902

Priori vs. likelihood: raw data





Prior density, $\xi < 0$

Raw data likelihood

For raw data, likelihood is out of priori domain!

Conclussions

- Model fits better log-scaled data (i.e. Weibull d.a. prior, GPD)
  Checked by:
  - predictive goodness of fit ;
  - discrepancy bayesian *p*-value.
- Are Gumbel/Frechet d.a. admisible models for natural (finite) phenomena?

For Further Reading I

Egozcue , J.J. and Ramis, C.
Bayesian hazard analysis of heavy precipitation in Eastern Spain.
Int. J. Climatol., **21**: 1263-1279,2001.

Egozcue, J.J. and Tolosana-Delgado, R. (2002)
Program BGPE: Bayesian Generalized Pareto Estimation.
Ed. Diaz-Barrero, J.L., ISBN 84-69999125,Barcelona, Spain.

Egozcue, J.J, Pawlowsky-Glahn, V. and Ortego M.I.
Wave-height hazard analysis in Eastern Coast of Spain. Bayesian approach using generalized Pareto distribution.
Advances in Geosciences, **2**: 25-30, 2005.

For Further Reading II


Romero, R. , Guijarro J.A. , Ramis, C. and Alonso, S.
A 30-years (1964-93) daily rainfall data base for the Spanish Mediterranean regions: first exploratory study.
Int. J. Climatol., **18**: 541-560,1998.


Bayarri,M.J. and Berger, J.O.
P-values for composite null models.
J. of the Am. Stat. Ass., **95**: 1127-1142, 2000.


Gelman, A. and Carlin, J.B. and Stern, H.and Rubin, D.B.
Bayesian data analysis.
Wiley, 1995.