Estimation of non-stationary GEV model parameters

S. El-Adlouni, T. Ouarda & X. Zhang, R. Roy & B. Bobée

Extreme Value Analysis
15-19 August 2005

Hydro Québec

Statistical Hydrology Chair (INRS-ETE)

Outline

- Problem definition
- Objectives
- General Extreme Value Distribution
- Non-stationary GEV model
- Parameter estimation
- Simulation based comparison of estimation methods
- Case study
- Conclusions

Position of the problem

- ✓ In frequency analysis, data must generally be independent and identically distributed (i.i.d) which implies that they must meet the statistical criteria of independence, stationarity and homogeneity
- ✓In reality, the probability distribution of extreme events can change with time
- ✓Need to develop frequency analysis models which can handle various types of non-stationarity (trends, jumps, etc.)

Objectives of the study

Develop tools for frequency analysis in a non-stationary framework

✓Include the potential impacts of climate change

✓ Explore the case of trends or dependence on covariables

✓ Bayesian framework

GEV distribution

Y is GEV (Generalised Extreme Value) distributed if :

$$F_{GEV}(y) = \exp\left[-\left(1-\frac{\kappa}{\alpha}(y-\mu)\right)^{1/\kappa}\right] \qquad 1-\frac{\kappa}{\alpha}(y-\mu)>0$$

$\mu(\in\mathbb{R}), \alpha\ (>0)$ et $\kappa(\in\mathbb{R})$

are respectively the location, scale and shape parameters.

Non-stationary GEV model

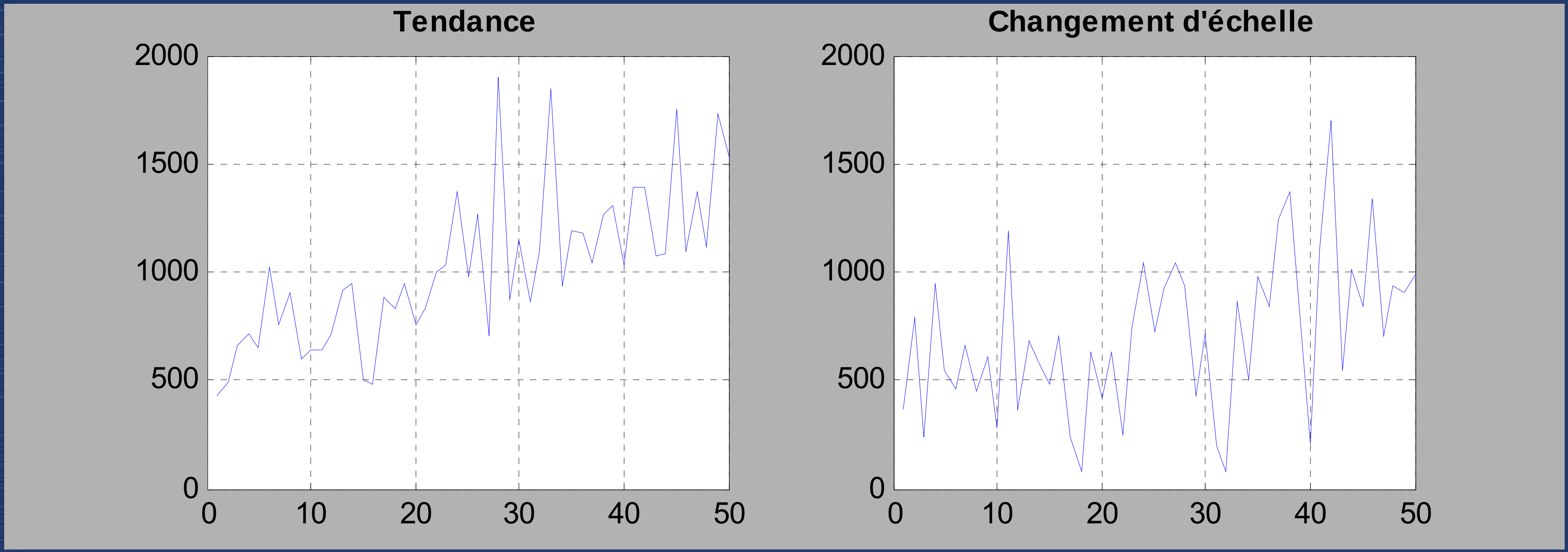
Non-stationary framework:

$$Y_t \sim GEV\left(\mu_t, \alpha_t, \kappa_t\right)$$

Parameters are function of time or other covariates.

Non-stationary GEV model

Illustration of the two types of non-stationarity



Non-stationary GEV model

I $$X_t \sim GEV_0(\mu, \alpha, \kappa)$$

Classic model : all parameters are constant.

II $$X_t \sim GEV_1(\mu_t = \beta_1 + \beta_2 Y_t, \alpha, \kappa)$$

The location parameter is a linear function of a covariable. The other two parameters are constant.

III $$X_t \sim GEV_2(\mu_t = \beta_1 + \beta_2 Y_t + \beta_3 Y_t^2, \alpha, \kappa)$$

The location parameter is a quadratic function of a temporal covariable. The other two parameters are constant.

Parameter Estimation

1. Maximum likelihood method (ML)
2. Bayesian model (Bayes)
3. Generalised maximum likelihood method (GML)

Maximum Likelihood Method

Properties of ML estimators

Under some regularity conditions, ML estimators have the desired optimality properties.

These regularity conditions are not met when the shape parameter is different from 0, since the support of the distribution depends on parameters (Smith 1985).

For small samples, the numerical resolution of the ML system can generate parameter estimators that are physically impossible and leads to very high quantile estimator variances.

Bayesian estimation : prior distribution

Prior distribution of parameter vector $\theta = (\mu, \alpha, \kappa)$

Fisher information matrix

$$I_{ij}(\theta) = E\left(-\frac{\partial^2 \ln f\left(\underline{y} \mid \theta\right)}{\partial\theta_i \, \partial\theta_j} \right)$$

Jeffrey's information prior

$$J(\theta) = \left| I(\theta) \right|^{\frac{1}{2}}$$

For the GEV distribution, the Fisher information matrix is given by Jenkinson (1969)

Bayesian estimation

GEV_0 model

In the absence of any additional information about the parameters (regional information, historic, expert opinion, etc.), we consider the Jeffrey's non-informative prior.

$$\pi_0\left(\mu,\alpha,\kappa\right)=J\left(\mu,\alpha,\kappa\right)$$

Bayesian estimation

GEV$_1$ model

$$\pi_1(\beta_1, \beta_2, \alpha, \kappa) = J(\beta_1, \alpha, \kappa)\, p(\beta_2)$$

With a vague prior for the parameter β_2

$$p(\beta_2) = N(0, \sigma^2) \quad \text{and} \quad \sigma = 100$$

Bayesian estimation

GEV$_2$ model

$$\pi_1(\beta_1, \beta_2, \beta_3, \alpha, \kappa) = J(\beta_1, \alpha, \kappa)\, p_{\beta_2}(\beta_2)\, p_{\beta_3}(\beta_3)$$

with

$$p_{\beta_2}(\beta_2) = N\left(0, \sigma_1^2\right) \qquad p_{\beta_3}(\beta_3) = N\left(0, \sigma_2^2\right)$$

and

$$\sigma_1 = \sigma_2 = 100$$

Generalised Maximum Likelihood Method

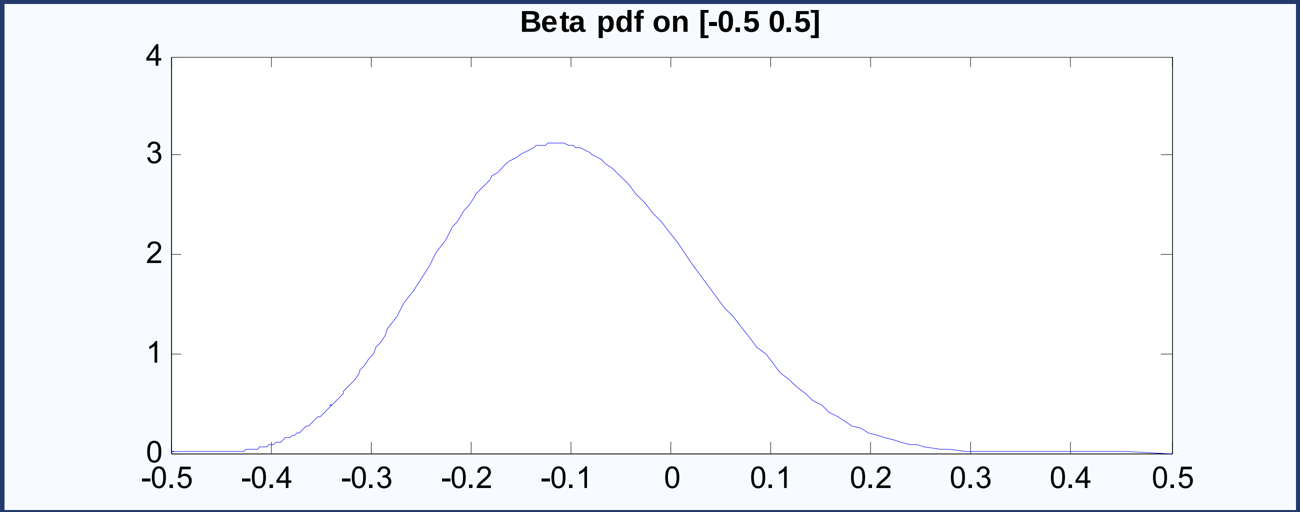
Stationary case

The GML method is based on the same principle than the ML method with an additional constraint on the shape parameter to restrain its domain.

Martins & Stedinger (2000) presented the GML approach for the case GEV0 with a Beta [-0.5;0.5] prior distribution for the shape parameter :

$$\pi_{\kappa}\left(\kappa\right) = Beta\left(u = 6, \mathrm{v} = 9\right)$$

Generalised Maximum Likelihood Method



prior distribution function of the shape parameter

$$E[\kappa] = -0.1 \quad \text{and} \quad Var[\kappa] = (0.122)^2$$

Generalised Maximum Likelihood Method

Non-stationary case

The GML method can be generalized to the non-stationary case:

1. Adopt the same prior distribution for the shape parameter,
2. Solve the equation system obtained by the ML method under this constraint.

The GML parameter estimators are the solution of the following optimisation problem :

$$\begin{cases} \max_{\theta} L_n\left(\underline{x};\theta\right) \\ \kappa \sim Beta\left(u, \mathrm{v}\right) \end{cases}$$

Generalised Maximum Likelihood Method

Non-stationary case

The solution of the optimisation problem is equivalent to the maximisation of the posterior distribution of the parameters conditionally to the data :

$$\pi\left(\theta \middle| \underline{x}\right) \propto L_n\left(\underline{x} \middle| \theta\right) \pi_\kappa\left(\kappa\right)$$

⇓

The GML estimator of the parameter vector is the mode of the posterior distribution.

Parameter and quantile estimation

ML : numerical solution : Newton-Raphson method.

GML & Bayes : Monte-Carlo Markov-Chain methods (MCMC)

The GML estimator corresponds to the mode of the posterior distribution,

The Bayesian estimator corresponds to the posterior mean.

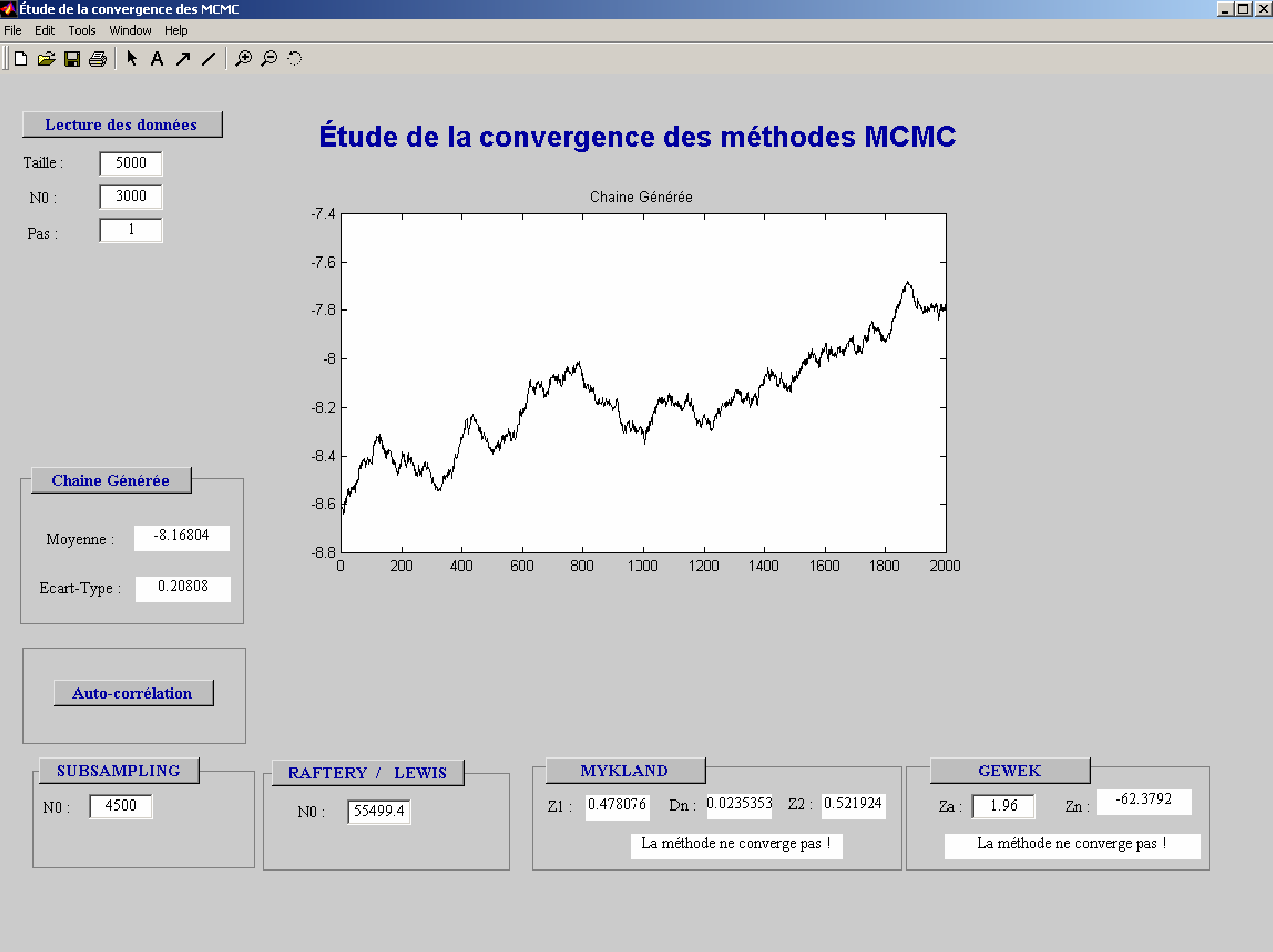
Parameter and quantile estimation

MCMC method adopted

For the GML and Bayesian method, the posterior distribution is simulated with the Metropolis-Hastings (M-H) algorithm (Gilks et al. 1996).

Chain size and burn-in period

Several techniques allow to check convergence of generated Markov Chain to the stationary distribution (El Adlouni et al. 2005). For all cases presented in this work, the convergence of the MCMC methods is obtained with a chain size of N=15000 and with a burn-in period of N0=8000.


Étude de la convergence des MCMC
File Edit Tools Window Help
Lecture des données
Taille : 5000
N0 : 3000
Pas : 1
Étude de la convergence des méthodes MCMC
Chaine Générée
-7.4
-7.6
-7.8
-8
-8.2
-8.4
-8.6
-8.8
0 200 400 600 800 1000 1200 1400 1600 1800 2000
Chaine Générée
Moyenne : -8.16804
Ecart-Type : 0.20808
Auto-corrélation
SUBSAMPLING
N0 : 4500
RAFTERY / LEWIS
N0 : 55499.4
MYKLAND
Z1 : 0.478076
Dn : 0.0235353
Z2 : 0.521924
La méthode ne converge pas !
GEWEK
Za : 1.96
Zn : -62.3792
La méthode ne converge pas !

Parameter and quantile estimation

Quantile estimation

Aside from parameter estimators, the MCMC algorithm iterations allow to obtain the conditional distribution of quantiles given an observed value y_0 of the covariate Y_t.

For each iteration of the MCMC algorithm $i=1,\ldots,N$ we compute the quantile $x^{(i)}_{p,y_0}$ corresponding to a non-exceedance probability p

$$x^{(i)}_{p,y_0} = \left(\mu_{y_0}{}^{(i)}\right) + \frac{\alpha^{(i)}}{\kappa^{(i)}}\left[1-\left(-\log\left(p\right)\right)^{\kappa^{(i)}}\right]$$

Conditional on the value y_0

Parameter and quantile estimation

Quantile estimation (cont.)

$\mu_{y_0}^{(i)}$ Is the location parameter conditional to a particular value y_0 of the covariate Y_t.

$\mu_{y_0}^{(i)} = \mu^{(i)}$ For the GEV0 model

$\mu_{y_0}^{(i)} = \beta_1^{(i)} + \beta_2^{(i)} y_0$ For the GEV1 model

$\mu_{y_0}^{(i)} = \beta_1^{(i)} + \beta_2^{(i)} y_0 + \beta_3^{(i)} {y_0}^2$ For the GEV2 model

Simulation based comparison

The three estimation methods are compared, for all three models, using Monte Carlo simulations.

The covariate Y_t represents time

$$Y_t = t$$

The following values of the shape parameter are considered:

$$\kappa = -0.1\ ,\ \kappa = -0.2 \quad \text{et} \quad \kappa = \text{-}0.3$$

Simulation based comparison

Methodology

Performance criteria are the bias and the RMSE of quantile estimates for different non-exceedance probabilities :

$$p = 0.5,\ 0.8,\ 0.9,\ 0.99 \text{ and } 0.999$$

Obtained for R=1000 samples of size $n=50$.

Simulation based comparison

Bias and RMSE of quantile estimates for the ML, GML and Bayesian approach and for model GEV0

GEV0		Bias			RMSE		
	p	ML	GML	Bayes	ML	GML	Bayes
$\kappa = -0.1$	0.5	0.02	0.01	0.01	0.35	0.17	0.25
	0.8	-0.03	0.05	0.02	0.44	0.33	0.47
	0.9	-0.05	0.04	0.08	0.45	0.45	0.63
	0.99	0.02	0.11	0.19	1.86	0.94	1.53
	0.999	0.71	0.22	0.38	6.01	1.60	1.96
$\kappa = -0.2$	0.5	-0.01	0.05	0.03	0.20	0.24	0.20
	0.8	-0.02	-0.03	0.05	0.35	0.33	0.42
	0.9	-0.01	-0.05	0.12	0.57	0.42	0.62
	0.99	0.57	-0.17	0.26	3.31	1.20	1.64
	0.999	1.72	-0.37	0.53	14.35	3.53	5.78
$\kappa = -0.3$	0.5	-0.04	-0.01	0.01	0.17	0.20	0.24
	0.8	-0.12	-0.04	0.08	0.35	0.39	0.49
	0.9	-0.16	-0.07	0.14	0.75	0.64	0.73
	0.99	0.19	-0.23	0.27	4.44	2.48	4.15
	0.999	1.96	-0.42	0.48	21.08	7.83	9.06

Simulation based comparison

Bias and RMSE of quantile estimates for the ML, GML and Bayesian approach and for model GEV1

GEV1		Bias			RMSE		
	p	ML	GML	Bayes	ML	GML	Bayes
$\kappa = -0.1$	0.5	0.06	0.01	0.04	0.41	0.39	0.32
	0.8	0.04	0.03	0.03	0.47	0.50	0.41
	0.9	-0.02	0.03	0.08	0.56	0.56	0.50
	0.99	-0.17	0.05	0.13	1.58	0.85	0.91
	0.999	-0.14	0.12	0.39	4.17	1.36	1.45
$\kappa = -0.2$	0.5	0.01	0.02	0.08	0.45	0.30	0.39
	0.8	0.02	0.05	0.07	0.53	0.51	0.53
	0.9	0.03	0.06	0.05	0.73	0.73	0.68
	0.99	0.26	-0.11	0.18	3.04	2.08	3.41
	0.999	1.74	-0.17	0.36	11.33	5.24	7.65
$\kappa = -0.3$	0.5	0.07	0.03	0.04	0.56	0.36	0.37
	0.8	0.04	0.04	0.08	0.66	0.59	0.57
	0.9	-0.03	0.04	0.19	0.88	0.83	0.79
	0.99	-0.64	-0.12	0.36	4.10	2.44	2.62
	0.999	-1.34	-0.61	0.82	17.95	7.06	8.89

Simulation based comparison

Bias and RMSE of quantile estimates for the ML, GML and Bayesian approach and for model GEV2

GEV2		Bias			RMSE		
	p	ML	GML	Bayes	ML	GML	Bayes
$\kappa = -0.1$	0.5	**-0.06**	**0.01**	**0.04**	**0.98**	**0.56**	**0.53**
	0.8	**-0.71**	**0.04**	**0.09**	**1.02**	**0.66**	**0.66**
	0.9	**-0.76**	**0.06**	**0.13**	**1.09**	**0.77**	**0.78**
	0.99	**-0.99**	**0.13**	**0.27**	**1.96**	**1.30**	**1.64**
	0.999	**-1.12**	**0.18**	**0.35**	**4.63**	**2.12**	**3.26**
$\kappa = -0.2$	0.5	**-0.81**	**0.10**	**0.12**	**1.22**	**0.64**	**0.72**
	0.8	**-0.79**	**0.16**	**0.17**	**1.23**	**0.80**	**0.86**
	0.9	**-0.80**	**0.21**	**0.34**	**1.30**	**0.97**	**1.07**
	0.99	**-0.93**	**0.43**	**0.63**	**2.83**	**1.94**	**2.93**
	0.999	**-1.79**	**0.82**	**0.95**	**8.93**	**3.73**	**5.56**
$\kappa = -0.3$	0.5	**-0.83**	**0.27**	**0.71**	**1.47**	**0.97**	**1.22**
	0.8	**-0.92**	**0.48**	**0.88**	**1.46**	**1.38**	**1.54**
	0.9	**-0.81**	**0.61**	**0.87**	**1.87**	**1.86**	**1.92**
	0.99	**-1.57**	**1.42**	**1.83**	**3.87**	**3.48**	**3.63**
	0.999	**-3.69**	**2.87**	**4.14**	**12.62**	**8.54**	**9.26**

Simulation based comparison

Results

- For the ML method :

1. GEV0 & GEV1 models: large RMSE is caused by the high variance.
2. GEV2 model: reduction of the variance for extreme quantiles.

- Bayesian approach: positive Bias and low RMSE for all quantiles.

Simulation based comparison

Results

- GML method

  1. Best bias performance for all three models.
  2. Generally low RMSE.
  3. Negative Bias for high skewness since the prior is centered on -0.1

Conclusion

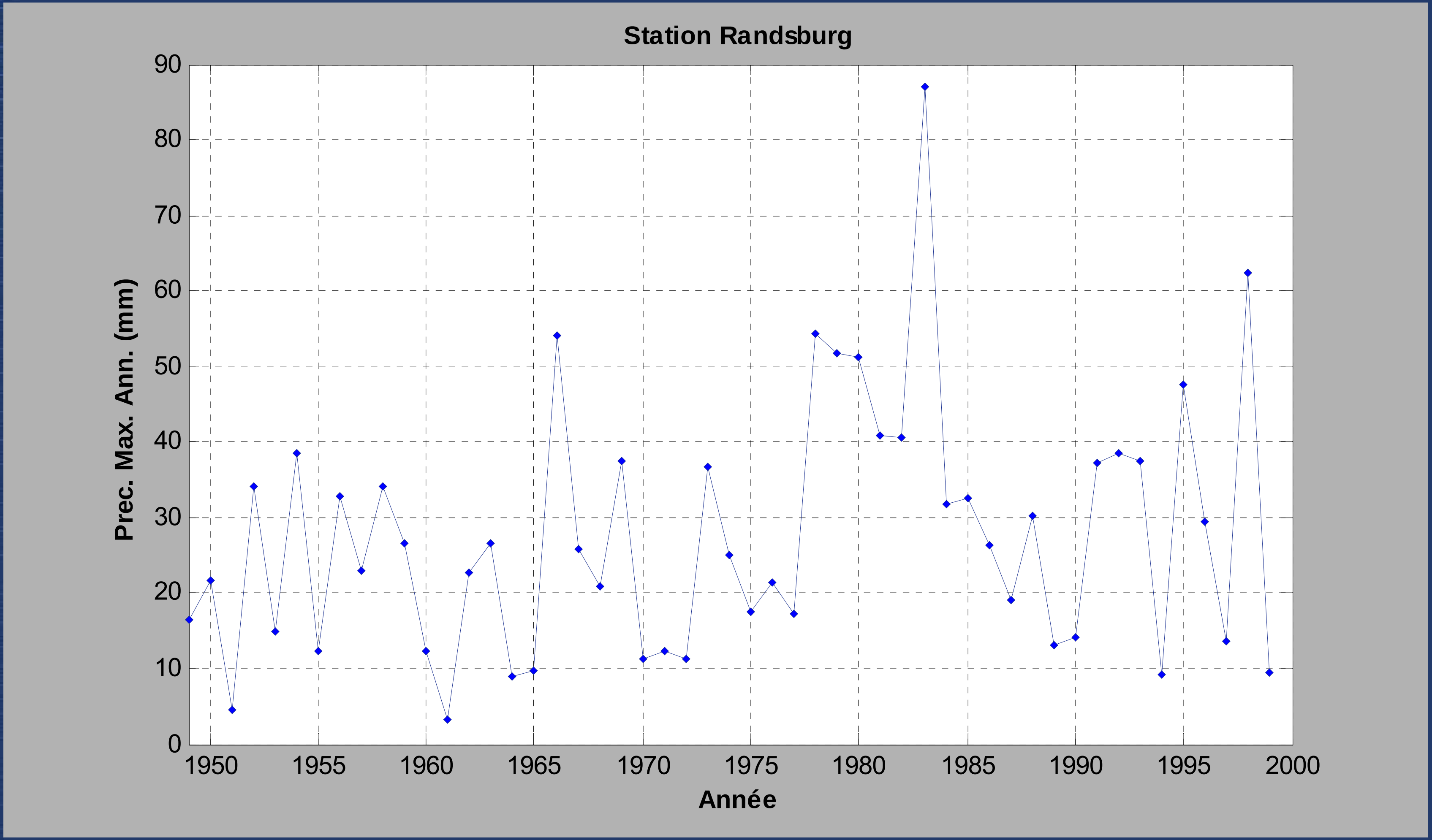
The GML method leads to best results for all three models.

Case study

Location of the Randsburg station in California

Code : Station 047253
Name : Randsburg
Latitude : 35.3700
Longitude : -117.650
Period : 1949-1999
Size of the series : n=51

Case study



Annual Max rainfall series at Randsburg

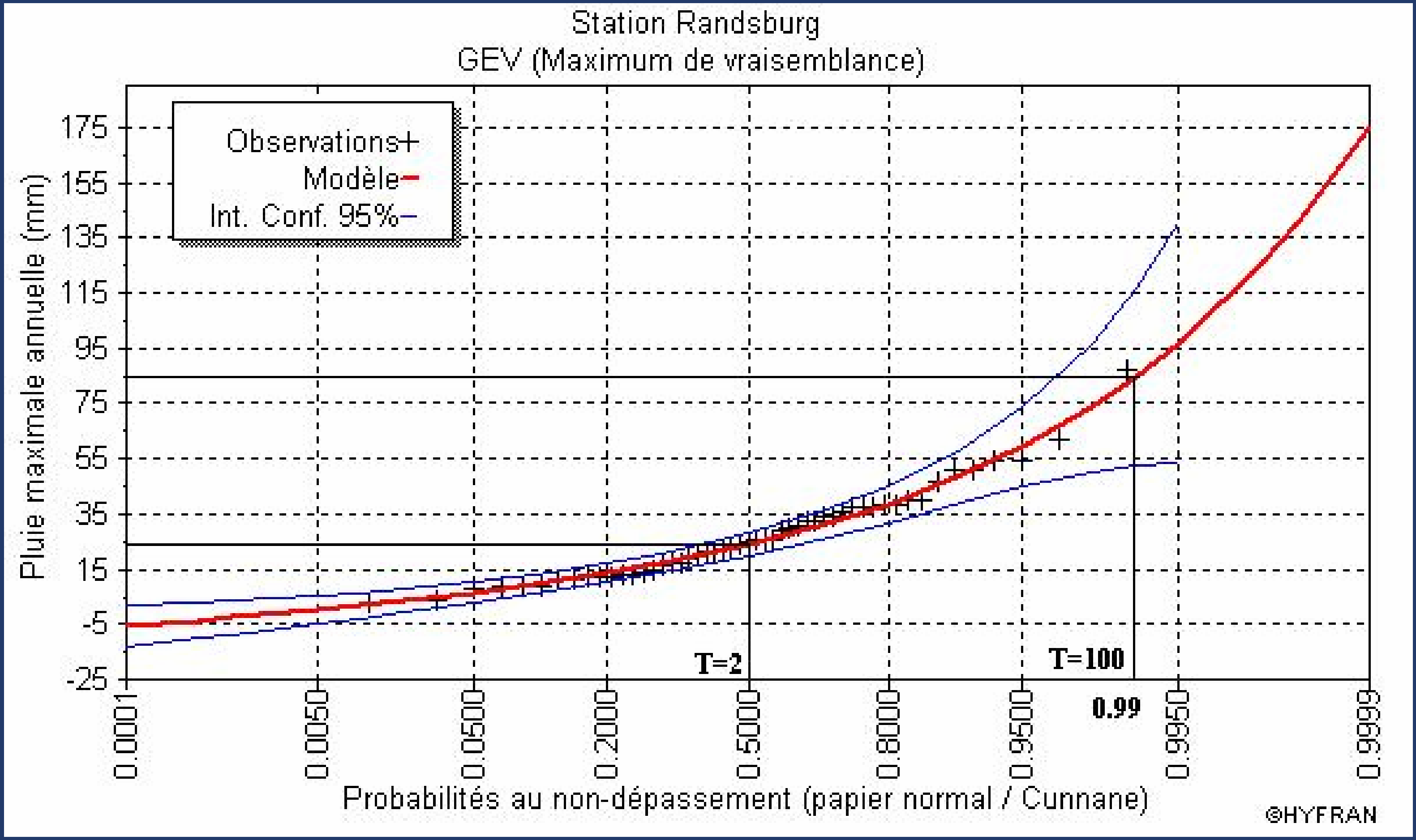
Case study: characteristics

Basic statistics	Max. Ann. Prec. at Randsburg
Number of observations	51
Minimum	3.00
Maximum	87.0
Mean	27.2
St. dev.	16.7
Median	25.0
Coefficient of variation (Cv)	0.612
Coefficient of skewness (Cs)	1.14
Coefficient of kurtosis (Ck)	4.52

Parameter estimators for the GEV (ML)

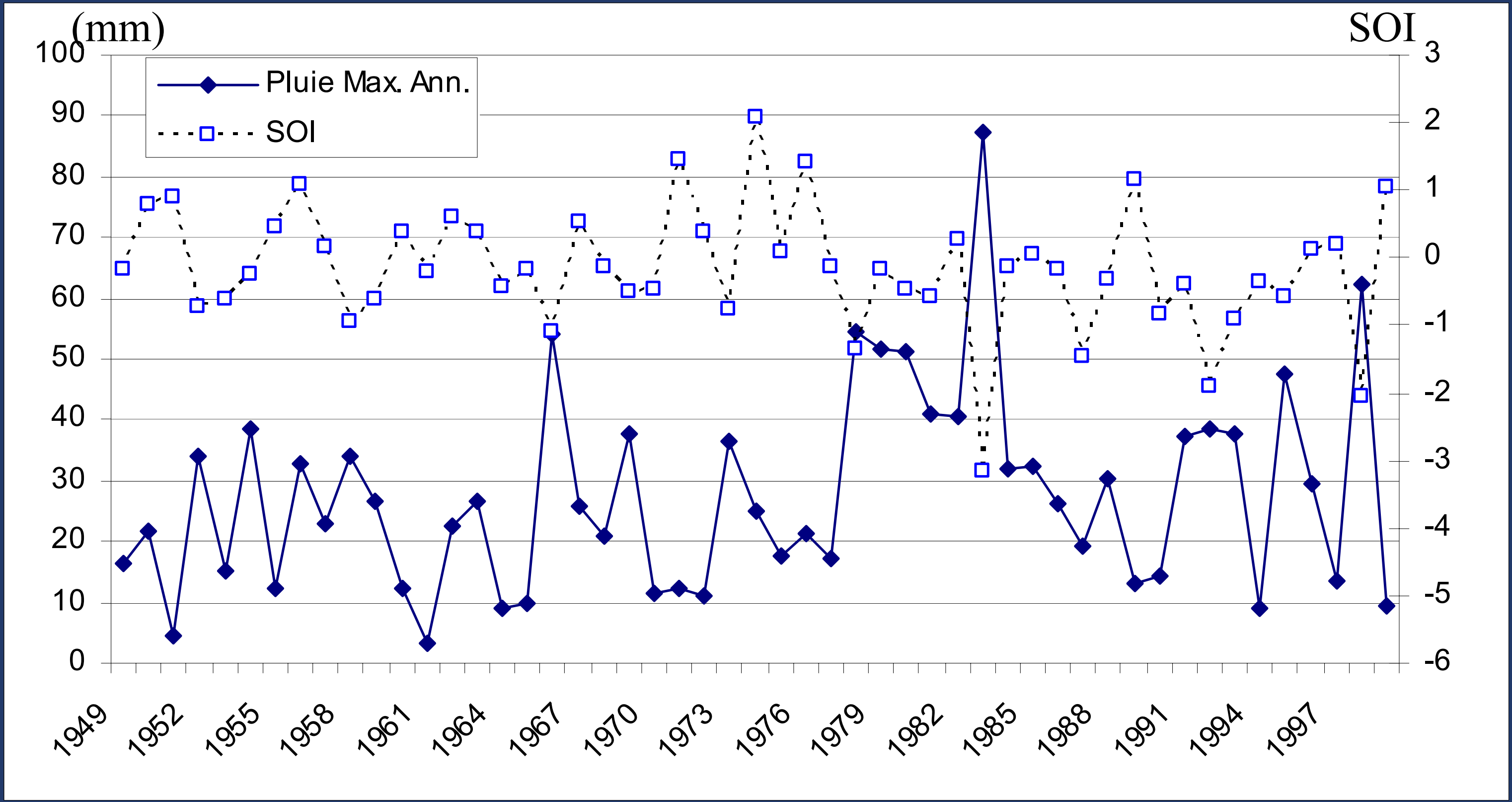
$$\mu = 19.35 \quad \alpha = 12.10 \quad \kappa = -0.07$$

Case study: ML fitting



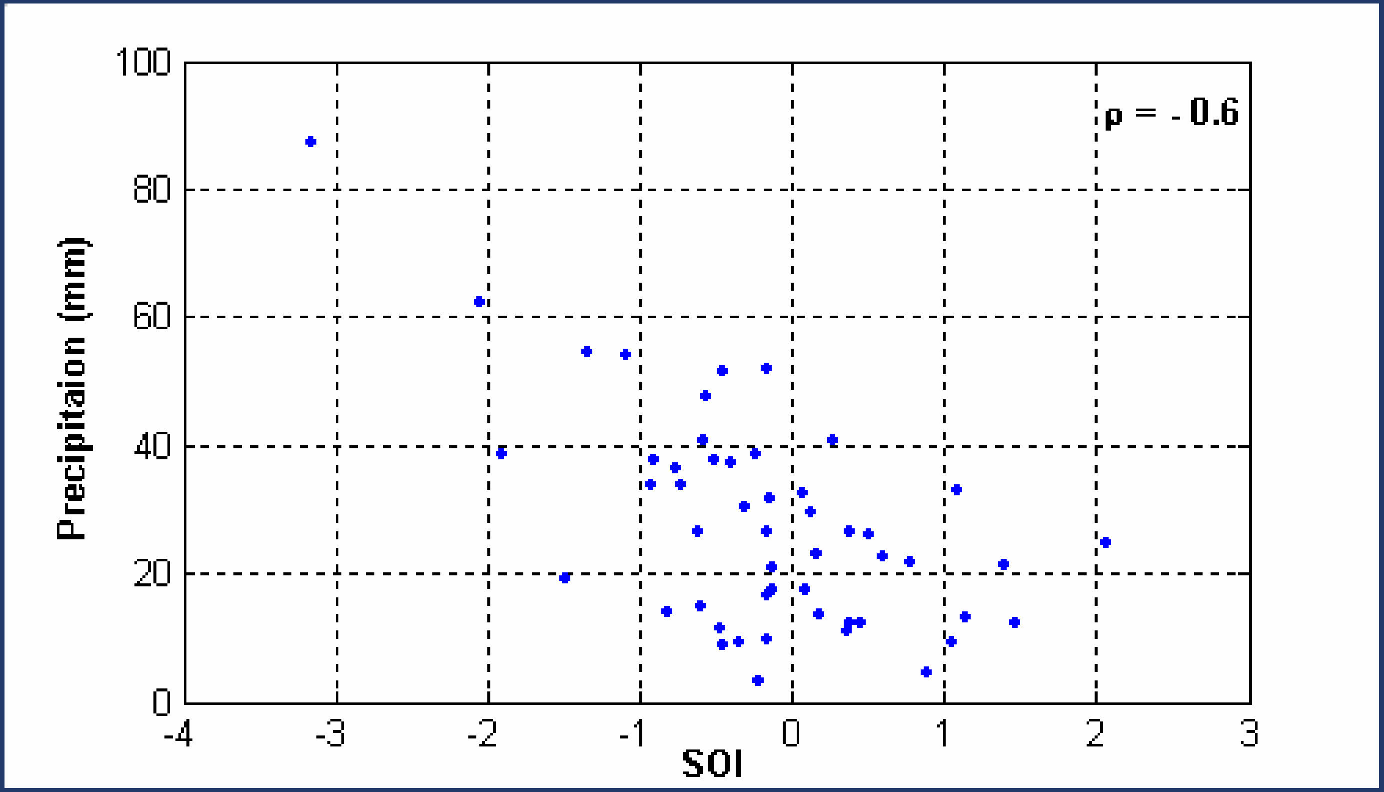
Fitting of the GEV model / Classic ML

Case study : Correlation with SOI



Annual Max rainfall series and SOI index at Randsburg

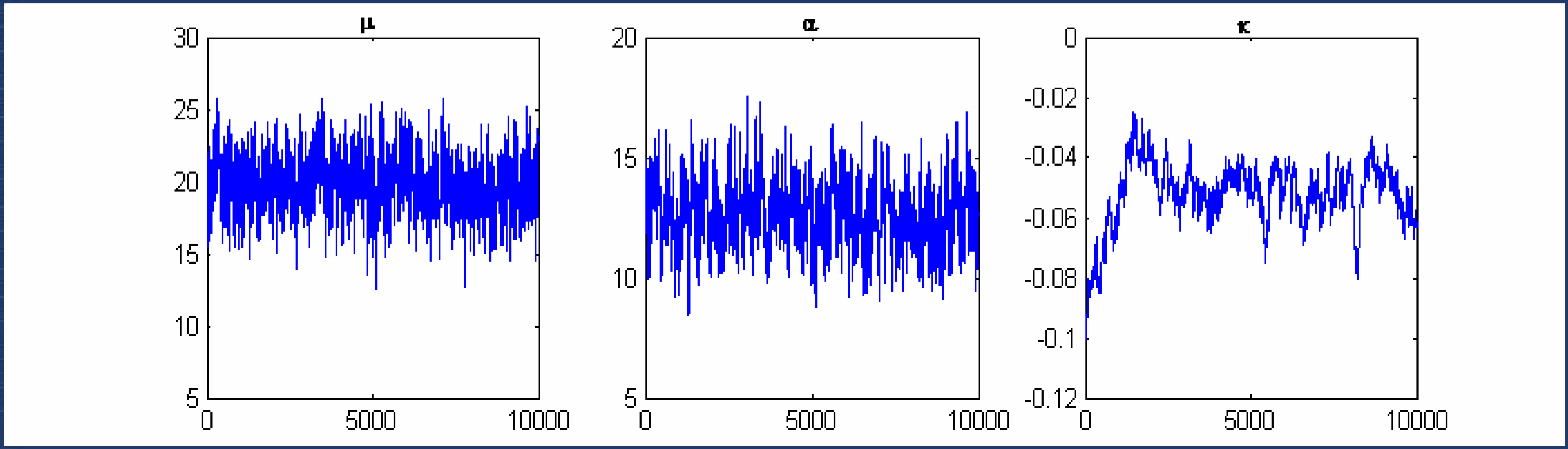
Case study : Correlation with SOI



Observed Annual Max rainfall and corresponding SOI value

Case study: linear dependence model

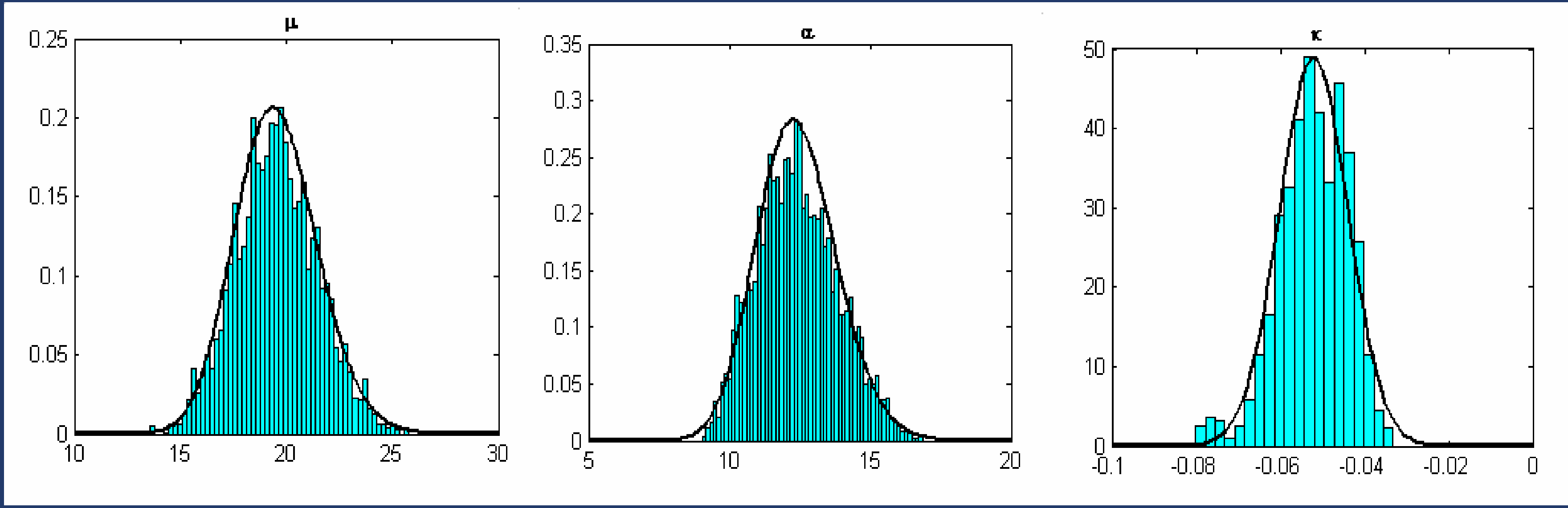
GML method: Outputs of the MCMC algorithm



MCMC algorithm iterations for the estimation of the GEV0 model parameters with the GML method.

Case study: linear dependence model

GML method: Outputs of the MCMC algorithm



Histogram of the GEV0 parameters posterior distribution, obtained with the last N-N0 iterations of the MCMC algorithm.

The GML estimator corresponds to the mode of the posterior distribution for each parameter.

Case study : model comparison

Maximized log-likelihood function and GML parameter estimators for each model

	l_n^*	β_1	β_2	β_3	α	κ
GEV_0	-209.96	19.52	-	-	12.36	-0.05
GEV_1	-206.86	18.89	-9.92	-	12.21	-0.07
GEV_2	-204.71	16.57	-10.61	3.03	12.14	-0.06

Case study : model comparison

A simple approach to compare the validity of two models M_1 and M_0 such that $M_0 \subset M_1$ is to use the deviance statistic defined as:

$$D = 2\left\{ l_n^*\left(M_1\right) - l_n^*\left(M_0\right) \right\}$$

l_n^* is the maximized log-likelihood function for each model

The statistic is Chi-squared distributed. The parameter of the Chi-deux distribution is the difference of the number of parameters of the two models M_1 et M_0.

Case study : model comparison

Comparison of models GEV0 and GEV1 indicates that the difference is significant, since the statistic value is $D = 6.2$

D is considered significant at the confidance level 95%

$$\Pr\left(\chi_1^2 \leq 6.2\right) = 0.9872$$

The GEV1 model provides a better representation of data variability than model GEV0.

Case study : model comparison

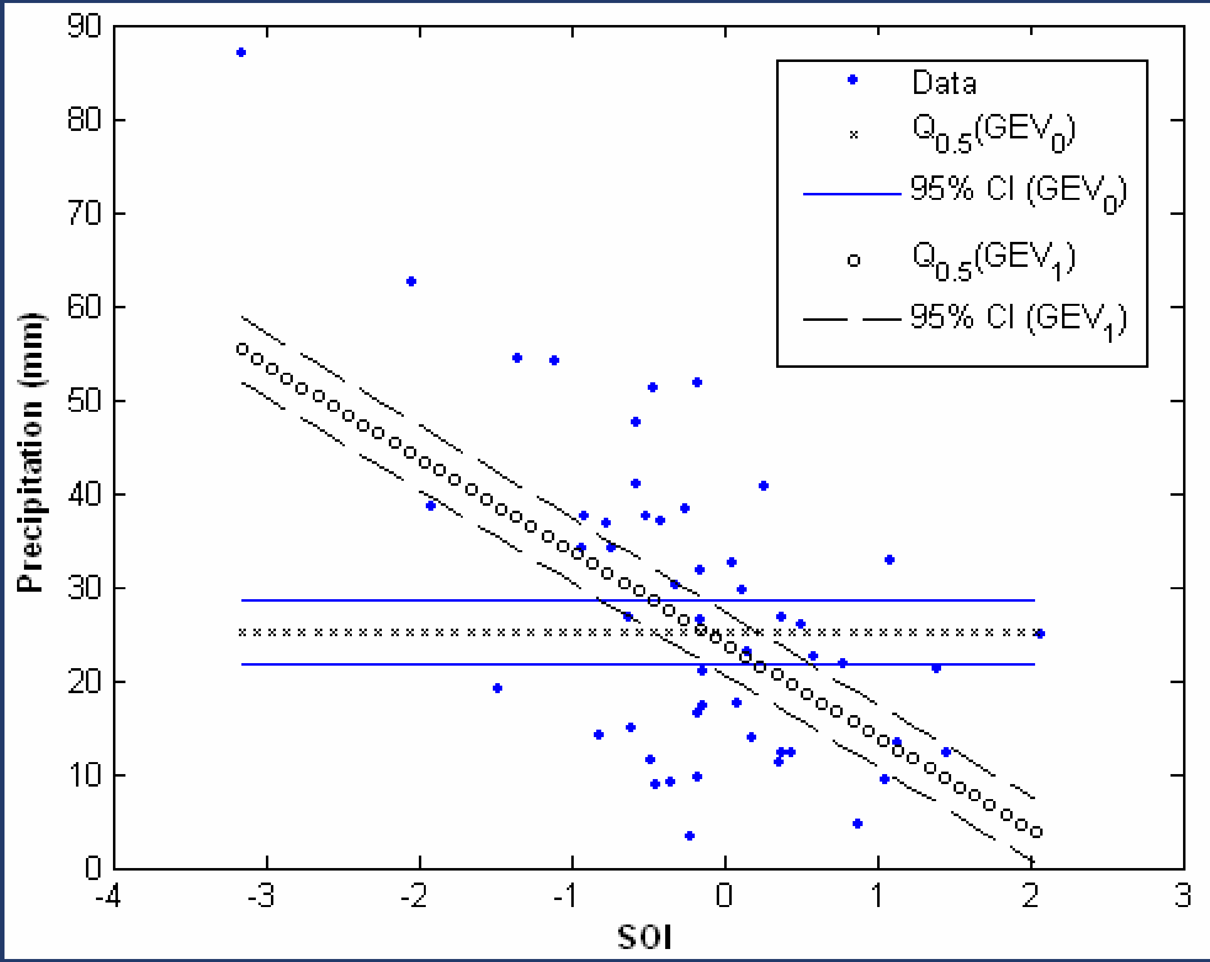
Comparison of models GEV1 and GEV2 leads to a statistic value of $D = 4.3$

with $\Pr\left(\chi_1^2 \leq 4.3\right) = 0.9619$

The difference between the two models is hence significant at the 95% level.

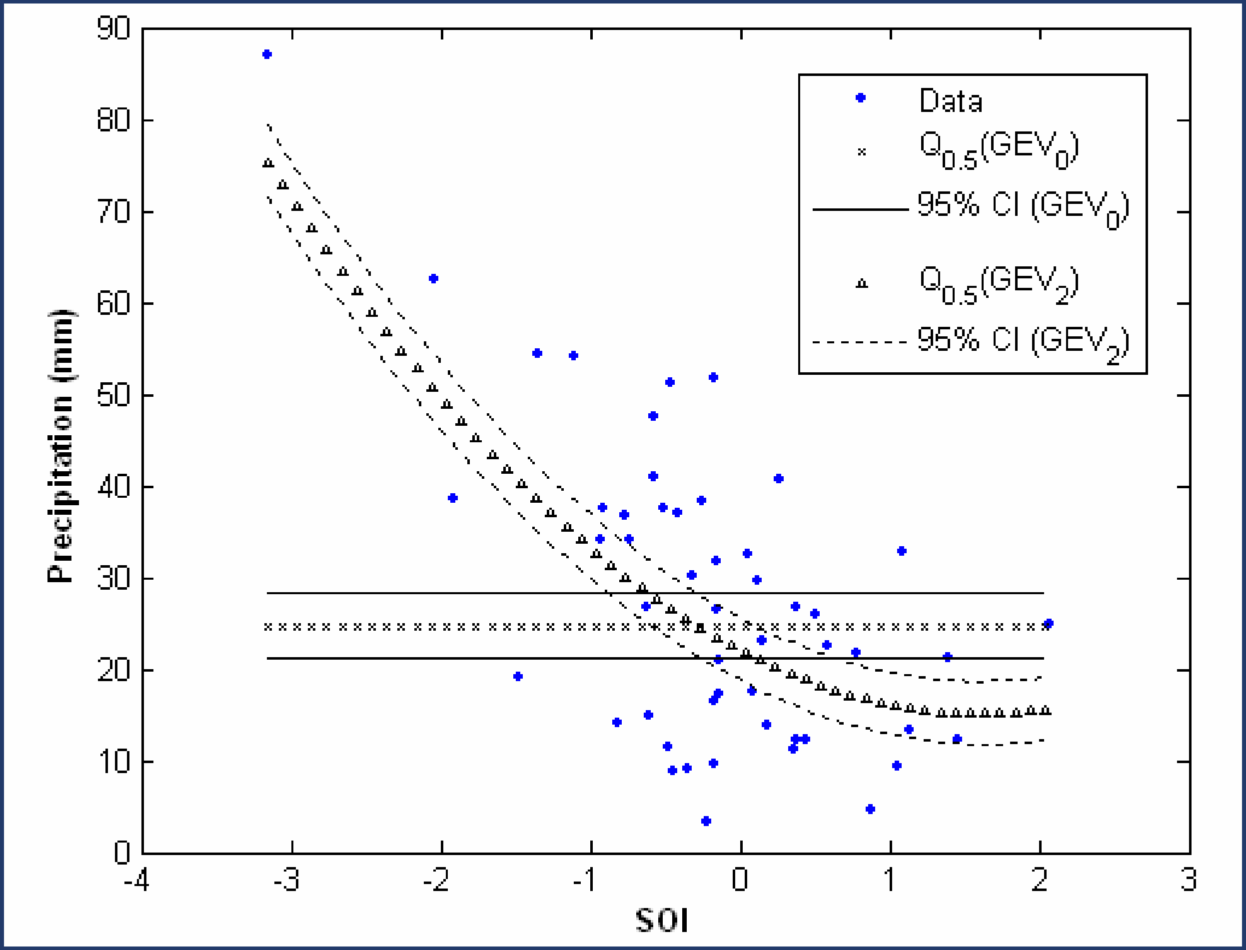
The GEV2 model is the most adequate to represent the dependence between the location parameter of the GEV and the SOI index.

Case study : Correlation with SOI



GEV0 and GEV1 Median estimators conditional to SOI values.

Case study : Correlation with SOI



GEV0 AND GEV2 Median estimators conditional to SOI values.

Case study : Correlation with SOI

GEV0, GEV1 and GEV2 Median estimators conditional to particular SOI values.

	SOI = - 3.16	SOI = 0.04	SOI = 2.04
GEV_0	24 (21-28)	24 (21-28)	24 (21-28)
GEV_1	54 (51-58)	23 (19-27)	4 (0.5-7)
GEV_2	77 (72-82)	21 (18-24)	17 (15-22)

Case study : Correlation with SOI

Results

- The difference is more important for lower SOI values which correspond to higher extremes of annual max. precipitations.
- GEV2 median estimate can be 3 times higher than GEV0 estimate.
- The linear trend model (GEV1) under-estimates the median for the higher SOI values.

Conclusions

✓ Advantages of the non-stationary GEV model in describing the variance.

✓ No assumption of normality as in classical models.

✓ Generalisation of the GML model to the non-stationary case provides efficient estimators for real hydro-meteorological data series.

✓ Use of MCMC methods allows a robust solution to the likelihood equation system and leads to <u>estimates of credibility intervals for parameters and quantiles.</u>

References

Coles G. S. (2001). An Introduction to Statistical Modeling of Extreme Values, Springer, 208 p.

El Adlouni, S., Favre, A-C. and Bobée, B. (2005). Comparison of methodologies to assess the convergence of Markov Chain Monte-Carlo Methods. Computational Statistics and Data Analysis (Under Press).

Hastings, W. (1970). Monte Carlo sampling methods using Markov Chains and their applications. Biometrika,57. 97-109.

Martins, E. S., J. R. Stedinger (2000) . Generalized Maximum Likelihood GEV Quantile Estimators for Hydrologic Data. Water Resources Research , vol.36., no.(3)., pp.737-744.

Scarf, P.A. (1992). Estimation for a four parameter generalized extreme value distribution, Comm. Stat. Theor. Meth., 21, 2185-2201.

Smith, R. L. (1985). Maximum likelihood estimation in a class of non-regular cases. Biometrika 72, 67-92.