How to choose threshold in a POT model?

Martin Schindler, Jan Picek, Jan Kyselý

e-mail: martin.schindler@tul.cz

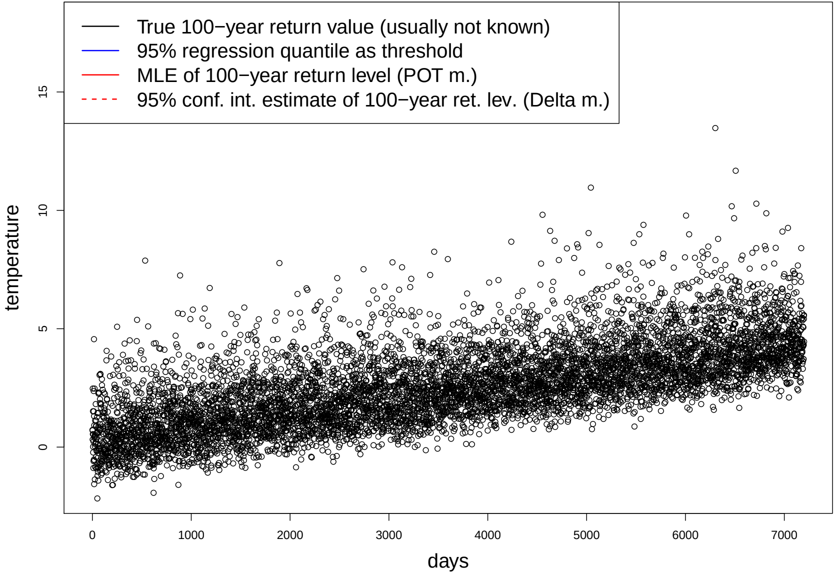
TECHNICAL UNIVERSITY OF LIBEREC

Robust, 8-14 September 2012

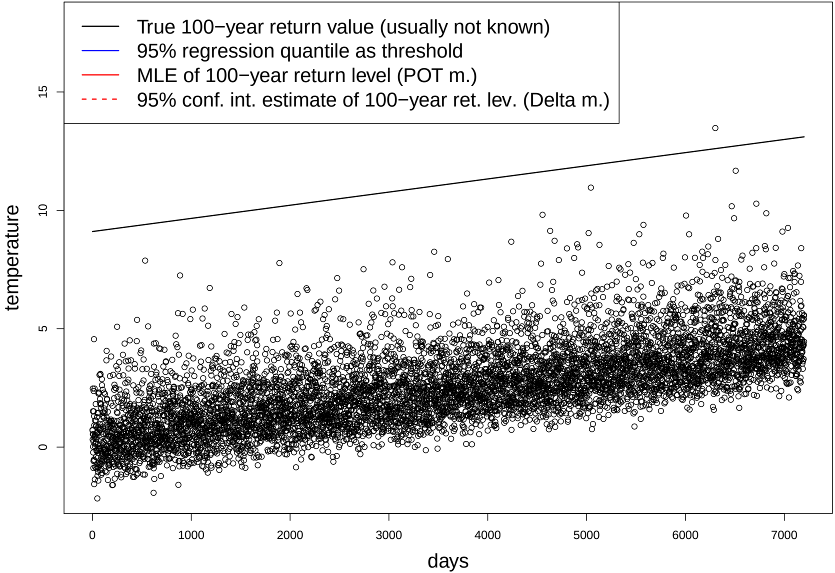
1 Introduction

Outline

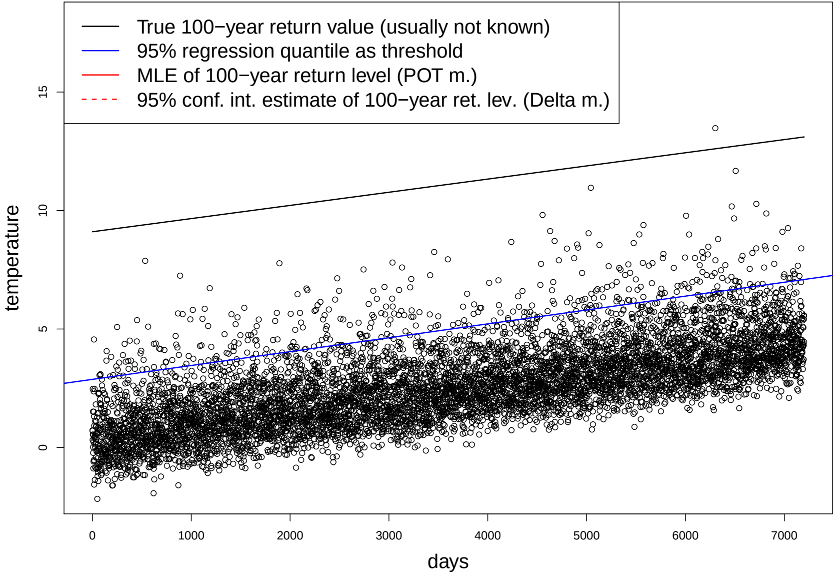
Max. daily temperatures with linear trend. Sample size $7200 = 80$ years $\times$ 90



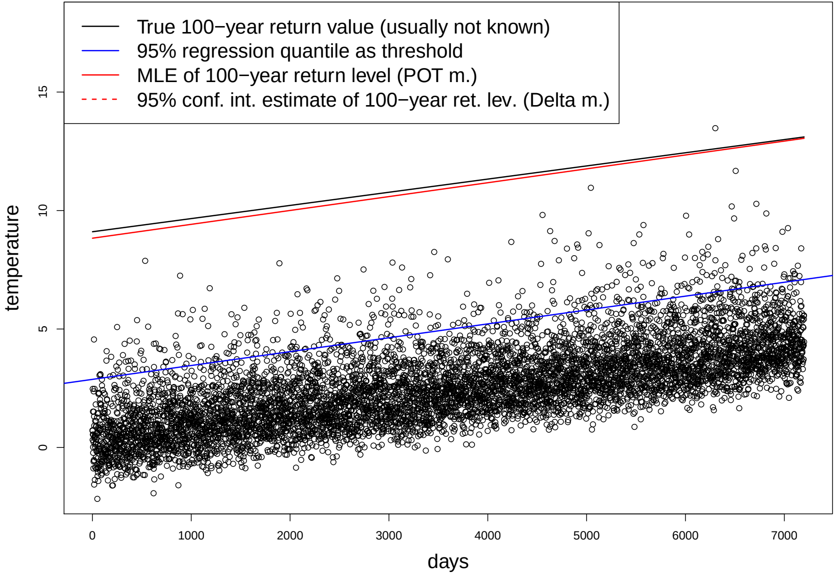
Max. daily temperatures with linear trend. Sample size $7200 = 80$ years $\times$ 90



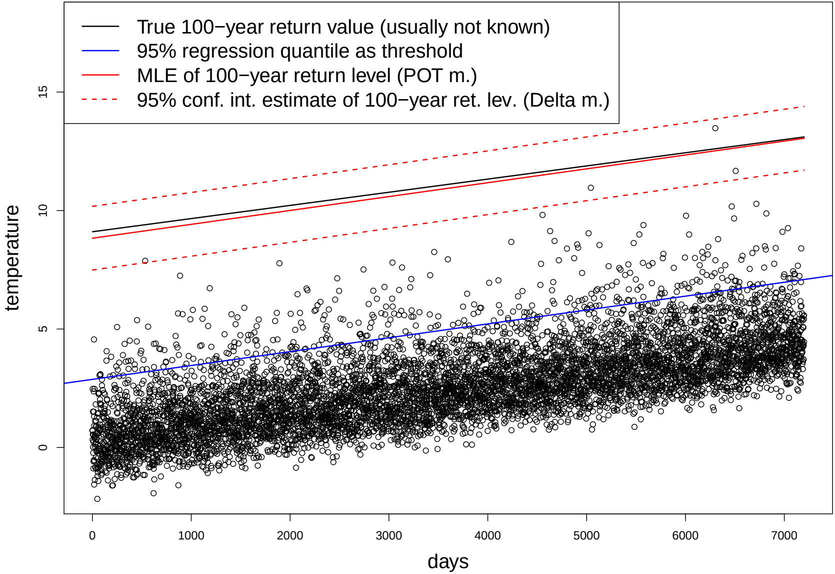
Max. daily temperatures with linear trend. Sample size $7200 = 80$ years $\times$ 90



Max. daily temperatures with linear trend. Sample size $7200 = 80$ years $\times$ 90



Max. daily temperatures with linear trend. Sample size $7200 = 80$ years $\times$ 90



Peak over threshold (POT) method

Data: daily maximum temperature, daily rainfall values.
We wish to estimate m-year return ”value” (quantile of a distribution) using POT method.

Peak over threshold (POT) method

Data: daily maximum temperature, daily rainfall values.
We wish to estimate m-year return ”value” (quantile of a distribution) using POT method.

Situation (assumptions):

- data $y_1, \ldots, y_n$ follow a model $y_i = e_i + \beta_0 + \beta_1 \cdot i, \qquad i = 1, \ldots, n,$ where e_i is a random sample with distribution function $F(x)$.
- m-year return ”level” is

$$F^{-1}\left(1 - \frac{1}{m}\right) + \beta_1 \cdot i, \qquad 1 \leq i \leq n$$

which is the population $(1 - \frac{1}{m})$-th regression quantile line.

Solution: natural choice for the threshold is an α-th regression quantile

$$(\hat{\beta}_0(\alpha), \hat{\beta}_1(\alpha))^T = \arg\min_{\mathbf{t}\in\mathbf{R}^2} \sum_{i=1}^{n} \rho_\alpha(y_i - t_0 - t_1 \cdot i), \text{ where}$$

$$\rho_\alpha(u) = |u|\{(1-\alpha)I[u<0] + \alpha I[u>0]\}, \quad u \in \mathbf{R},\ \alpha \in (0,1)$$

Solution: natural choice for the threshold is an α-th regression quantile

$$(\hat{\beta}_0(\alpha), \hat{\beta}_1(\alpha))^T = \arg\min_{\mathbf{t}\in\mathbf{R}^2} \sum_{i=1}^{n} \rho_\alpha(y_i - t_0 - t_1 \cdot i), \text{ where}$$

$$\rho_\alpha(u) = |u|\{(1-\alpha)I[u<0] + \alpha I[u>0]\}, \quad u \in \mathbf{R},\ \alpha \in (0,1)$$

- $\hat{\beta}_0(\alpha)$ estimates $\beta_0 + F^{-1}(\alpha)$
- $\hat{\beta}_1(\alpha)$ estimates β_1

Solution: natural choice for the threshold is an α-th regression quantile

$$(\hat{\beta}_0(\alpha), \hat{\beta}_1(\alpha))^T = \arg\min_{\mathbf{t} \in \mathbf{R}^2} \sum_{i=1}^{n} \rho_\alpha(y_i - t_0 - t_1 \cdot i), \text{ where}$$

$$\rho_\alpha(u) = |u|\{(1-\alpha)I[u < 0] + \alpha I[u > 0]\}, \quad u \in \mathbf{R}, \ \alpha \in (0,1)$$

- $\hat{\beta}_0(\alpha)$ estimates $\beta_0 + F^{-1}(\alpha)$
- $\hat{\beta}_1(\alpha)$ estimates β_1

Goal: Find the α such that the α-th regression quantile line $\hat{\beta}_0(\alpha) + \hat{\beta}_1(\alpha) \cdot i$ is the "optimal" threshold for the POT method.
Criterion: Maximization of the probability that a confidence belt covers the real return "level".
Methods: MLE is used to estimate the GPD parameters, and Delta method to construct the confidence belts for return "levels".

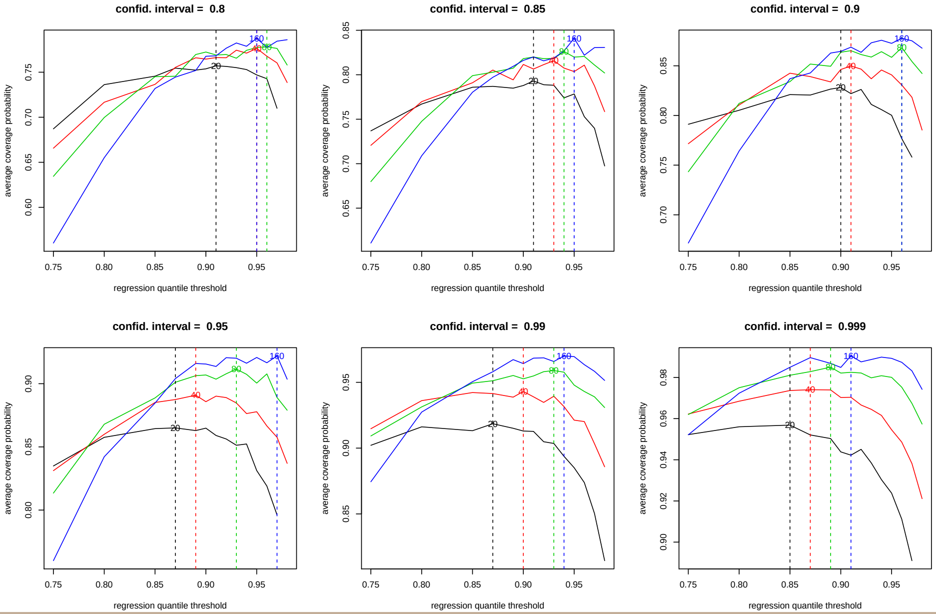
Outline

Monte Carlo simulations

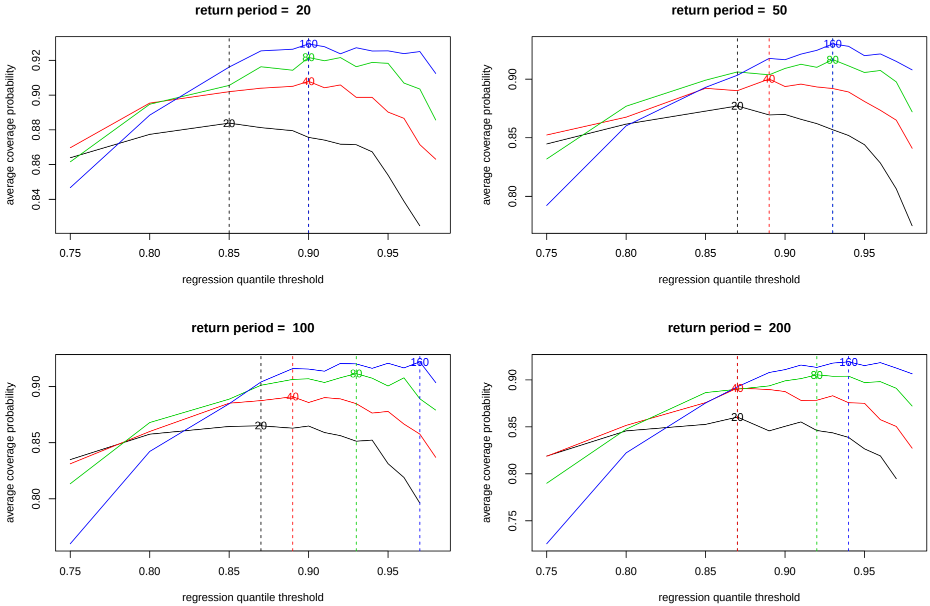
Parameters of simulated data:

- Choice of $F(x)$: standard Gumbel distribution
- Sample size $n = (20, 40, 80, 160)$-years $\cdot\ 90$
- We set trend $\beta_1 = 0.05/90$
- We wish to estimate $(20, 50, 100, 200)$-year return ”levels”
- We use $(75, 80, 85, 87, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98)\%$ regression quantile as threshold
- We compute $(80, 85, 90, 95, 99, 99.9)\%$ confidence belts to estimate the return ”levels”
- For every setting of the parameters we generate 6600 sets of data.

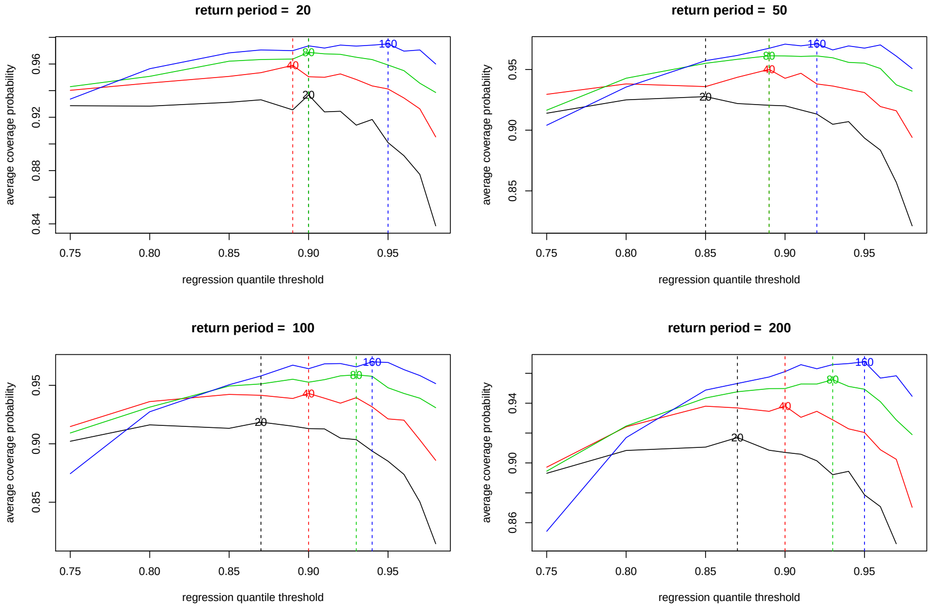
Average coverage probability for different sample size. Return "level": 100 years



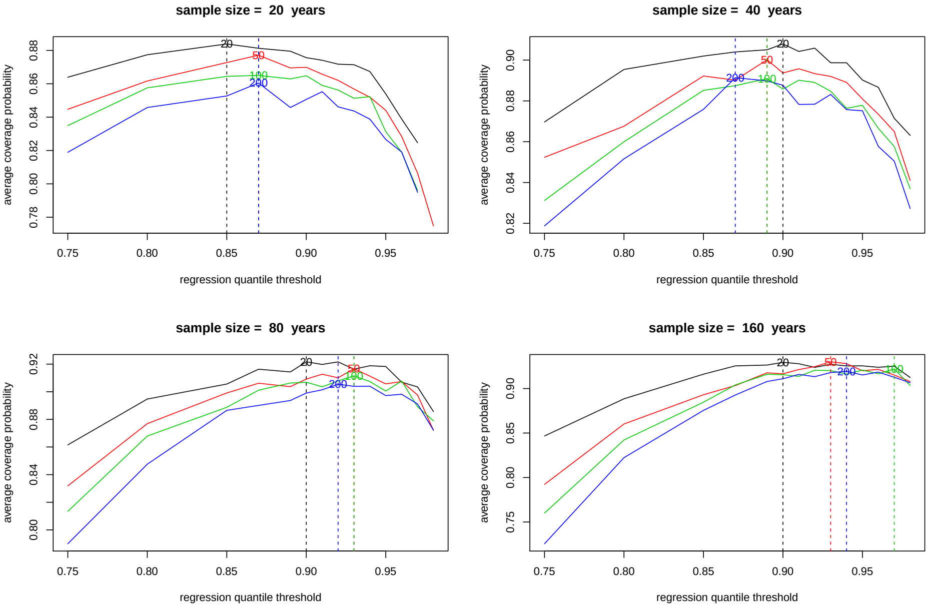
Average coverage probability for different sample size. Confidence interval: 95%



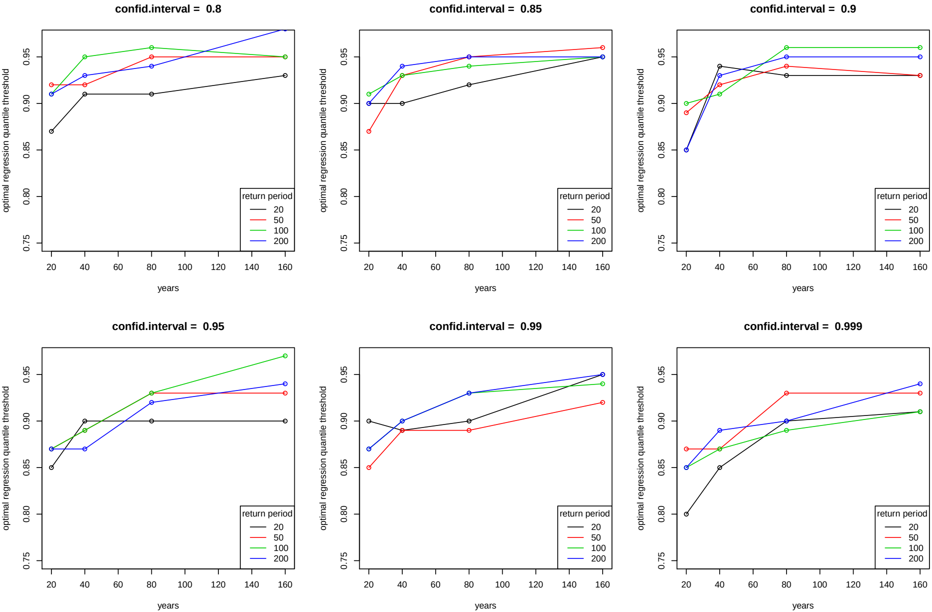
Average coverage probability for different sample size. Confidence interval: 99%



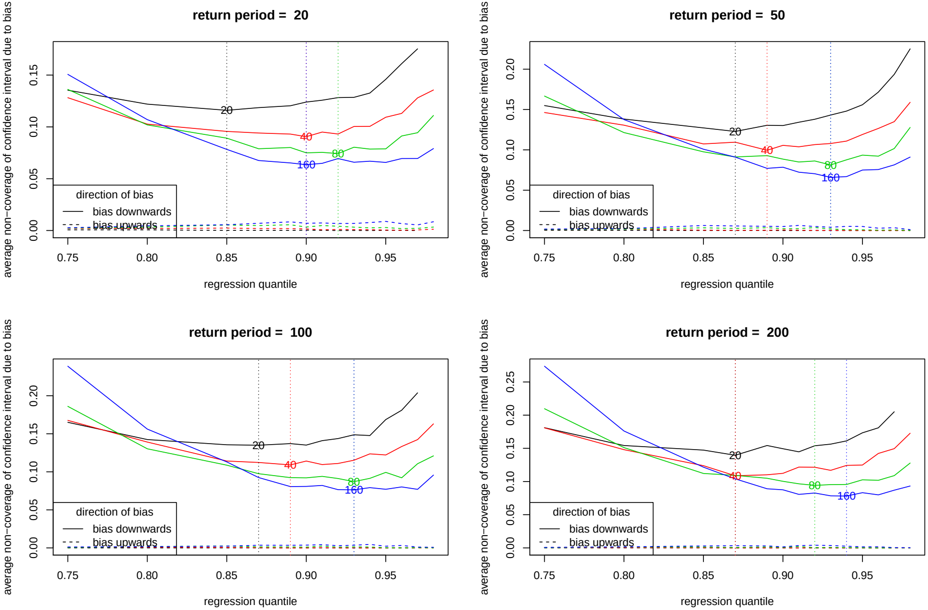
Average coverage probability for different return periods. Confidence interval: 95%



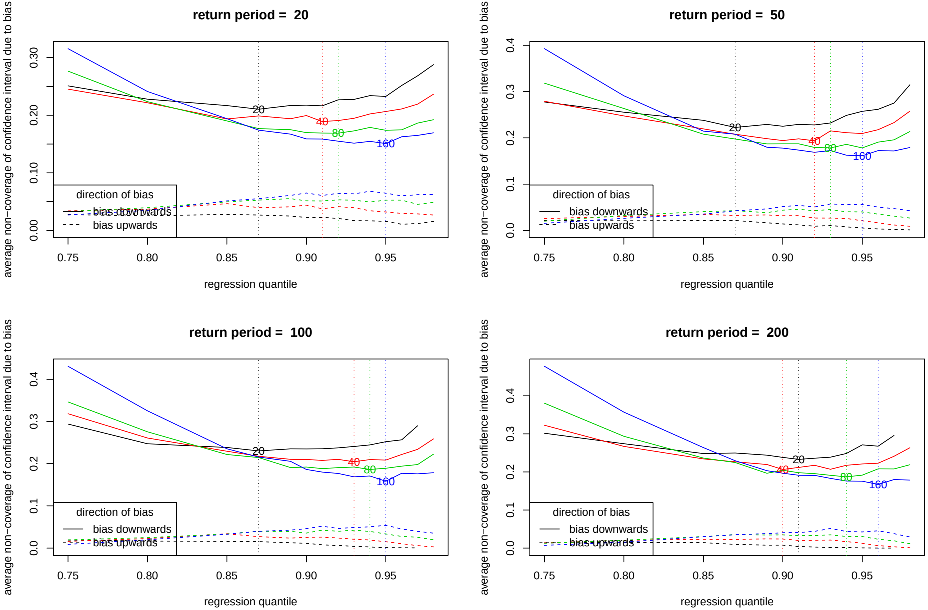
Optimal regression quantile wrt. sample size



Average non-coverage prob. for different sample sizes. Confidence interval: 95%



Average non-coverage prob. for different sample sizes. Confidence interval: 80%



Summary of the results

- Usually the optimal threshold is lower than 95% regression quantile

Summary of the results

- Usually the optimal threshold is lower than 95% regression quantile
  - Optimal: approx. 90% regression quantile, unless low confidence int. and very large sample size
  - Sample size $\searrow$ $\Longrightarrow$ Optimal regression quantile $\searrow$
  - Confidence int. $\nearrow$ $\Longrightarrow$ Optimal regression quantile $\searrow$

Summary of the results

- Usually the optimal threshold is lower than 95% regression quantile
  - Optimal: approx. 90% regression quantile, unless low confidence int. and very large sample size
  - Sample size $\searrow \quad \Longrightarrow \quad$ Optimal regression quantile $\searrow$
  - Confidence int. $\nearrow \quad \Longrightarrow \quad$ Optimal regression quantile $\searrow$

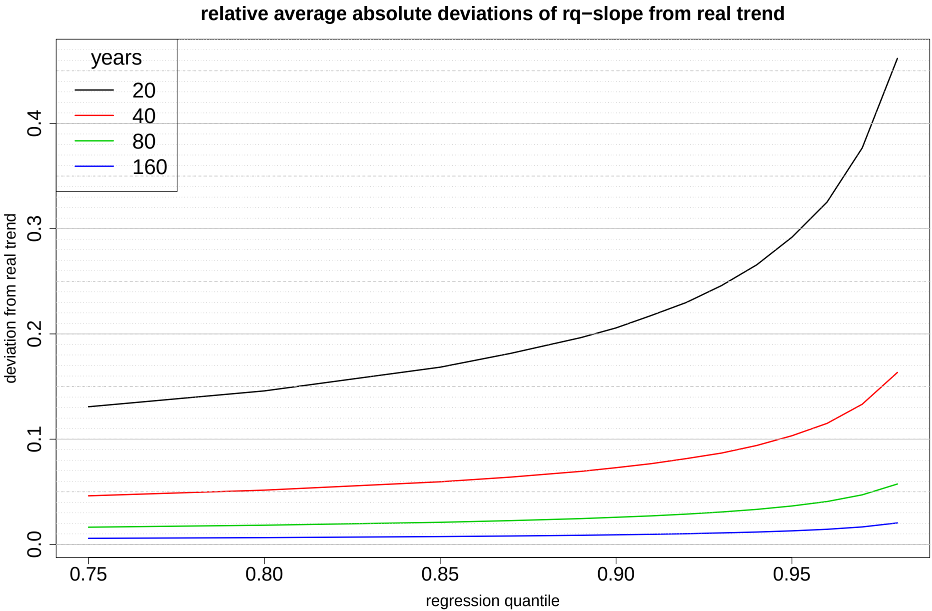
- Real coverage probability of confid. int. is much lower than expected.

Summary of the results

- Usually the optimal threshold is lower than 95% regression quantile
  - Optimal: approx. 90% regression quantile, unless low confidence int. and very large sample size
  - Sample size $\searrow$ $\Longrightarrow$ Optimal regression quantile $\searrow$
  - Confidence int. $\nearrow$ $\Longrightarrow$ Optimal regression quantile $\searrow$
- Real coverage probability of confid. int. is much lower than expected.
  - For real 95% confidence we need to construct approx. 99% confid. int.
  - Sample size $\searrow$ $\Longrightarrow$ Real confidence $\searrow$
  - Return period $\nearrow$ $\Longrightarrow$ Real confidence $\searrow$
  - Confidence interval is biased downwards, i.e. return level is too often higher than the confidence interval indicate.

Outline

How deviations of rq slope from real trend effect the optimal choice? IN NO WAY!



Noncoverage induced by variance of reg. quant. is constant

- For our model we can show that the α-th regression quantile $\hat{\beta}_1(\alpha)$ is approximately normal $\mathcal{N}\left(\mu = \beta_1, \sigma^2 = \frac{\alpha(1-\alpha)}{f^2(F^{-1}(\alpha))}\frac{12}{n(n^2-1)}\right)$
- Width of a confidence int. (belt) is proportional to $1/\sqrt{(1-\alpha)n}$
- For normal variable, mean absolute deviation from mean is $\sigma\sqrt{2/\pi}$

Noncoverage induced by variance of reg. quant. is constant

- For our model we can show that the α-th regression quantile $\hat{\beta}_1(\alpha)$ is approximately normal $\mathcal{N}\left(\mu = \beta_1, \sigma^2 = \frac{\alpha(1-\alpha)}{f^2(F^{-1}(\alpha))} \frac{12}{n(n^2-1)}\right)$
- Width of a confidence int. (belt) is proportional to $1/\sqrt{(1-\alpha)n}$
- For normal variable, mean absolute deviation from mean is $\sigma\sqrt{2/\pi}$

$\Rightarrow$ e.g., for fixed α and F, if we quadruple sample size to $4n$, the conf. int. gets $2\times$ narrower and $4\times$ longer which compensates with the fact that mean absolute deviation of $\hat{\beta}_1(\alpha)$ from real trend β_1 gets smaller $\sqrt{4^3}\times = 8\times$.

Noncoverage induced by variance of reg. quant. is constant

- For our model we can show that the α-th regression quantile $\hat{\beta}_1(\alpha)$ is approximately normal $\mathcal{N}\left(\mu = \beta_1, \sigma^2 = \frac{\alpha(1-\alpha)}{f^2(F^{-1}(\alpha))}\frac{12}{n(n^2-1)}\right)$
- Width of a confidence int. (belt) is proportional to $1/\sqrt{(1-\alpha)n}$
- For normal variable, mean absolute deviation from mean is $\sigma\sqrt{2/\pi}$

$\Rightarrow$ e.g., for fixed α and F, if we quadruple sample size to $4n$, the conf. int. gets $2\times$ narrower and $4\times$ longer which compensates with the fact that mean absolute deviation of $\hat{\beta}_1(\alpha)$ from real trend β_1 gets smaller $\sqrt{4^3}\times = 8\times$.

- For F (standard) Gumbel distribution $\frac{\alpha(1-\alpha)}{f^2(F^{-1}(\alpha))} = \frac{1-\alpha}{\alpha(\ln\alpha)^2}$
- Expanding the reciprocal around $\alpha = 1$ we get

$$\frac{1-\alpha}{\alpha(\ln\alpha)^2} \doteq \frac{1}{(1-\alpha)-\frac{(1-\alpha)^3}{12}-\frac{(1-\alpha)^4}{12}-\frac{13(1-\alpha)^5}{180}-\frac{11(1-\alpha)^6}{180}} \overset{\alpha\in(0.7,1)}{\doteq} \frac{1}{1-\alpha}$$

Noncoverage induced by variance of reg. quant. is constant

- For our model we can show that the α-th regression quantile $\hat{\beta}_1(\alpha)$ is approximately normal $\mathcal{N}\left(\mu = \beta_1, \sigma^2 = \frac{\alpha(1-\alpha)}{f^2(F^{-1}(\alpha))}\frac{12}{n(n^2-1)}\right)$
- Width of a confidence int. (belt) is proportional to $1/\sqrt{(1-\alpha)n}$
- For normal variable, mean absolute deviation from mean is $\sigma\sqrt{2/\pi}$

$\Rightarrow$ e.g., for fixed α and F, if we quadruple sample size to $4n$, the conf. int. gets $2\times$ narrower and $4\times$ longer which compensates with the fact that mean absolute deviation of $\hat{\beta}_1(\alpha)$ from real trend β_1 gets smaller $\sqrt{4^3}\times = 8\times$.

- For F (standard) Gumbel distribution $\frac{\alpha(1-\alpha)}{f^2(F^{-1}(\alpha))} = \frac{1-\alpha}{\alpha(\ln\alpha)^2}$
- Expanding the reciprocal around $\alpha = 1$ we get

$$\frac{1-\alpha}{\alpha(\ln\alpha)^2} \doteq \frac{1}{(1-\alpha)-\frac{(1-\alpha)^3}{12}-\frac{(1-\alpha)^4}{12}-\frac{13(1-\alpha)^5}{180}-\frac{11(1-\alpha)^6}{180}} \overset{\alpha\in(0.7,1)}{\overline{\overline{\quad\cdot\quad}}} \frac{1}{1-\alpha}$$

$\Rightarrow$ e.g., for fixed sample size n, if we quadruple $(1-\alpha)$ to $4(1-\alpha)$, the conf. int. gets $2\times$ narrower which is again compensated by $\sqrt{4}\times = 2\times$ smaller absolute dev. of $\hat{\beta}_1(\alpha)$.

Noncoverage induced by variance of reg. quant. is constant

- For our model we can show that the α-th regression quantile $\hat{\beta}_1(\alpha)$ is approximately normal $\mathcal{N}\left(\mu = \beta_1, \sigma^2 = \frac{\alpha(1-\alpha)}{f^2(F^{-1}(\alpha))}\frac{12}{n(n^2-1)}\right)$
- Width of a confidence int. (belt) is proportional to $1/\sqrt{(1-\alpha)n}$
- For normal variable, mean absolute deviation from mean is $\sigma\sqrt{2/\pi}$

$\Rightarrow$ e.g., for fixed α and F, if we quadruple sample size to $4n$, the conf. int. gets $2\times$ narrower and $4\times$ longer which compensates with the fact that mean absolute deviation of $\hat{\beta}_1(\alpha)$ from real trend β_1 gets smaller $\sqrt{4^3}\times = 8\times$.

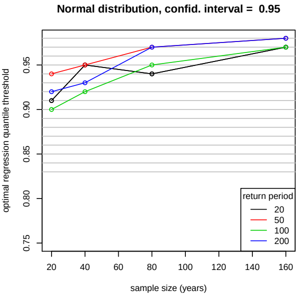
- For F (standard) Gumbel distribution $\frac{\alpha(1-\alpha)}{f^2(F^{-1}(\alpha))} = \frac{1-\alpha}{\alpha(\ln\alpha)^2}$
- Expanding the reciprocal around $\alpha = 1$ we get

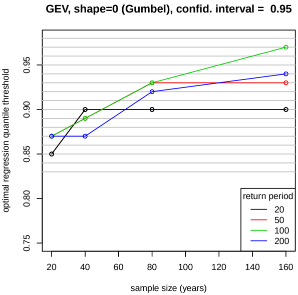
$$\frac{1-\alpha}{\alpha(\ln\alpha)^2} \doteq \frac{1}{(1-\alpha) - \frac{(1-\alpha)^3}{12} - \frac{(1-\alpha)^4}{12} - \frac{13(1-\alpha)^5}{180} - \frac{11(1-\alpha)^6}{180}} \overset{\alpha\in(0.7,1)}{\doteq\!=} \frac{1}{1-\alpha}$$

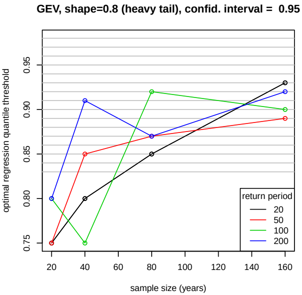
$\Rightarrow$ e.g., for fixed sample size n, if we quadruple $(1-\alpha)$ to $4(1-\alpha)$, the conf. int. gets $2\times$ narrower which is again compensated by $\sqrt{4}\times = 2\times$ smaller absolute dev. of $\hat{\beta}_1(\alpha)$.

Conclusion: The optimal quantile as threshold is the same as if there was no trend in data and we used ordinary quantiles as threshold.

Other distributions F







Form of $\frac{\alpha(1-\alpha)}{f^2(F^{-1}(\alpha))}$ for different distributions F

Normal distribution	GEV,shape = 0 (Gumbel)	GEV,shape = 0.8 (heavy)
$\approx \frac{1}{\sqrt{1-\alpha}}$	$= \frac{1-\alpha}{\alpha(\ln \alpha)^2} \approx \frac{1}{1-\alpha}$	$\approx \frac{1}{(1-\alpha)^{2.6}}$

Outline

Future work

- Take data with different structure
  - Investigate autocorrelated data.
  - Different forms of trend.
- Use other methods to estimate the return ”levels”. Bootstrap?

Bibliography

Koenker R. and Bassett G. (1978).
Regression quantiles.
Econometrica, **46**, 33-50.

Kyselý J., Picek J., Beranová R. (2010).
Estimating extremes in climate change simulations using the peaks-over-threshold method with a non-stationary threshold.
Global and Planetary Change, **72**, 55-68.