Royal Netherlands
Meteorological Institute
Ministry of Infrastructure and the Environment

Threshold selection for regional peaks-over-threshold data

Martin Roth

roth@knmi.nl

Joint work with A. Buishand and G. Jongbloed

Wij zijn het KNMI

Het weer is grillig, de bodem beweegt en het klimaat verandert. Voor onze veiligheid en welvaart moeten we weten welke risico's en kansen dit oplevert. En: hoe we ons het beste kunnen voorbereiden. Die kennis heeft het KNMI in huis als het nationale kennis- en datacentrum voor weer, klimaat en seismologie. Betrouwbaar, onafhankelijk en gericht op wat Nederland nodig heeft.

EVT related problems

- Combined extremes
- Climate model validation
- Improvement of extreme weather predictions
- Earthquake analysis

Storms, flooding prompt Dutch evacuations

5 January 2012 | Europe

Share

Dutch villagers have been asked to leave their homes and farms because strong winds and heavy rain have led to fears of coastal flooding.

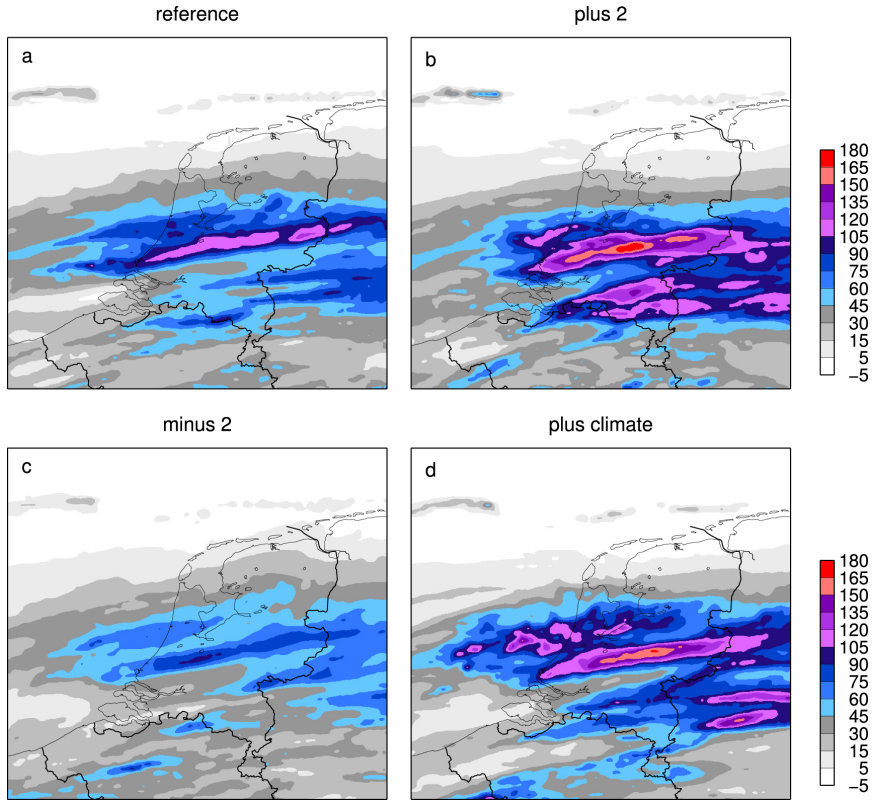
Around 100 villagers had to leave their homes as a dyke came close to collapse

Some farmers in Tolbert in the northern province of Groningen refused to leave their cattle, despite reports that an inland dyke was at risk of breaking.

There are flood warnings for the Dutch west coast. Gales have also battered the UK, Belgium, France and Germany.

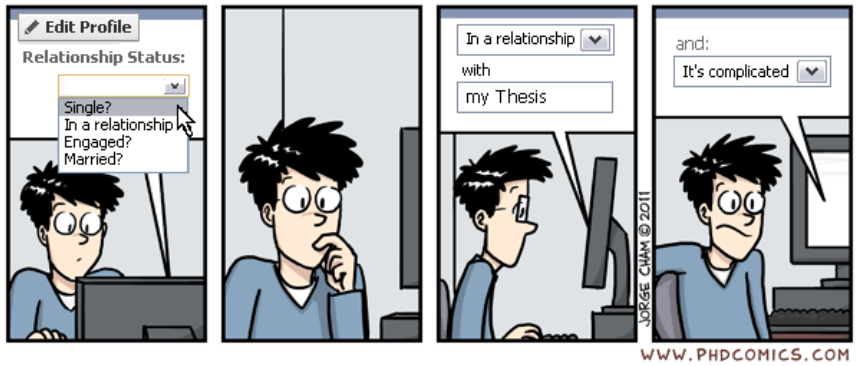
EVT related problems

- Combined extremes
- Climate model validation
- Improvement of extreme weather predictions
- Earthquake analysis



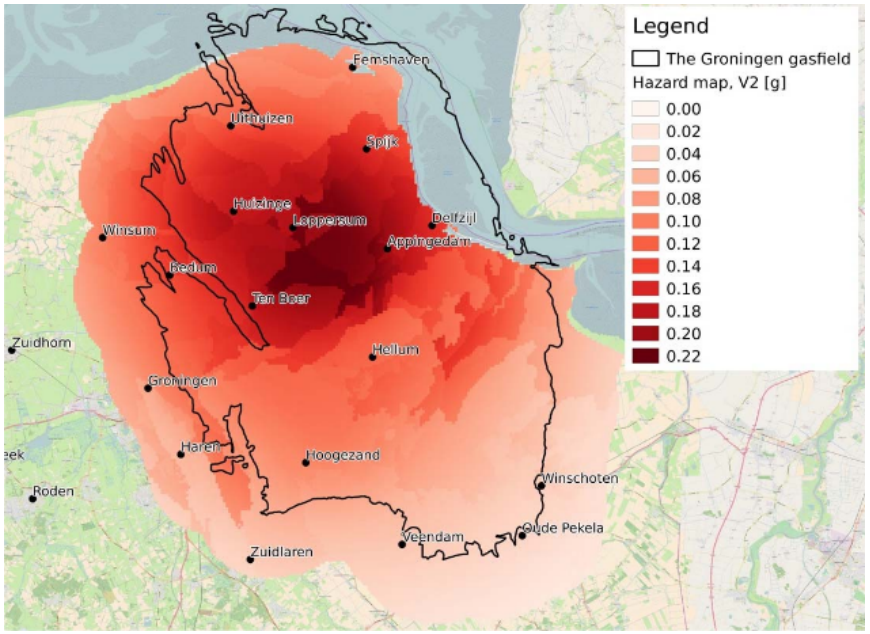
EVT related problems

- Combined extremes
- Climate model validation
- Improvement of extreme weather predictions
- Earthquake analysis

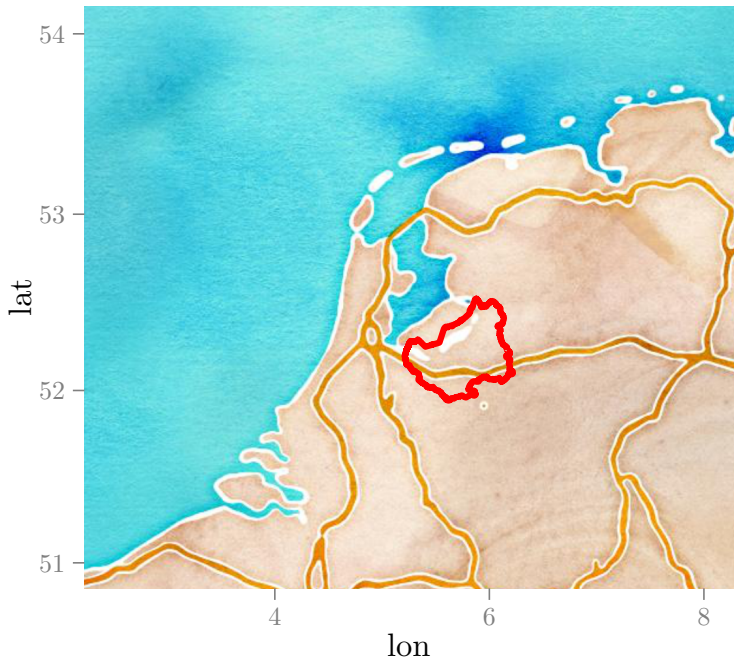


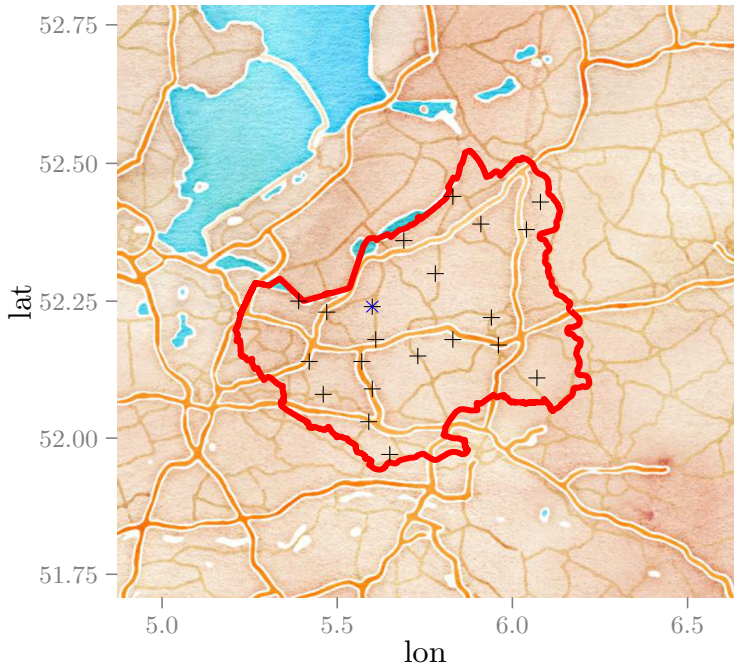
EVT related problems

- Combined extremes
- Climate model validation
- Improvement of extreme weather predictions
- Earthquake analysis



Study Setting





Waterboard Vallei & Veluwe
21 daily precipitation series for 1951–2009

Buishand et al. (2013): Homogeneity of precipitation series in the Netherlands and their trends in the past century. *Int. J. Climatol.,*

RFA

Regional frequency analysis (RFA)

- Estimate the frequency of rare events at one site by using data from several sites
- For instance by assuming that the marginal distributions are identical apart from a scaling factor

RFA

Regional frequency analysis (RFA)

- Estimate the frequency of rare events at one site by using data from several sites
- For instance by assuming that the marginal distributions are identical apart from a scaling factor

$$r_s(T) = \begin{cases} u_s \left(1 - \frac{\sigma_s / u_s}{\xi} \left[1 - (\lambda_s T)\right]^{\xi}\right), & \xi \neq 0, \\ u_s \left(1 + \sigma_s / u_s \log\left(\lambda_s T\right)\right), & \xi = 0. \end{cases}$$

RFA

Regional frequency analysis (RFA)

- Estimate the frequency of rare events at one site by using data from several sites
- For instance by assuming that the marginal distributions are identical apart from a scaling factor

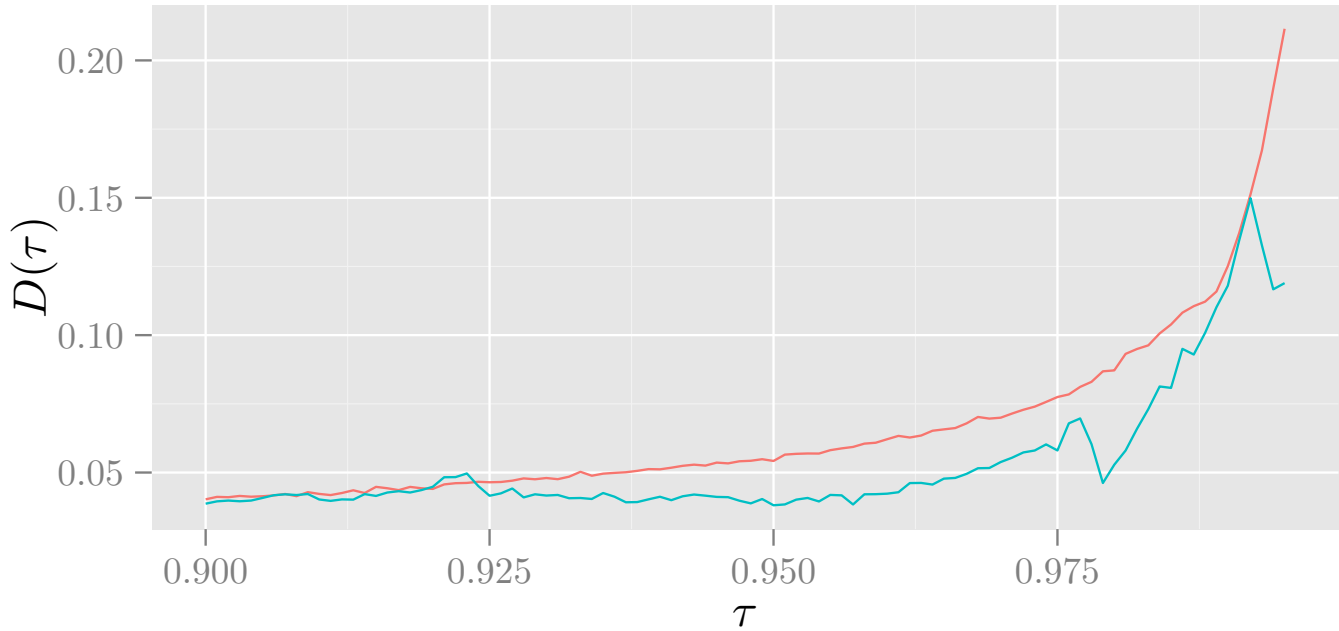
$$r_s(T) = \begin{cases} u_s \left(1 - \frac{\gamma}{\xi}\left[1 - (\lambda T)\right]^{\xi}\right), & \xi \neq 0, \\ u_s \left(1 + \gamma \log\left(\lambda T\right)\right), & \xi = 0. \end{cases}$$

Threshold selection

Trade-off situation

- Low thresholds lead to bias in the analysis of the excesses
- High thresholds result in high parameter estimation uncertainty

Threshold selection



KS statistic for the precipitation data in Putten as function of the non-excess probability τ, together with the 95% critical values.

Threshold selection

Trade-off situation

- Low thresholds lead to bias in the analysis of the excesses
- High thresholds result in high parameter estimation uncertainty

Regional scale

- Most studies use the same quantile threshold
- Influence of threshold level is not often studied

Smooth splicing

Any continuous distribution F on $[0, \infty)$ can be written as

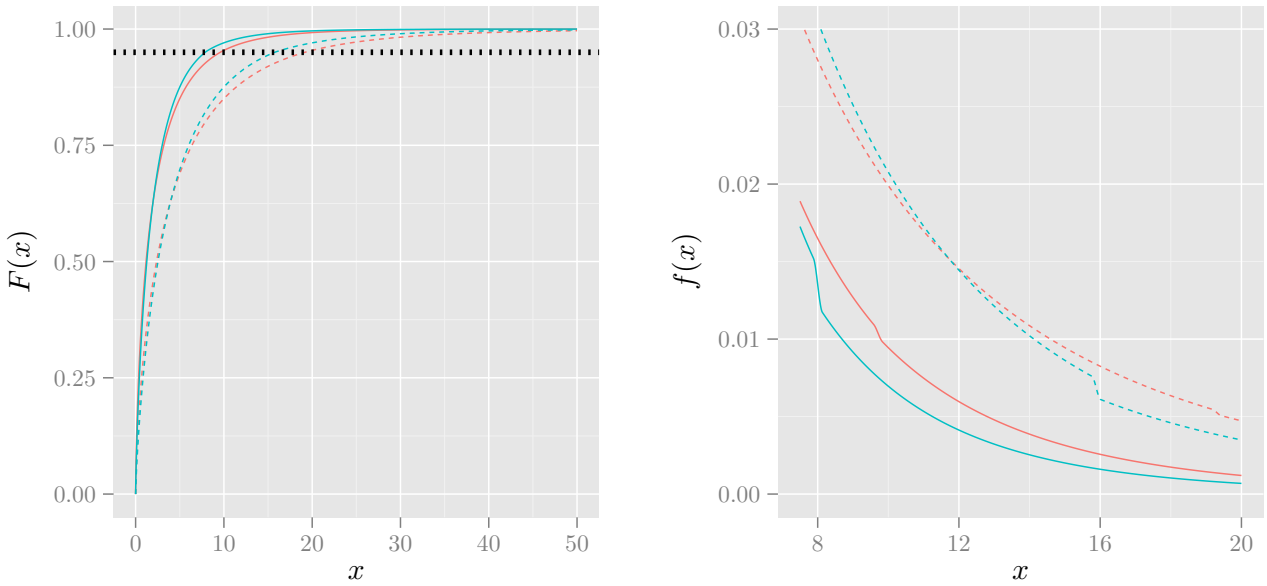
$$F(x) = 1 - \exp\Big(-\int_0^x h(u)\,du\Big),$$

where h is the hazard rate of the distribution. Define

$$h(x) := \eta\left(\frac{x-u}{\varepsilon}\right) h_1(x) + \left(1 - \eta\left(\frac{x-u}{\varepsilon}\right)\right) h_2(x),$$

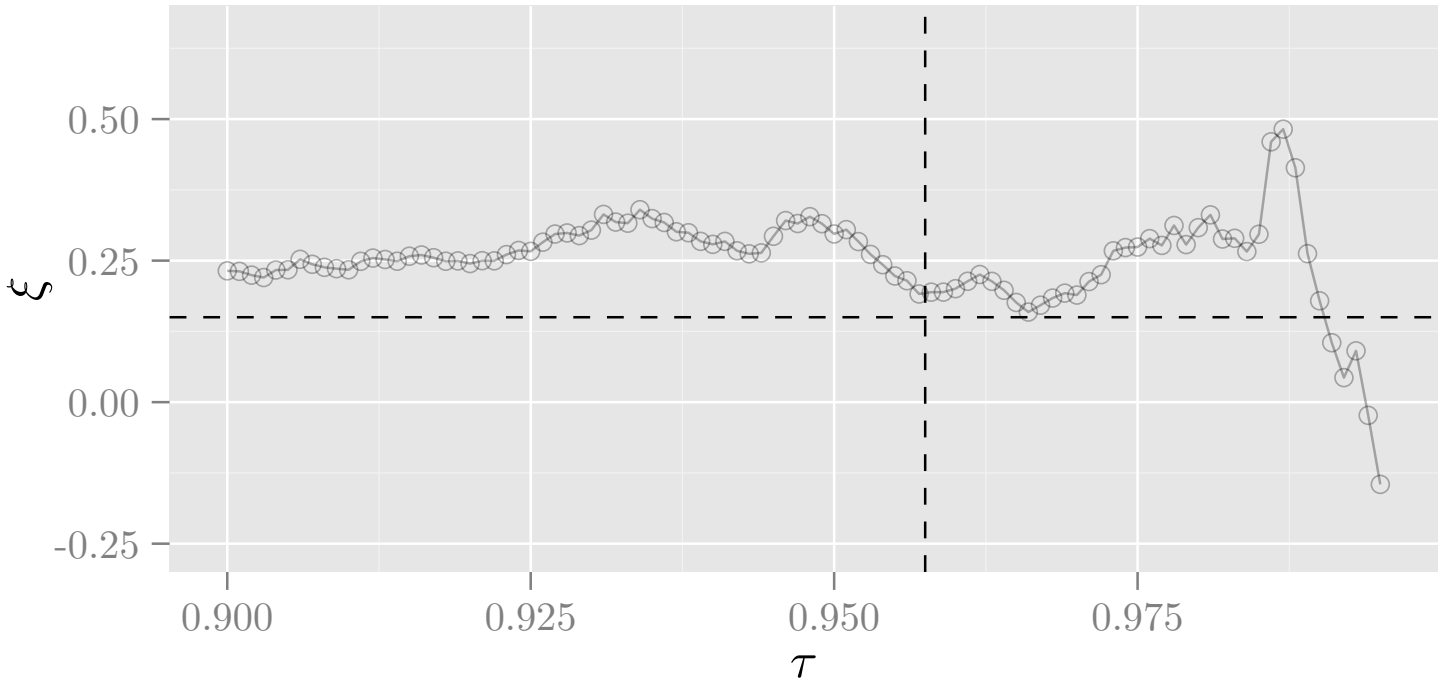
with h_1 the hazard rate of some bulk distribution, h_2 the hazard rate of the GPD, and η some smooth transition function.

Weibull-GPD model



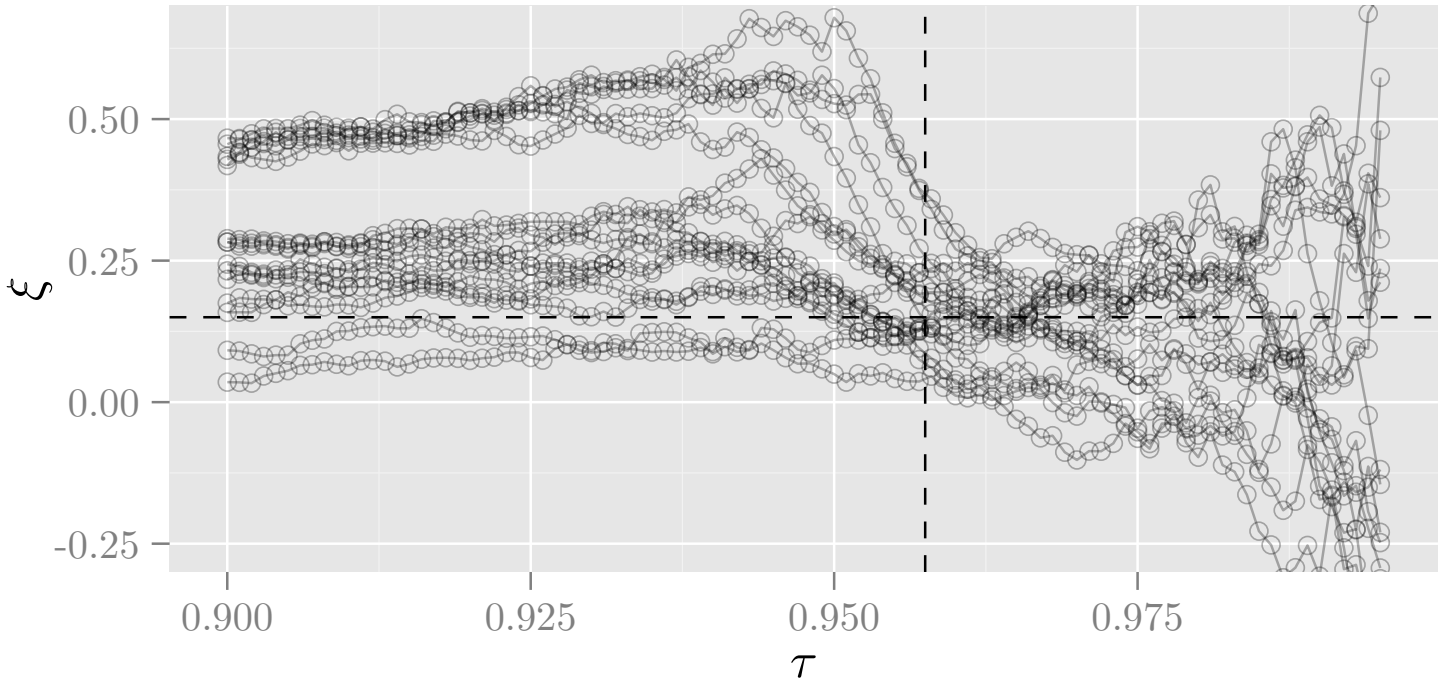
Cdf and density of different hybrid Weibull-GPD distributions.

Threshold Stability Plot



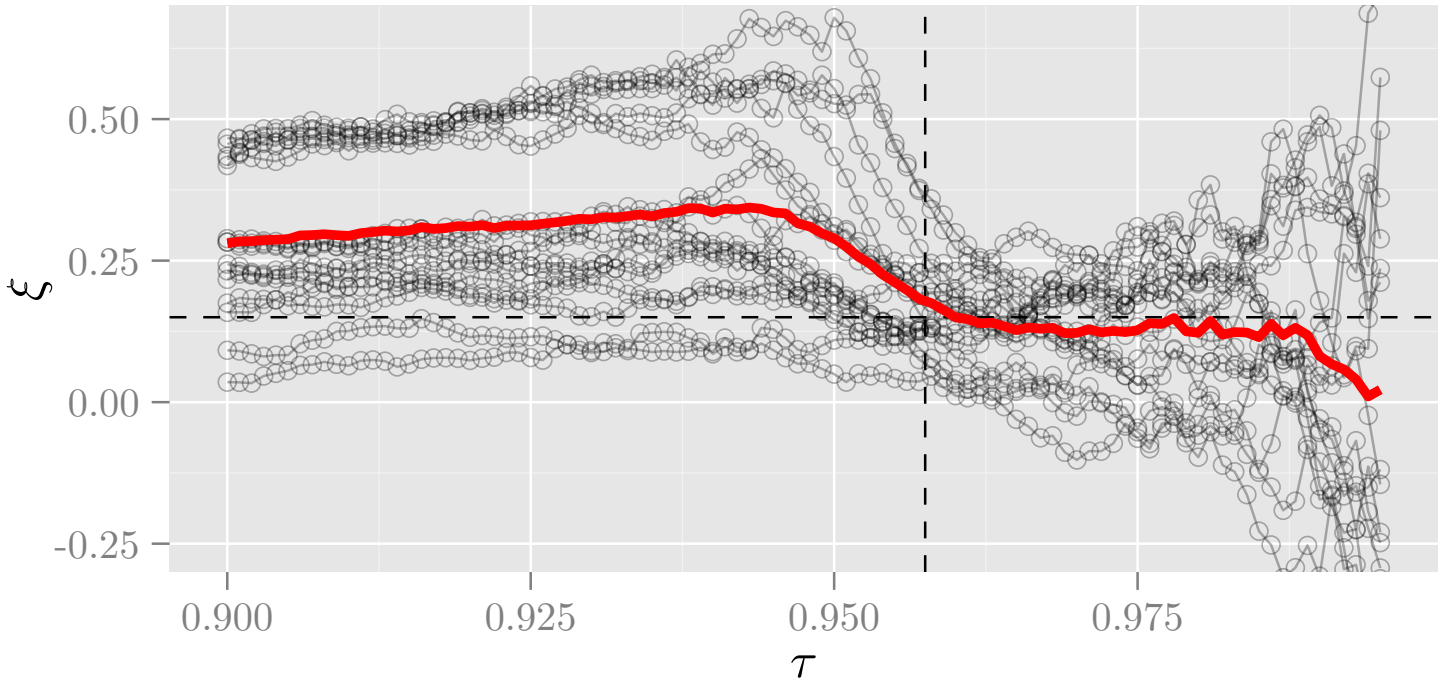
τ is the non-exceedance probability. The vertical line indicates the start of the GPD tail and the horizontal line shows the true shape parameter of the simulated data.

Threshold Stability Plot



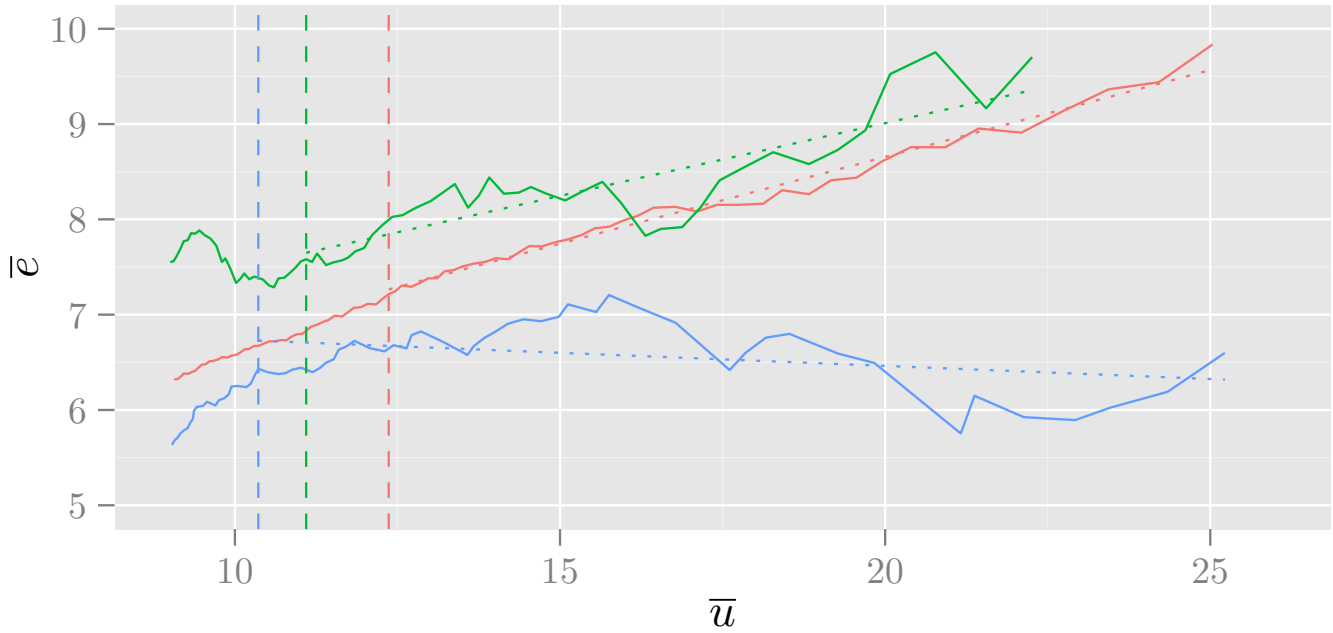
τ is the non-exceedance probability. The vertical line indicates the start of the GPD tail and the horizontal line shows the true shape parameter of the simulated data.

Threshold Stability Plot – Averaged



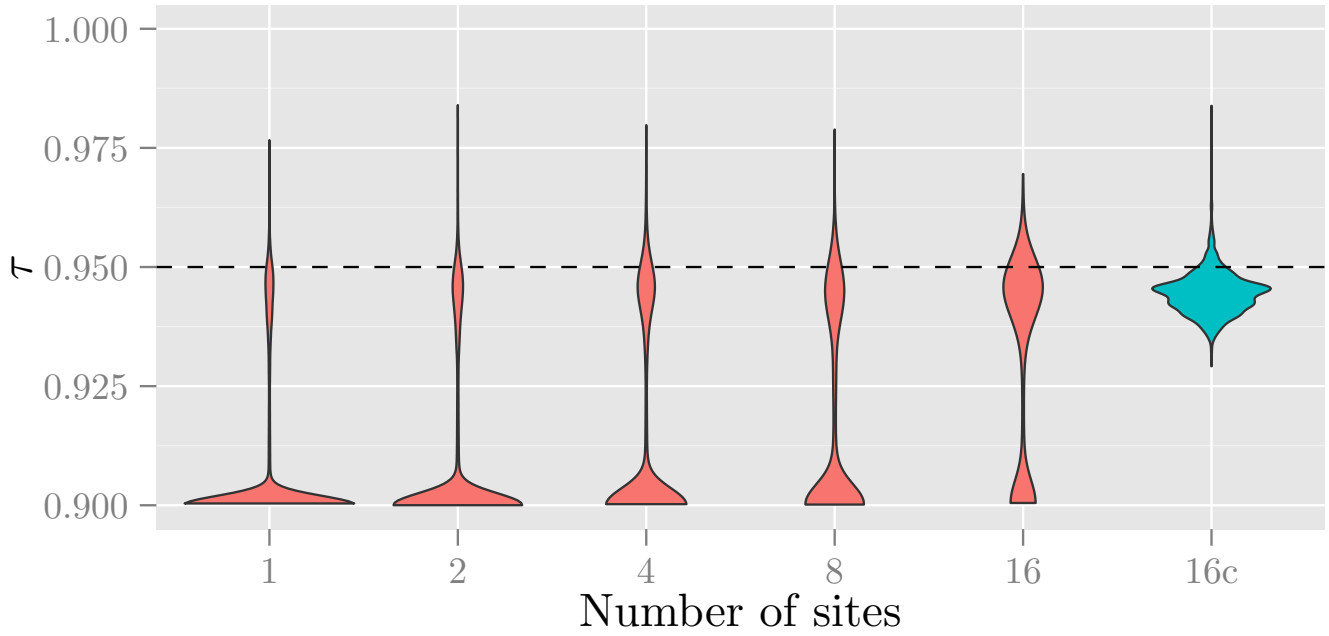
τ is the non-exceedance probability. The vertical line indicates the start of the GPD tail and the horizontal line shows the true shape parameter of the simulated data.

Mean excess plot



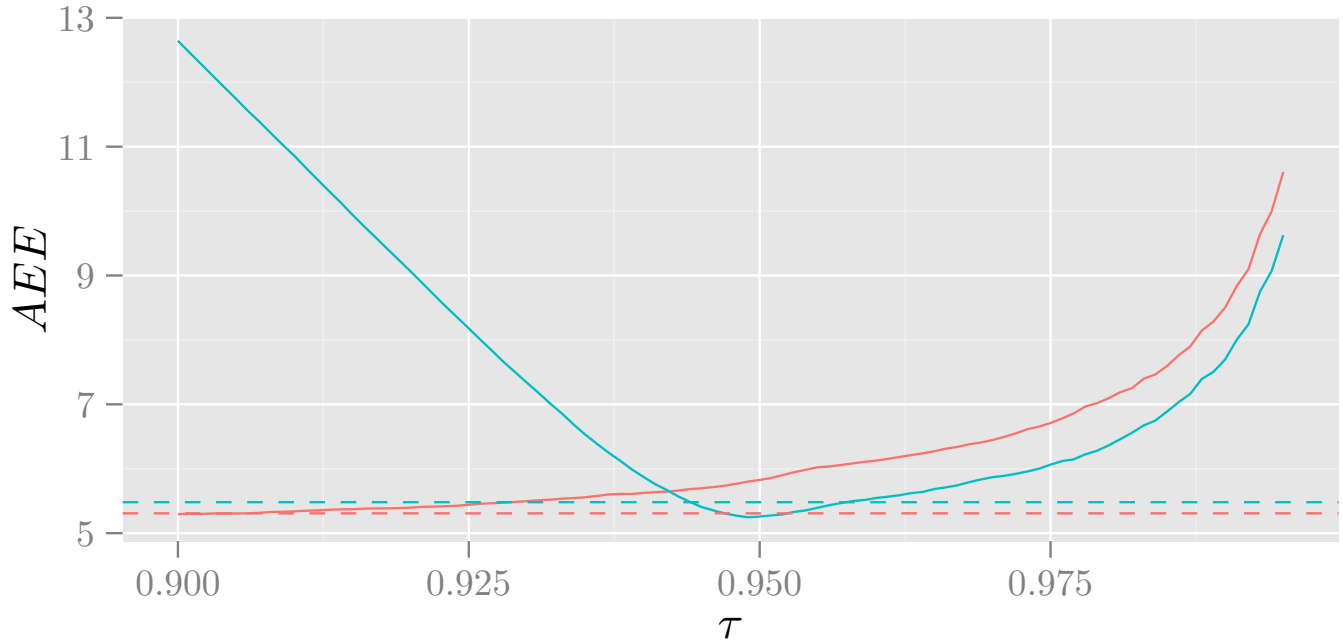
Mean excess plot for 1 (blue), 8 (green), and 16 (red) sites.

Averaged KS statistic



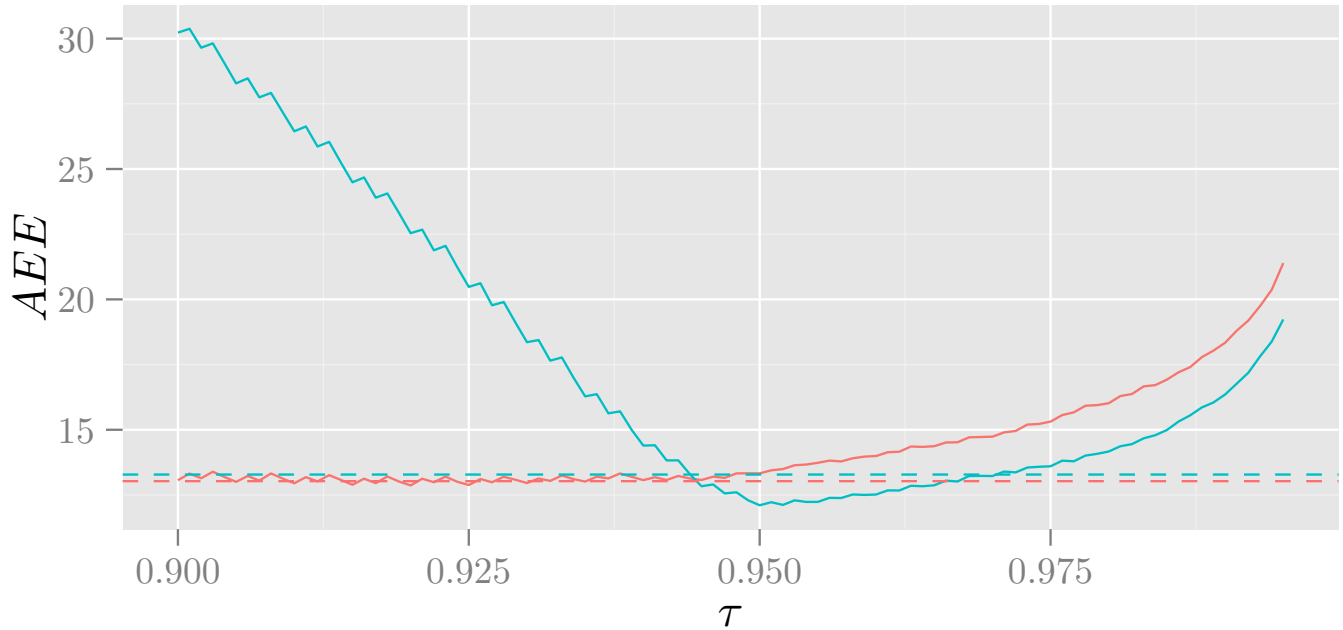
Violin plots of the selected threshold based on the lowest value of τ for which the average KS statistic is not significant at the 5% level.

Effect of Bulk–Tail Transition



AEE in the 5-year return level as a function of the probability τ for the simulated data (blue - bulk and tail are different, red - similar). The dashed horizontal lines indicate the AEE of the selected threshold.

Effect of Bulk–Tail Transition



AEE in the 50-year return level as a function of the probability τ for the simulated data (blue - bulk and tail are different, red - similar). The dashed horizontal lines indicate the AEE of the selected threshold.

Spatial dependence model

Copula approach

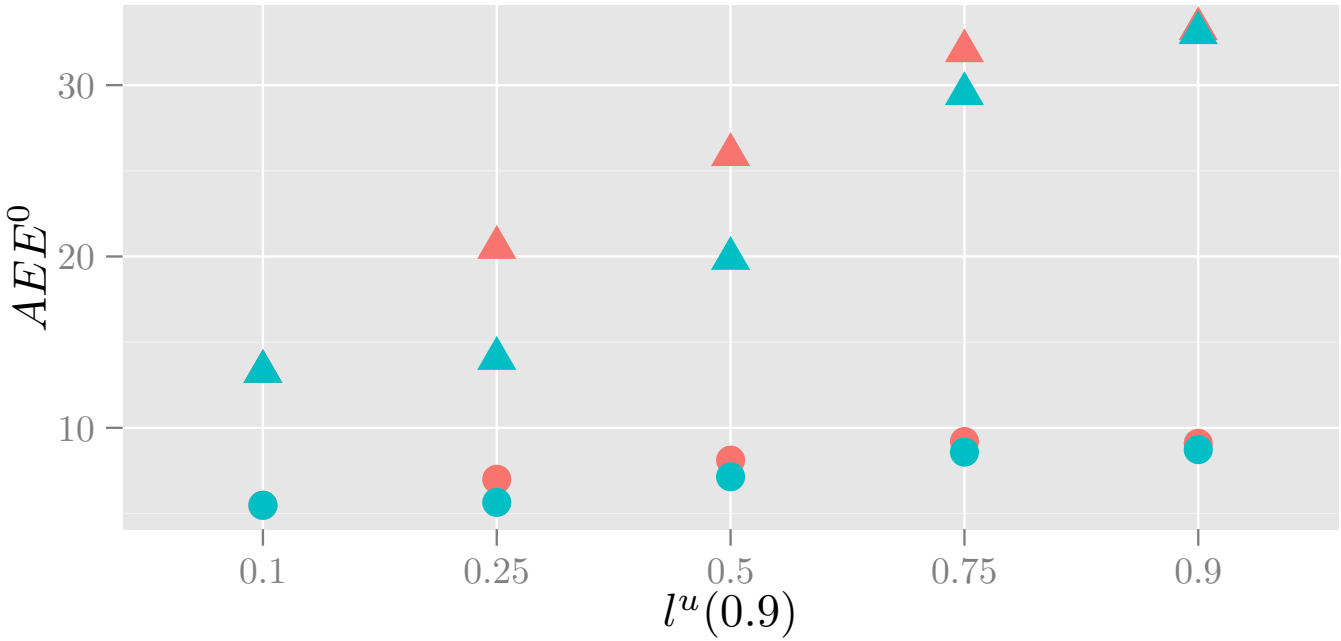
- Normal copula $\Rightarrow$ weak tail dependence
- Gumbel copula $\Rightarrow$ strong tail dependence

Quantile based measure of tail dependence

$$l^{\mathrm{u}}(\tau) := P\left(X_1 > F_1^{-1}(\tau) | X_2 > F_2^{-1}(\tau)\right)$$

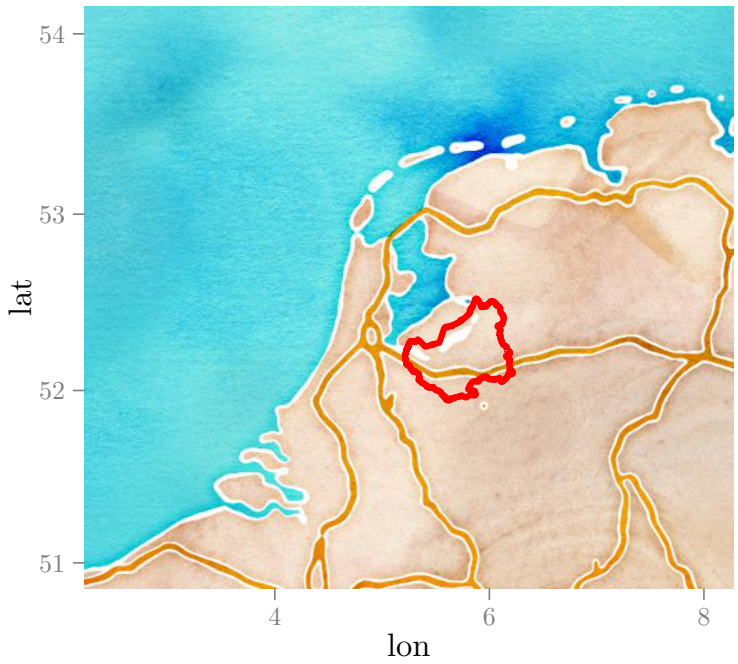
Copula parameter can be chosen such that $l^{\mathrm{u}}(0.9)$ is the same.

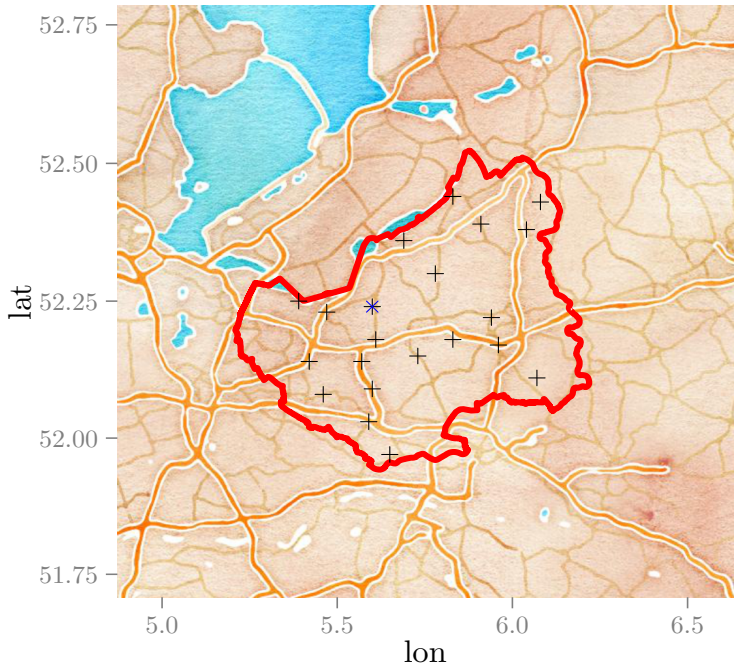
Effect of Spatial Dependence



Averaged Euclidean error of the 5- (dots) and 50-year (triangles) return level for simulated data (red - Gumbel copula, blue - normal copula)

Study Setting

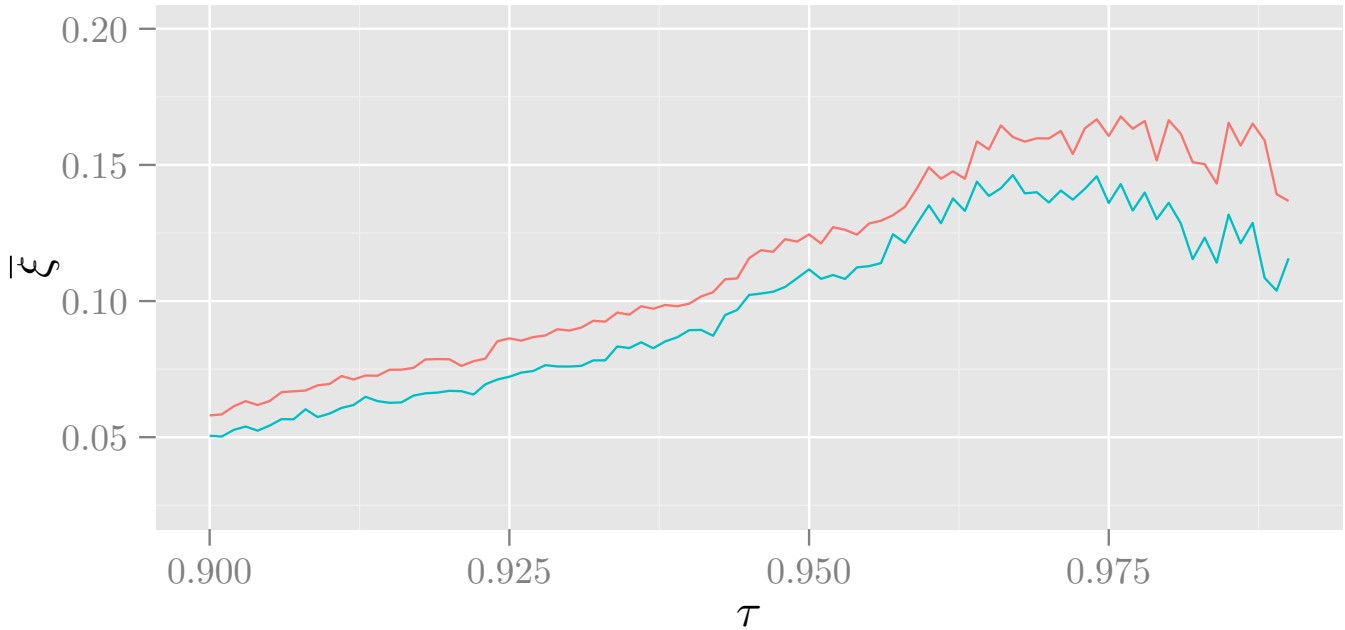




Waterboard Vallei & Veluwe
21 daily precipitation series for 1951–2009

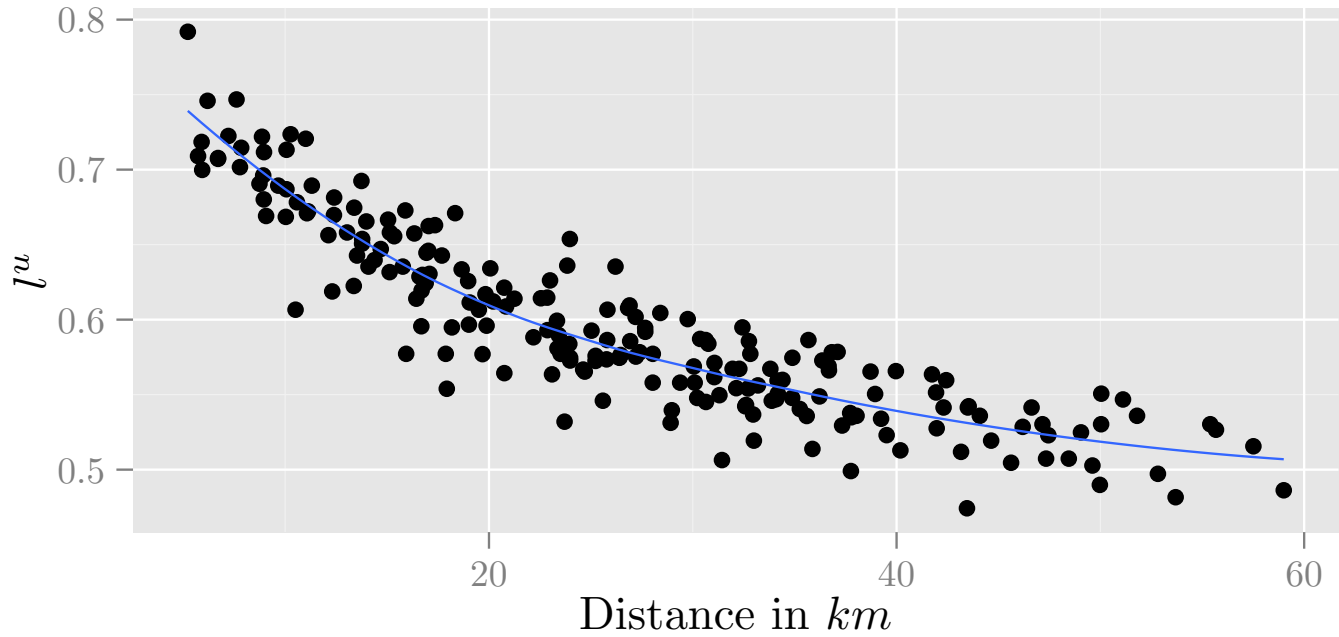
Buishand et al. (2013): Homogeneity of precipitation series in the Netherlands and their trends in the past century. *Int. J. Climatol.,*

Spatially averaged TS plot



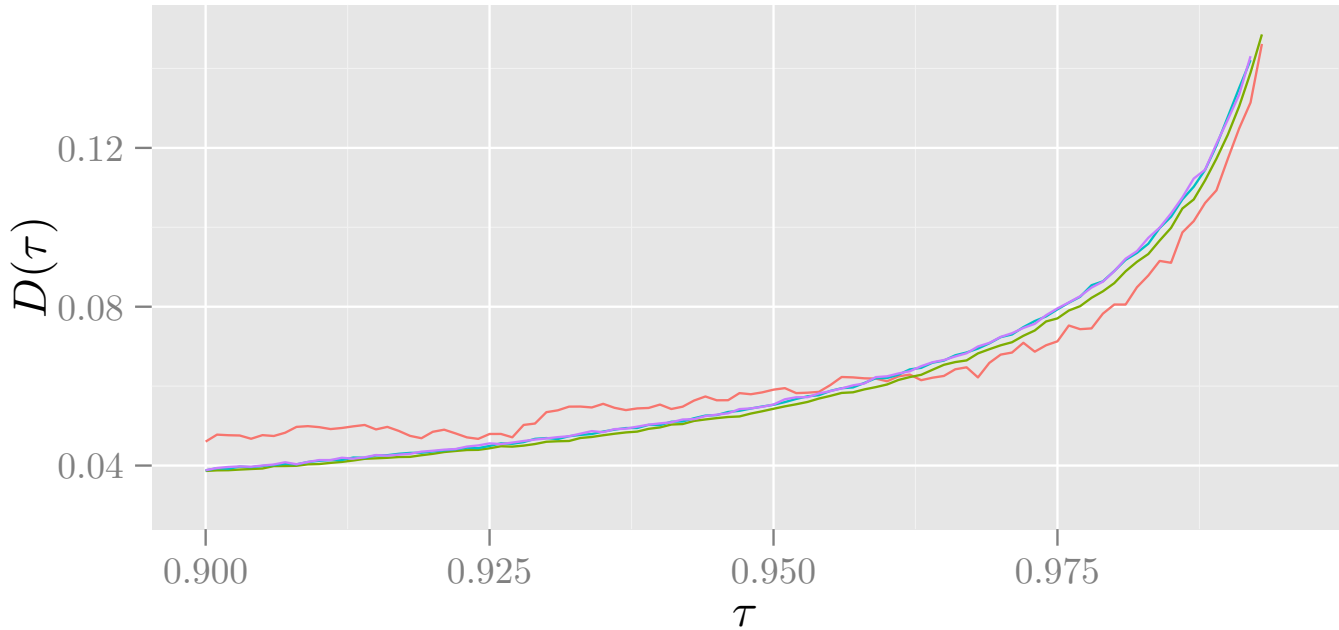
Threshold stability plot for the rainfall data.

Quantile dependence in the data



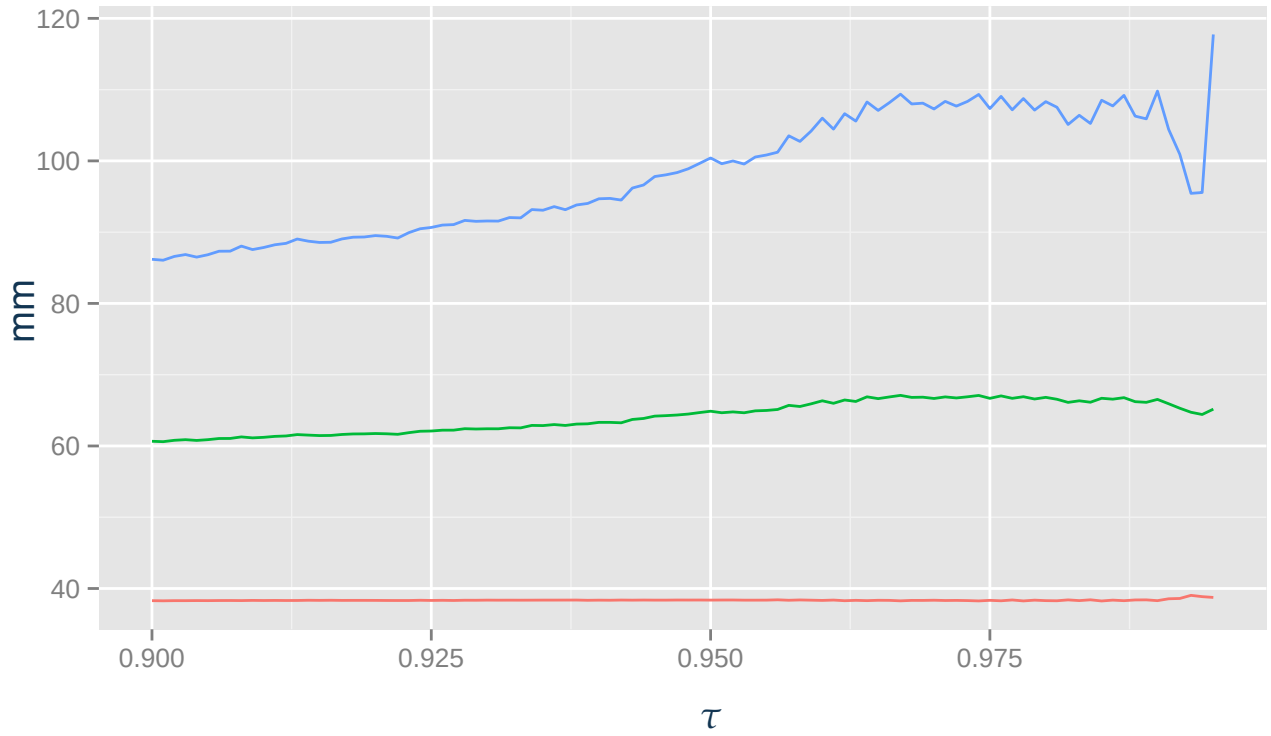
Dependence in the 0.9 quantile of the rainfall data versus distance.

Spatially averaged KS statistic



Spatially averaged KS statistic for the three different dependence models.

Effect on Return Levels



Average return level as a function of the non-exceedance probability of the selected threshold for return periods of 5 (red), 50 (green) and 500 (blue) years.

Conclusion / Summary

Use RFA principles also for threshold selection!

GoF based approach not restricted to the GPD.

Outlook

- Inspect the role of selection criterion
- Extend the approach to incorporate trends

Roth, M., G. Jongbloed, and T. A. Buishand (2016),
Threshold selection for regional peaks-over-threshold data.
Journal of Applied Statistics, 43:1291–1309.

Kolmogrov-Smirnov statistic

(1) Fix τ_0 and the corresponding threshold u.

(2) Estimate the parameters based on the n excesses above u.

(3a) Simulate n values from the corresponding GPD.

(3b) Sample $\tilde{n}$ from $B(T, 1-\tau_0)$. Simulate $\tilde{n}$ values from the corresponding GPD. Scales to regional setting.

(4) Calculate KS statistic for the simulated data.

(5) Repeat steps 3 and 4 a thousand times and take the 0.95-quantile of the bootstrapped statistic as critical value for D_n at τ_0.

(6) Repeat procedure for every τ in a reasonable range.

Averaged Euclidean Error (AEE)

$$AEE(\tau) := \frac{1}{B} \sum_{i=1}^{B} ||\hat{\boldsymbol{r}}_i(\tau) - \boldsymbol{r}||_2,$$

where $||.||_2$ denotes the Euclidean (or l_2) norm, B gives the number of bootstrap samples and $\boldsymbol{r}$ specifies some return level.