Modeling clusters of extreme values

Johanna G. Nešlehová

Joint work with
Jonathan Jalbert, Orla Murphy and Christian Genest

McGill University, Montréal, Canada

Wirtschaftsuniversität Wien, Austria

1. Extreme events

Major natural catastrophes in 2016:

- Earthquakes (Italy, Indonesia, Japan, New Zealand, Taiwan)
- Floods (Germany, Louisiana)
- Wildfires (California, Tennessee)
- Storms and hurricanes (Jonas, Matthew)
- Famines, etc.

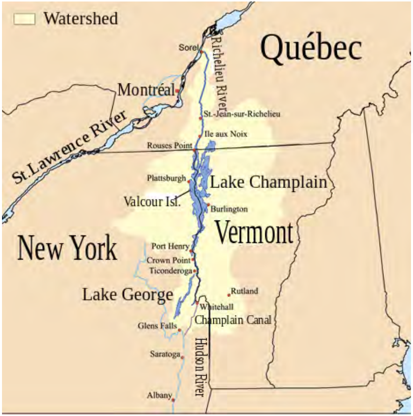
They had disastrous human, economic, and financial consequences.

The 2011 Richelieu River flood

- It lasted two months (mid-April to mid-June).
- Thousands of citizens were evacuated.
- Damages were estimated at $100 million USD.

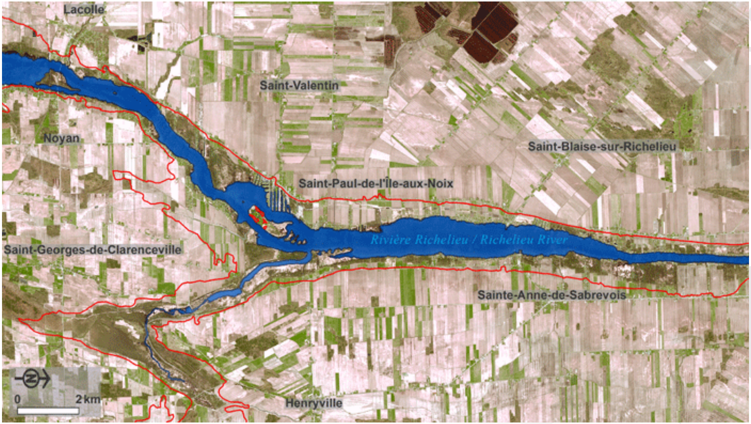
ⓒBernard Brault - The Canadian Press

The Richelieu River Watershed



- The Richelieu is the only outlet of the watershed.
- Lake Champlain acts as a natural buffer against a flood surge in the river.
- More than 90% of the water passing through the Richelieu comes from Lake Champlain.
- The river discharges are strongly correlated with the lake levels (Riboust & Brissette, 2015).

Animation


Lacolle
Saint-Valentin
Noyan
Saint-Blaise-sur-Richelieu
Saint-Paul-de-l'Île-aux-Noix
Saint-Georges-de-Clarenceville
Rivière Richelieu / Richelieu River
Sainte-Anne-de-Sabrevois
Henryville
0
2 km

No reliable return period for the 2011 flood has been provided sofar.

Extremes need EVT

- Standard statistical techniques work well for large data sets and focus on common features.
- In contrast, natural disasters are usually due to observations that are atypical and rare.
- To guard against future extreme events, risk measures are typically computed at a high level (high quantiles, long return periods).
- Careful extrapolation beyond observed data can be accomplished using extreme-value theory (EVT) methods.

Standard EVT techniques

The block maxima method proceeds as follows:

- The data are grouped into n consecutive blocks of equal size m.
- The maximum in each group is computed, yielding n block maxima

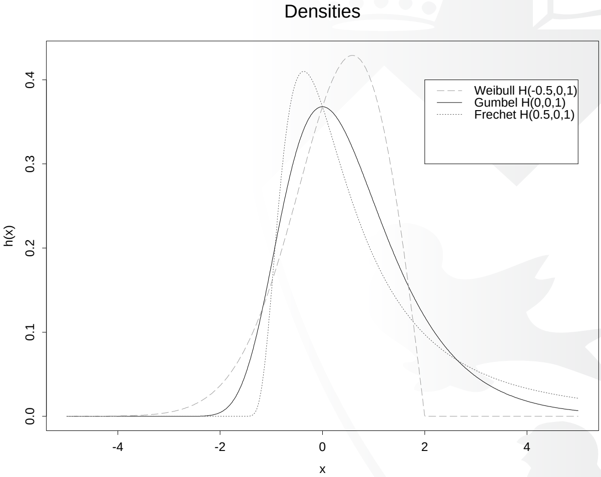
  $$M_1, \ldots, M_n$$

  which are regarded as independent and identically distributed.
- Using the Fisher–Tippett Theorem, the extreme–value distribution

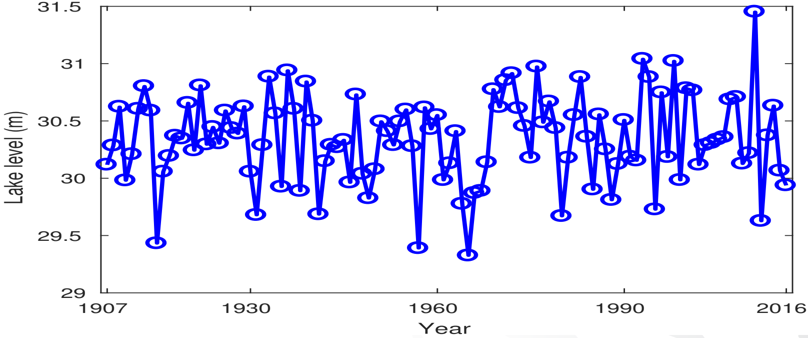
  $$H(x;\mu,\sigma,\xi) = \begin{cases} \exp\left[-\left\{1+\xi\,\frac{(x-\mu)}{\sigma}\right\}^{-1/\xi}\right] & \text{if } \xi \neq 0, \\ \exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\} & \text{if } \xi = 0, \end{cases}$$

  where $1+\xi(x-\mu)/\sigma > 0$, is fitted to $M_1, \ldots, M_n$.

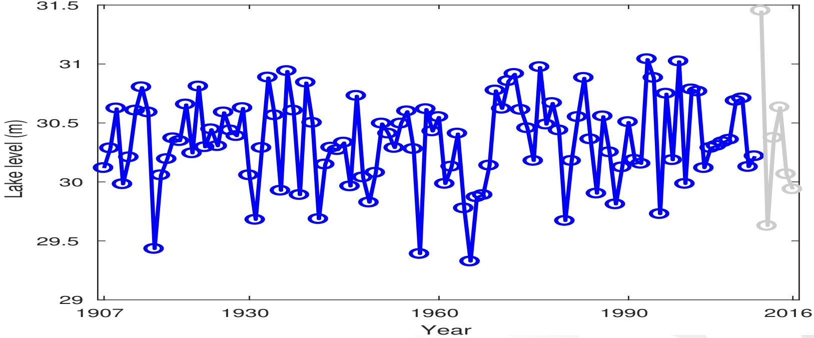
Graph of the EVT density



The lake level annual maxima series



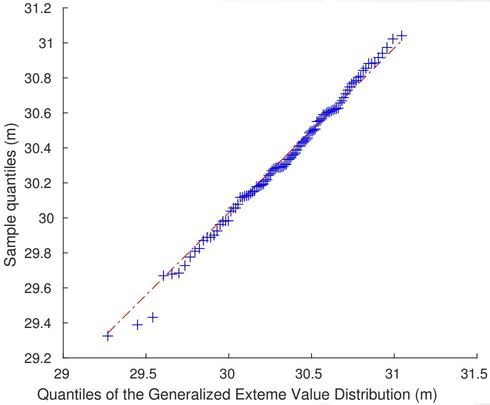
The lake level annual maxima series



Fitted GEV with the annual maxima before 2011

The parameter estimates with 95% credibility intervals are

$$\hat{\mu} = 30.2 \in (30.1, 30.3),\ \hat{\sigma} = 0.39 \in (0.34, 0.44),\ \hat{\xi} = -0.44 \in (-0.56, -0.32).$$



Red flags

- The shape parameter estimate $\hat{\xi}$ is negative. This means that the distribution of the annual maximum lake level is bounded above.

- According to the fitted model,

$$\Pr(M > 31.2) = 0$$

  However, the lake level reached in the 2011 flood was 31.45m.

- ML estimation is irregular when $\xi < -.5$ (Smith 1985). This is a possibility here, given that $\hat{\xi} = -0.44 \in (-0.56, -0.32)$.

Black Swan!

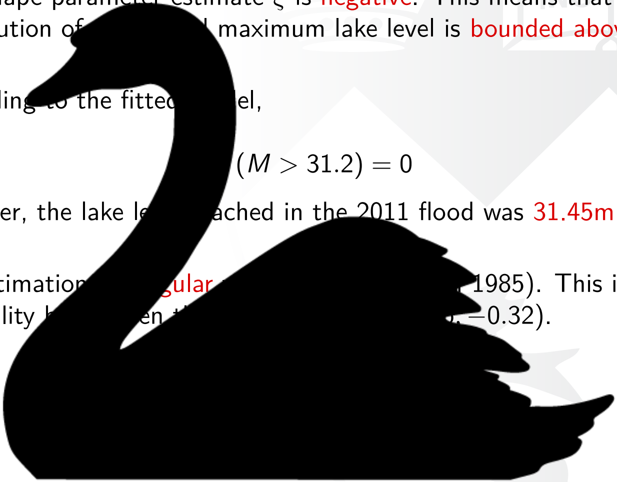
- The shape parameter estimate $\hat{\xi}$ is negative. This means that the distribution of [illegible] maximum lake level is bounded above.

- According to the fitted model,

$$(M > 31.2) = 0$$

However, the lake le[illegible]ached in the 2011 flood was 31.45m.

- ML estimation [illegible]gular [illegible] 1985). This is a possibility b[illegible]en t[illegible] -0.32).



Variables that affect the lake level

- The spring freshet is the result of the snowmelt and the occurrence of precipitation during this period.
- Riboust & Brissette (2015) showed that the key factor for explaining the flood severity is the amount of spring precipitation.
- Precipitation at Burlington, VT, can be used as a proxy for precipitation levels in the watershed.
- The snowpack on the watershed and the temperature do not have a significant impact on the maximal water level.

Precipitation data

- Consider the spring daily precipitation series recorded at Burlington, VT, from 1883 to 2010.

- We consider spring precipitation, April 1 to June 30, i.e., the critical period of snowmelt when large precipitation can trigger a flood.

- The spring precipitation maxima series can be assumed stationary according to the Mann–Kendall stationarity test.

Standard EVT techniques (cont'd)

The peaks-over-threshold method (POT) proceeds as follows:

1) Choose a high threshold u.

2) Consider excesses of the observations $Z_1, \dots, Z_n$ above u, viz.

$$W_j = Z_j - u \quad \text{for} \quad Z_j > u$$

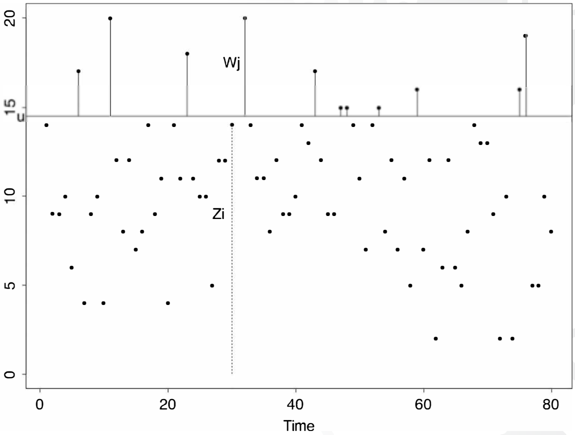
which are regarded as independent and identically distributed.

3) Using the Pickands–Balkema–de Haan Theorem, the generalized Pareto distribution

$$G(w; \beta, \xi) = \begin{cases} 1 - (1 + \xi w/\beta)^{-1/\xi} & \text{if } \xi \neq 0, \\ 1 - \exp(-w/\beta) & \text{if } \xi = 0, \end{cases}$$

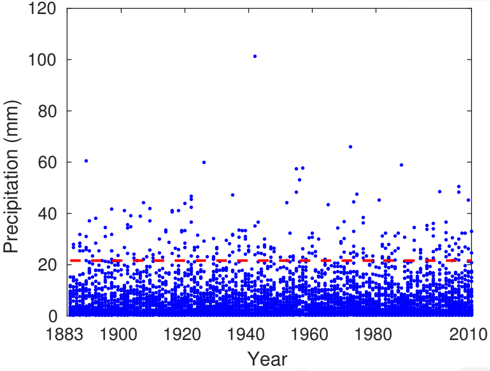
where $1 + \xi w/\beta > 0$, is fitted to $W_1, \dots, W_m$.

Illustration of the POT method



POT method applied to the Burlington data

The threshold was set at the 95th percentile of positive precipitation amounts, i.e., $u = 21.6$mm.



Declustering the series

- The series of exceedances for the Burlington precipitation data exhibits autocorrelation.
- This autocorrelation is typically removed using the runs method.
- Any two consecutive threshold exceedances separated by r or more non-exceedances are considered to belong to different clusters.
- The POT method is then applied to cluster maxima.

Results of the POT analysis

The runs method with lag $r = 1$ was used.

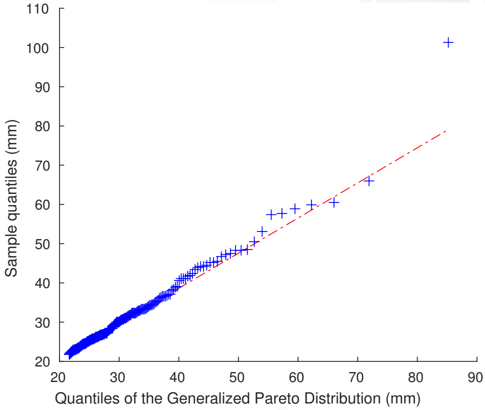
Using a Bayesian analysis, the parameter estimates along with the 95% credibility intervals were

$$\hat{\beta} = 8.5394 \in (7.0168, 10.2379), \quad \hat{\xi} = 0.0655 \in (-0.0454, 0.2045).$$

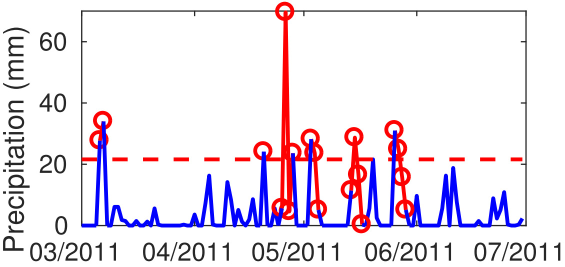
The return period for the extreme rainfall of 69.6mm that occurred on 26 April, 2011, is estimated to be $T = 66$ years.

The model seems to fit well (see next slide), yet no flood was ever observed that matches the 2011 flood in magnitude.

Results of the POT analysis

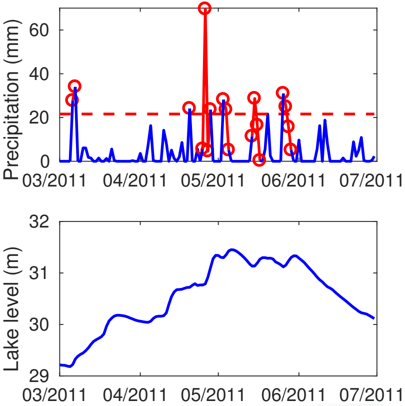


What happened in the spring of 2011?



10 threshold exceedances occurred, forming 6 clusters
(the threshold was fixed at the 95th percentile of positive precipitation)

What happened in the spring of 2011?



The lake reached its historical level after a 4-day streak of precipitation (103mm of rain fell during these 4 days)

The need for more sophisticated methodology

- Events like the 2011 flood are largely due to heavy rainfalls occurring in clusters over several days.
- To estimate the return period of the 2011 flood, the entire clusters of extreme precipitation should be taken into account.
- The conditional exceedance model of Heffernan and Tawn (2004) could be used to this end, but it leads to underestimation in the present context.
- Instead, we propose a new modeling strategy inspired by the M3 Dirichlet model proposed by Süveges and Davison (2012).

What is an M3 process?

- Assume independent shock sequences of i.i.d. unit Fréchet variables

$$(Z_{1,i}), \quad (Z_{2,i}), \quad (Z_{3,i}), \quad \ldots$$

- Take a filter matrix with elements $a_{\ell k} \geq 0$, $\ell \in \mathbb{N}$, $k \in \mathbb{Z}$, such that

$$\sum_{\ell=1}^{\infty} \sum_{k=-\infty}^{\infty} a_{\ell k} = 1.$$

- The M3 process (Maxima of Moving Maxima) is then given by

$$Y_i = \max_{k \in \mathbb{Z}} \big(\max_{\ell \in \mathbb{N}} a_{\ell k} Z_{\ell, i-k} \big), \quad i \in \mathbb{Z}.$$

- We will only use finitely many sequences and a finite window, i.e.,

$$a_{\ell k} > 0 \quad \Rightarrow \quad \ell \in \{1, \ldots, L\}, \quad k \in \{1, \ldots, K\}.$$

Profiles and polar coordinates

Suppose that a single extreme event occurs at time t in the ℓth shock sequence. The profile $(Y_{t+1}, \ldots, Y_{t+K})$ is then of the form

$$(Y_{t+1}, \ldots, Y_{t+K}) = (a_{\ell 1} Z_{\ell,t}, \ldots, a_{\ell K} Z_{\ell,t}).$$

Introduce the normalized profile, viz.

$$(V_{t+1}, \ldots, V_{t+K}) = \frac{1}{\sum_{k=1}^{K} Y_{t+k}} (Y_{t+1}, \ldots, Y_{t+K})$$

and observe that (Zhang & Smith, 2004), for every $\ell \in \{1, \ldots, L\}$,

$$\Pr\left\{ (V_{t+1}, \ldots, V_{t+K}) = \frac{1}{\sum_{k=1}^{K} a_{\ell k}} (a_{\ell 1}, \ldots, a_{\ell K}) \text{ infinitely often} \right\} = 1.$$

Morale:

> *If the data are indeed from an M3 process and if the threshold is suitably selected, one can hope to identify the signatures and their respective probabilities.*

Ansatz of Süveges and Davison (2012)

1) Identify observations that exceed a given threshold u.

2) Select c index sequences $\mathcal{P}_1, \dots, \mathcal{P}_c$ of equal length K around the threshold excesses, defining the clusters of extremes.

3) Regard the vectors $(Y_j : j \in \mathcal{P}_i)$ as (noisy) profiles of an M3 process.

4) For each $i \in \{1, \dots, c\}$, set

$$S_i = \sum_{j \in \mathcal{P}_i} Y_j$$

and regard $V_i = (Y_j / S_i : j \in \mathcal{P}_i) \in \mathcal{S}_K$ as noisy, finite-sample counterparts of the shock type signatures.

5) Model the distribution of the V_i's through a finite Dirichlet mixture. The filter matrix is then recovered from the parameter estimates.

Comments on the Süveges–Davison approach

- ✓ The Dirichlet mixture is intended to model the different signatures.
- ✓ All profiles need to be of the same length and the selection of $\mathcal{P}_i$ involves an iterative k-means clustering algorithm.
- ✓ For the Burlington precipitation data, the method did not yield satisfactory results:
  - (i) it is difficult to identify physically meaningful profiles while avoiding excessive profile length and overlaps;
  - (ii) forcing the profiles to be of the same length leads to clusters that include days without rain (exact zeros);
  - (iii) the Dirichlet mixture does not fit well.

Towards a new model

In the present context, only the total precipitation per cluster matters, e.g., 103mm for the 4-day streak that triggered the 2011 flood.

Cluster identification:

- ✓ Define a *cluster of high precipitation* as the streak of consecutive rainy days containing at least one exceedance above a high threshold u.
- ✓ Each cluster is thus separated from any other by at least one day without rain.
- ✓ Cluster lengths may vary. No cluster of high precipitation contains days with no rain.

Multivariate regular variation

Consider the daily precipitation amounts $\boldsymbol{Y} = (Y_j : j \in \mathcal{C})$ in a cluster $\mathcal{C}$.

Denote the cluster maximum and sum by

$$M = \max(Y_j : j \in \mathcal{C}), \quad S = \sum_{j \in \mathcal{C}} Y_j.$$

Suppose that $\boldsymbol{Y}$ is multivariate regularly varying, i.e.,

$$\frac{\Pr(\|\boldsymbol{Y}\|_\infty > yt, \boldsymbol{Y}/\|\boldsymbol{Y}\|_\infty \in \cdot)}{\Pr(\|\boldsymbol{Y}\|_\infty > t)} \rightsquigarrow y^{-\alpha}\sigma(\cdot)$$

for some $\alpha > 0$ and a probability distribution σ on the unit simplex

$$\{\mathbf{x} \in [0,1]^{|\mathcal{C}|} : \|\boldsymbol{x}\|_\infty = 1\}.$$

Consequences for model building

- ✓ Keep in mind that we only need to model the cluster sum S.
- ✓ Conditionally on $M > u$ for some high threshold u, M and $\boldsymbol{Y}/M$ are roughly independent. Thus also M and

$$P = \frac{M}{\sum_{j \in \mathcal{C}} Y_j} = \frac{M}{S}$$

are approximately independent conditionally on $M > u$.

- ✓ Observe that $S = M \times (1/P)$.

We propose to scale up M with an independent factor $1/P \geq 1$.

The random scale model

1) Identify clusters $\mathcal{C}_1, \ldots, \mathcal{C}_c$ of extreme precipitation based on a choice of threshold u. The cluster lengths are allowed to vary.

2) For each cluster $(Y_j, j \in \mathcal{C}_i)$, compute

$$M_i = \max(Y_j : j \in \mathcal{C}_i)$$

and fit a GPD distribution to $M_1, \ldots, M_c$ by the POT approach, viz.

$$\Pr(M_i - u \leq w | M_i > u) \approx \begin{cases} 1 - (1 + \xi w/\beta)_+^{-1/\xi} & \text{if } \xi \neq 0, \\ 1 - \exp\left(-w/\beta\right) & \text{if } \xi = 0. \end{cases}$$

The random scale model (cont'd)

3) For each cluster, compute the proportions

$$P_i = \frac{M_i}{S_i} = \frac{\max_{j \in \mathcal{P}_i} Y_j}{\sum_{j \in \mathcal{P}_i} Y_j}$$

and note that $1/|\mathcal{C}_i| \leq P_i \leq 1$.

4) Model $P_1, \ldots, P_c$ with a 1-inflated scaled beta distribution, viz.

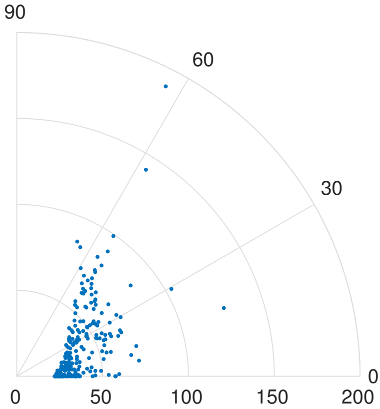
$$\mathcal{IB}(p \mid \omega, \theta, \alpha, \beta) = \omega \, \delta_{\{1\}}(p) + (1 - \omega) \, \mathcal{B}^*_{(\theta,1)} \, (p \mid \alpha, \beta, \theta) \, .$$

Here, $\mathcal{B}^*_{(\theta,1)}(p \mid \alpha, \beta)$ denotes the density of the random variable

$$(1 - \theta)X + \theta,$$

where X has a $\mathcal{B}(\alpha, \beta)$ distribution.

Extreme precipitation clusters for the Burlington data



The clusters were defined as streaks of rainy days separated by at least one day with no rain and containing at least one value above 21.6mm.

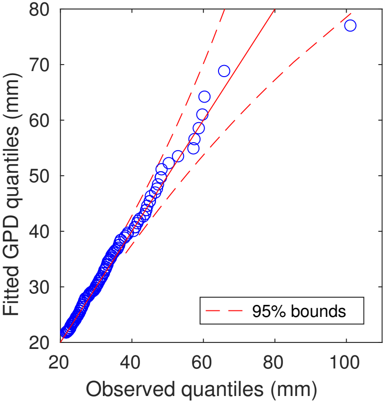
There were $c = 241$ clusters; 51 of length 1 and 20 with 2+ excesses.

Displayed are the pairs

$$(S_i \cos(\Theta_i), S_i \sin(\Theta_i)),$$

where $\Theta_i = \arccos(P_i)$.

Extreme-value model for the cluster maxima



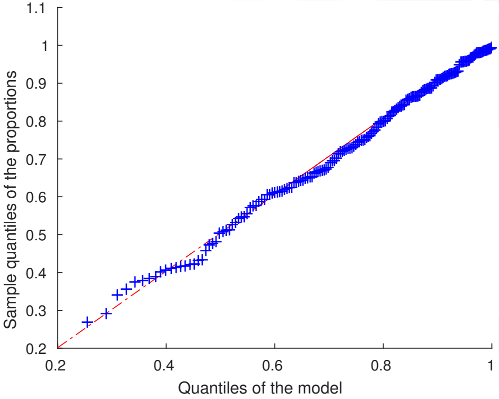
Parameter estimates and 95% credibility intervals

$$\hat{\beta} = 8.6086 \in (7.1258, 10.2472),$$

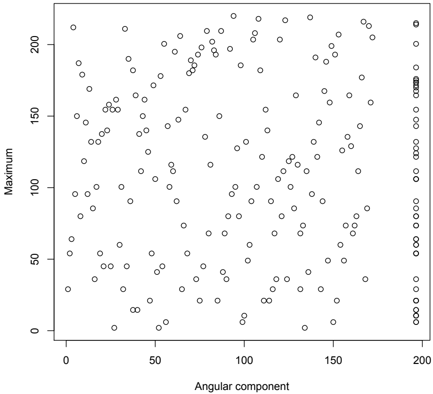
$$\hat{\xi} = 0.0630 \in (-0.0464, 0.2056).$$

They are nearly the same as in the traditional POT approach.

Model for the angular component of the maximum

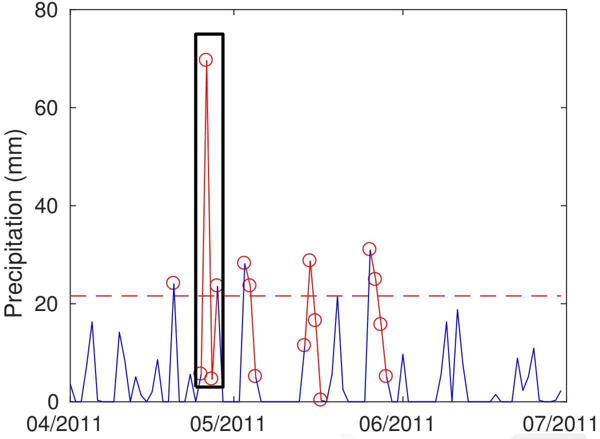


Independence between the components



There is no evidence of dependence between the radial and angular component.

Probability that a cluster sum exceeds 103mm



Probability that a cluster total exceeds 103mm

This probability is estimated by simulation.

- Simulate the cluster maximum exceedance with the following predictive distribution [the prior was $f(\beta,\xi) \propto 1/\beta$]:

$$f(m) = \int\int \mathcal{GP}(m|\beta,\xi) \times f(\beta,\xi \mid y)\, d\beta\, d\xi.$$

- Simulate the angular component of the maximum with the following predictive distribution:

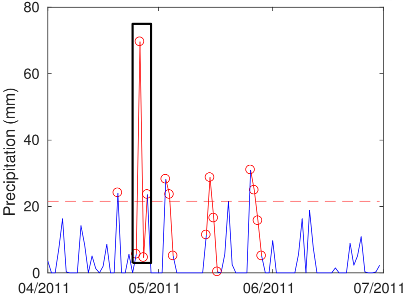
$$f(p) = \int\cdots\int \mathcal{IB}(p \mid \omega,\theta,\alpha,\beta) \times f(\omega,\theta,\alpha,\beta \mid y)\, d\omega\, d\theta\, d\alpha\, d\beta.$$

- Compute the cluster sum: $s = (m+u)/p$.

Probability that a cluster total exceeds 103mm

Through simulation, the probability that a cluster precipitation total exceeds 103mm is

$$\Pr(S > 103\text{mm}) \approx \frac{1}{32} \approx 3.125\%.$$



As more than one cluster can occur per spring, the return period of a cluster exceeding 103mm is less than 32 years.

Probability of a wetter spring than in 2011

During the spring of 2011, a total of 318mm came from clusters of high precipitation.

The number of clusters during the spring can be generated using the Poisson distribution as in the POT model.

The probability that high clusters of precipitation exceed an accumulation of 318mm in a single spring is approximately $1/365$, which corresponds to a return period of 365 years.

Conclusion

In the 2011 flood of the Richelieu River, the clusters of high precipitation are responsible for the 67-day flood.

While clusters of high precipitation had occurred before, it was their frequency that was problematic in 2011.

The modeling of the clusters gives realistic estimates of the return period of the 2011 event, which is 365 years.

Considering the complete period of observation, from 1883 to 2016, the return period of the 2011 event is then estimated at 234 years.

Research funded by

CANADA
RESEARCH
CHAIRS

Fonds de recherche
sur la nature
et les technologies
Québec

CANSSI
INCASS

NSERC
CRSNG

McGill