Modelling clusters of precipitation extremes, with an application to the 2011 Lake Champlain flood

Jonathan Jalbert
École polytechnique, Montréal, Québec, Canada
E-mail: jonathan.jalbert@mail.mcgill.ca
Orla A. Murphy, Christian Genest, and Johanna G. Nešlehová
McGill University, Montréal, Québec, Canada

Summary. Lake Champlain is a natural freshwater lake that straddles the Canada-US border in Eastern North America. In the spring of 2011, its water level was at a record high, and heavy rainfall occurring in several streaks of consecutive days caused massive floods in the surrounding valley and along the Richelieu River (Québec, Canada). Extreme-value analysis of this unprecedented event thus requires a model for clusters of high precipitation. One such modelling strategy is proposed here. It relies on a decomposition of clusters into polar coordinates. An extreme-value distribution is used to model the radial component, while the model for the angular component is based on a 1-inflated mixture of scaled Beta distributions. It is shown that the new model gives a more sensible estimate of the return period of the precipitation that triggered the 2011 Richelieu Valley flood than other existing extreme-value models that take clustering of extremes into account.

Keywords: Clusters of extremes; High precipitation; Peaks-Over-Threshold; Time series extremes

1. Introduction

Lake Champlain is a natural freshwater lake located primarily in the Eastern United States, whose only outlet is the Richelieu River (Québec, Canada). In the spring of 2011, the lake level reached an unprecedented high, leading to a major flood in its surroundings and in the Richelieu Valley. The flood stage was reached on April 14 and continued for over two months, forcing the evacuation of thousands of citizens and causing an estimated 100 million USD in damages (International Joint Commission, 2013). The Richelieu River's 2011 peak discharge of 1542m^3/s was far beyond its mean annual peak discharge of 920m^3/s.

Around 90% of the Richelieu River's streamflow comes from the Lake Champlain watershed, and hence the river's discharge is strongly correlated with the lake's water level. Riboust and Brissette (2015) showed that the lake's water level measured at the gage station in Burlington, Vermont, is a particularly suitable proxy for the Richelieu River discharge. Figure 1 shows the annual maximum water level of Lake Champlain as measured since 1907 at this station. To see whether the 2011 historical high of 31.45m could be predicted from this record, one could fit a generalized extreme-value distribution (GEV) to the annual maxima from the period 1907–2010, spanning 104 years.

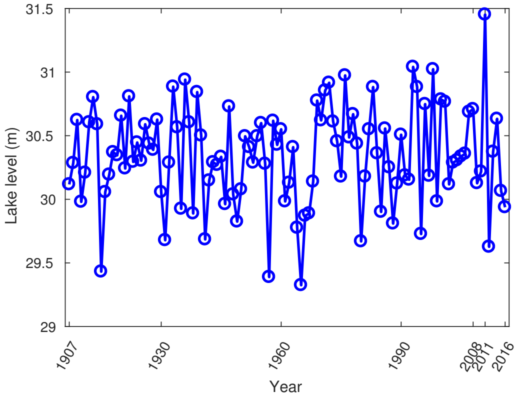

Fig. 1: Lake Champlain annual water level maxima recorded at the Burlington gauge station.

Recall, e.g., from the books of Coles (2001) or Beirlant et al. (2004), that the GEV distribution function $H_{\mu,\sigma,\xi}$ with location parameter $\mu \in \mathbb{R}$, scale parameter $\sigma > 0$ and shape parameter $\xi \in \mathbb{R}$ is given by

$$H_{\mu,\sigma,\xi}(z) = \begin{cases} \exp\Big\{-\Big(1+\xi\,\dfrac{z-\mu}{\sigma}\Big)^{-1/\xi}\Big\} & \text{whenever } 1+\xi(z-\mu)/\sigma > 0 \text{ if } \xi \neq 0, \\ \exp\Big\{-\exp\Big(-\dfrac{z-\mu}{\sigma}\Big)\Big\} & \text{whenever } z \in \mathbb{R} \text{ if } \xi = 0. \end{cases}$$

The maximum likelihood estimates of these parameters are $(\hat{\mu}, \hat{\sigma}, \hat{\xi}) = (30.2,\ 0.392,\ -0.440)$. The fact that $\hat{\xi}$ is negative means that the fitted GEV distribution has a finite upper endpoint, estimated at $\hat{\mu} + \hat{\sigma}/\hat{\xi} = 31.1$m. The 2011 peak water level thus lies outside of the support of the fitted GEV distribution. In other words, this classical GEV analysis deems the 2011 event impossible, a situation which is sometimes referred to as a *Black Swan*.

In view of this simple analysis, it is not surprising that the return period for the 2011 event has proven hard to estimate from historical data. Clearly, it is insufficient to consider only the lake's annual water level maxima. Although daily water levels at the Burlington station are available, this time series is difficult to handle statistically as it shows substantial seasonality and autocorrelation; this is apparent from the bottom panel of Figure 2.

As an alternative, we propose to focus on daily precipitation as measured at Burlington (Vermont) during the critical period of snowmelt in the spring when large precipitation events can trigger a flood. Using a hydrological model, Riboust and Brissette (2016) showed that it is indeed precipitation that has the most critical influence on floods for this watershed. Although the spring freshet in northern watersheds is generally the result of the snowmelt and concurrent precipitation, the snowpack seems to have played a minor role in the 2011 Richelieu Valley spring flood. For example, the largest snowpack was actually recorded during the spring of 2008, and yet it was a normal year for the annual water level maximum (see Figure 1). More importantly, Riboust and Brissette (2016) combined the 2008 snowpack observations with the 2011 precipitation series in their hydrological model and found that

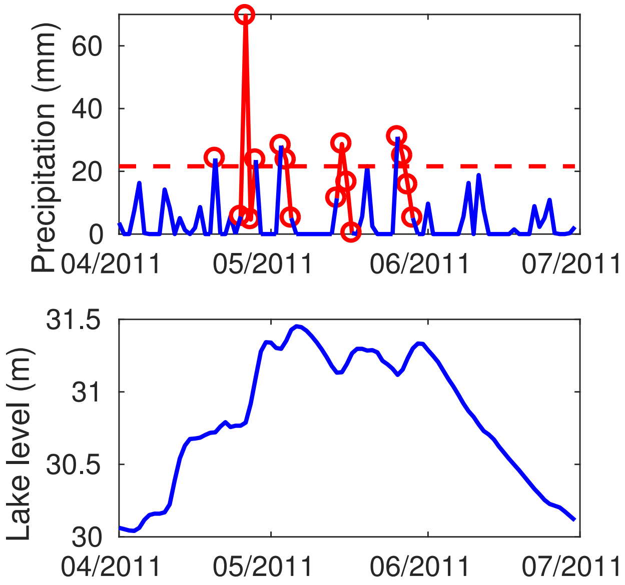

Fig. 2: Daily precipitation at Burlington Airport (VT) and the daily Lake Champlain water levels for the spring of 2011

the simulated flood was not much larger than the actual 2011 flood. They also noted that the spring temperature did not play a major role. Additional evidence in favour of using precipitation as an explanatory variable for the Lake Champlain water level is provided in Figure 3, which shows a boxplot of the spring rainfall accumulations recorded at the Burlington Airport station from 1884 to 2011. The 2011 value is marked by a cross.

The 2011 spring daily precipitation recorded at the Burlington Airport station is shown in the top panel of Figure 2. The red dashed line is the 95th centile of nonzero precipitation for the months of April to June for the years 1884–2016, which constitutes the entire record. It can be seen that in 2011, eight threshold exceedances occurred during this 4-month period. An important stylized fact of this series is that the threshold exceedances mainly occur in clusters. In fact, there were 261 exceedances in the entire series, but only 51 of these were single-day events with no rain on the previous day or on the next. All other exceedances occurred in streaks of consecutive rainy days. In Spring 2011, six clusters were observed; these are identified in red in Figure 2. From the bottom panel of that figure, one can also see that the lake level rose sharply following the 4-day cluster which cumulated a total of 103mm of precipitation, and only began to sink gradually after the heavy spring rains passed.

To assess the flood risk properly, it is thus crucial to take entire clusters of extreme precipitation into account, as the total accumulation per cluster can be much larger than the cluster maximum. The classical Peaks-Over-Threshold (POT) model alone does not suffice to compute the return period of the extreme seasonal rain accumulation observed in

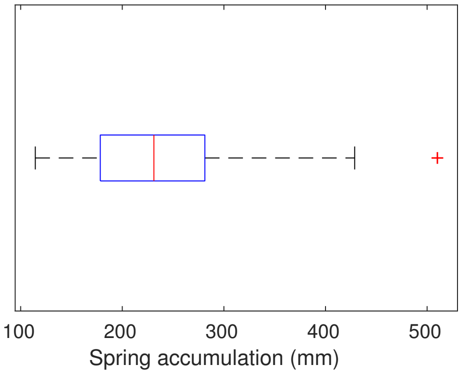

Fig. 3: Spring rainfall accumulations from 1884 to 2011 at Burlington, Vermont

2011, marked by a cross in Figure 3. The latter only considers the frequency and severity of cluster maxima, while in this application, rain accumulation in each cluster is needed.

In this article, we propose a novel extension of the POT model that does account for cluster precipitation totals. In this new model, which we call the random scale model, each cluster maximum is scaled up by an independent random factor. This is done carefully so that the extremal behaviour of the cluster sum is preserved. This approach is simple to implement, and can be justified through multivariate regular variation. As we demonstrate, it works very well for the Burlington precipitation data and leads to a realistic estimate of the return level of the 2011 flood, which other models have hitherto failed to provide.

The rest of the article is organized as follows. In Section 2, we give a detailed description of the Burlington precipitation data, define clusters of high precipitation, and model the severity and frequency of cluster maxima using the classical POT approach. The new random scale model is then presented in Section 3, and justified theoretically through multivariate regular variation. The Burlington precipitation data are then analyzed using this model in Section 4. Various diagnostic plots show that the model fit is good. In the same section, we also present a calculation of the return period of the 2011 spring events. Relationships with the M3 process and alternative models based on the conditional approach of Heffernan and Tawn (2004) are discussed in Section 5. Conclusions are presented in Section 6.

2. Data description and classical POT analysis

2.1. Data description

In this article, we consider daily precipitations in mm for the months of April to June, for the period 1884–2016. A majority of the measurements were recorded at the weather station located at the Burlington Airport, Vermont. The station is still active today and the period of record began in 1940. Another station located in Burlington, 3km from the airport, recorded daily precipitation from 1884 to 1943. The two station records were pooled to provide a longer dataset: the data prior to 1943 come from the Burlington station and

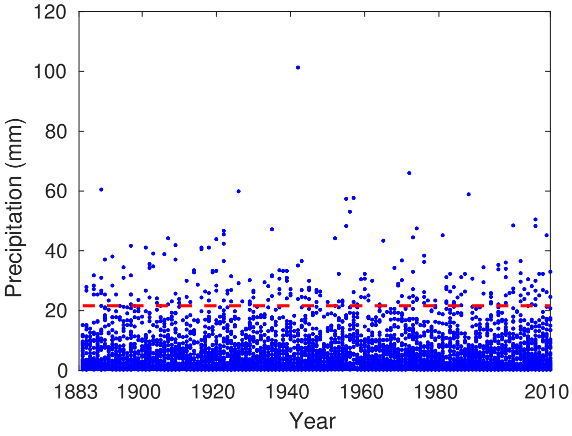

Fig. 4: Daily precipitation series at Burlington Airport, Vermont. The threshold $u = 21.6$mm is indicated by a red dashed line.

the remaining data from the Burlington Airport station. The homogeneity assumption in extreme values of the pooled dataset was checked (not shown). In particular, 1943 is not a change-point in the pooled series of the annual maxima. Both sub-series of annual maxima are stationary according to the Mann–Kendall stationary test (p-value $= 0.53$ and 0.24 for the first and second sub-series, respectively). The entire pooled series of spring precipitation can also be assumed stationary (p-value $= 0.52$). These data are freely available from the NOAA's National Climatic Data Center website (`https://www.ncdc.noaa.gov/`).

2.2. The classical POT model

In what follows, it will be convenient to 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. This definition of clusters differs from the classical runs method (O'Brien, 1987; Smith and Weissman, 1994), which puts threshold exceedances in the same cluster unless they are separated by at least r nonexceedances.

Using the 95% centile of nonzero daily precipitation amounts as the threshold $u = 21.6$mm, 233 exceedances were recorded in the period 1884–2010. The series is displayed in Figure 4, along with the threshold. There were 220 clusters of high precipitation; 51 of these were of length 1 day and 20 contained more than one exceedance. By comparison, the runs method with $r = 1$ identifies 222 clusters. The cluster maxima obtained by the two methods are essentially the same, however; in two instances only, two clusters identified by the runs method ended up being merged into a single cluster of high precipitation.

Let $M_1, \ldots, M_{220}$ be the maxima pertaining to the 220 clusters of high precipitation, and let $M_1 - u, \ldots, M_{220} - u$ be the corresponding excesses above the threshold u. In the classical POT approach, these excesses are modelled with the Generalized Pareto (GP) distribution

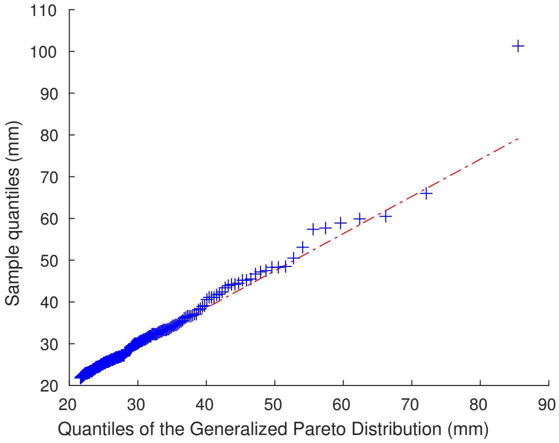

Fig. 5: QQ-plot of the GP distribution fitted to the 220 cluster maxima

with scaling parameter $\sigma > 0$ and shape parameter $\xi \in \mathbb{R}$, i.e., for each $i \in \{1, \ldots, 220\}$,

$$\Pr(M_i - u \leq z \mid M_i > u) \approx \begin{cases} 1 - (1 + \xi z/\sigma)^{-1/\xi} & \text{whenever } 1 + \xi z/\sigma > 0 \text{ if } \xi \neq 0, \\ 1 - \exp(-z/\sigma) & \text{whenever } z \in \mathbb{R} \text{ if } \xi = 0. \end{cases}$$

Assume an improper prior given, for all $\sigma > 0$ and $\xi \in \mathbb{R}$, by $f_{(\sigma,\xi)}(\sigma, \xi) \propto 1/\sigma$. Note that this prior yields a proper posterior as long as the sample size is greater than 2 (Northrop and Attalides, 2016), which is the case here. Bayesian estimates and associated 95% credible intervals for the parameters are then given by

$$\hat{\sigma} = 8.6086 \in (7.1258, 10.2472), \quad \hat{\xi} = 0.0630 \in (-0.0464, 0.2056).$$

The QQ-plot displayed in Figure 5, which is based on the Bayesian point estimates of σ and ξ, suggests an adequate fit.

As for the frequency of clusters of high precipitation, it can be adequately modelled with a homogeneous Poisson point process with intensity $\lambda > 0$. If an improper prior $f_\lambda(\lambda) \propto 1/\lambda$ is assumed, the posterior for λ is then a Gamma distribution, viz.

$$f_{(\lambda|\boldsymbol{Y}=\boldsymbol{y})}(\lambda) = \mathcal{G}(\lambda \mid 220; 11{,}557),$$

where 220 corresponds to the number of cluster maxima and 11,557 corresponds to the number of days of observations (127 years with 91 spring days per year). Here, $\mathcal{G}(\cdot \mid a; b)$ denotes the gamma density function with mean a/b.

Although the fit of the POT model seems adequate, the latter does not suffice to compute the return period for the events that triggered the 2011 flood. For example, the POT model can be used to estimate the return period for the extreme rainfall of 69.6mm that occurred on April 26, 2011 to be 66 years. This may seem low, but it does make good sense given that rainfalls of similar (or even higher) magnitude were already recorded; see Figure 4. However, no flood was ever observed that matches the 2011 flood in magnitude. From Figure 3, it is rather the spring precipitation accumulation T of 510mm that was unusually high in 2011. Because the POT method only models the frequency and severity of cluster maxima, what happens within a cluster of high precipitation is unaccounted for. In particular, one cannot compute the probability of $T > 510$mm from the POT model.

3. Random scale model for cluster accumulation

We now propose a new, simple extension of the POT model to account for the total precipitation amount within each cluster of high precipitation. The idea consists of scaling up each cluster maximum M by an independent random factor in order to model the cluster sum S.

3.1. Derivation of the random scale model

Let $Y_1, Y_2, \ldots$ be a stationary time series of non-negative measurements. In the present context, the values Y_j are daily precipitations. Suppose that n clusters of high precipitation, say $\mathcal{C}_1, \ldots, \mathcal{C}_n$, were identified using some high threshold u. For each $i \in \{1, \ldots, n\}$, let $\boldsymbol{Y}_i = (Y_j : j \in \mathcal{C}_i)$ be the vector of daily precipitation amounts corresponding to cluster $\mathcal{C}_i$. Within this cluster, the distribution of the cluster sum

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

could be deduced from a model for the entire vector $\boldsymbol{Y}_i$. This is cumbersome, however, particularly because the length L_i of $\boldsymbol{Y}_i$ depends on i, and also because $Y_j > u$ for at least one, but not necessarily all, $j \in \mathcal{C}_i$. This means that one could not resort, e.g., to a multivariate extreme-value model such as the tail model of Ledford and Tawn (1996).

Luckily, a full model for $\boldsymbol{Y}_i$ is not needed here. Instead, for each $i \in \{1, \ldots, n\}$, assume that the vector $\boldsymbol{Y}_i$ of length $L_i = \ell_i$ is multivariate regularly varying (Resnick, 1987). This implies that if $\|\cdot\|_\infty$ denotes the max-norm, there exists, for each $i \in \{1, \ldots, n\}$, a real $\eta > 0$ and a probability distribution ς on the unit simplex $\{\mathbf{x} \in [0,1]^{\ell_i} : \|\boldsymbol{x}\|_\infty = 1\}$ such that

$$\frac{\Pr(\|\boldsymbol{Y}_i\|_\infty > yt, \boldsymbol{Y}_i/\|\boldsymbol{Y}_i\|_\infty \in \cdot)}{\Pr(\|\boldsymbol{Y}_i\|_\infty > t)} \rightsquigarrow y^{-\eta}\varsigma(\cdot) \tag{1}$$

for all $y > 0$ as $t \to \infty$, where $\rightsquigarrow$ denotes weak convergence. By Corollary 5.18 in Resnick (1987), $\boldsymbol{Y}_i$ is in the domain of attraction of a multivariate extreme-value distribution. Furthermore, Eq. (1) implies that the cluster maximum $M_i = \|\boldsymbol{Y}_i\|_\infty$ is in the domain of attraction of the Fréchet distribution with parameter η. More interestingly, if $M_i > u$ for some high threshold u, M_i and $\boldsymbol{Y}_i/M_i$ are nearly independent. Thus conditionally on $M_i > u$, one also has approximate independence between M_i and

$$P_i = \frac{M_i}{\sum_{j \in \mathcal{C}_i} Y_j} = \frac{M_i}{S_i}\,.$$

Now suppose that the threshold u is high enough that the independence between M_i and P_i can be assumed to hold, at least approximately. Because $\mathcal{C}_i$ is a cluster of high precipitation, the condition $M_i > u$ is automatically satisfied. Thus, upon writing

$$S_i = M_i \times (1/P_i), \tag{2}$$

we propose to model S_i by scaling up the cluster maximum M_i with an independent multiplicative factor $1/P_i \geq 1$.

Remark 1. Let $\boldsymbol{Y}$ in $\mathbb{R}^\ell$ be a multivariate regularly varying random vector with non-negative components. Set $M = \max(Y_1, \ldots, Y_\ell)$, and $S = Y_1 + \cdots + Y_\ell$. Then there exists a Radon measure Q on $\mathbb{R}^\ell \setminus \{\mathbf{0}\}$ such that $\Pr(\boldsymbol{Y}/t \in \cdot)/\Pr(M > t) \Rightarrow Q$ as $t \to \infty$, where $\Rightarrow$ refers to vague convergence. As shown, e.g., by Jessen and Mikosch (2006),

$$\lim_{t\to\infty} \frac{\Pr(S > t)}{\Pr(M > t)} = \kappa \equiv Q\{(x_1, \ldots, x_\ell) \in (0,\infty)^\ell : x_1 + \cdots + x_\ell > 1\}.$$

It may happen that $\kappa = 0$ but if not, S and M are then tail equivalent; in fact, they are both in the domain of attraction of the Fréchet distribution with the same shape parameter. This tail equivalence between S and M is preserved in the random scale model $S = M \times (1/P)$, where P and M are independent and $P \geq 1$, provided that M is in the domain of attraction of the Fréchet distribution with shape parameter η and $\mathrm{E}(1/P^{\eta+\epsilon}) < \infty$ for some $\epsilon > 0$. This result, which follows from Breiman's Lemma (Jessen and Mikosch, 2006, Lemma 4.2), holds in particular when P is bounded above.

3.2. Choice of distributions

In order to model cluster sums by scaling up cluster maxima through Eq. (2), one needs to choose distributions for M_i and $1/P_i$ for each $i \in \{1, \ldots, n\}$. As already seen in Section 2, the POT model can be used to select a conditional distribution of M_i given $M_i > u$.

To propose a suitable model for P_i, note first that the distribution of P_i depends on the cluster length L_i. When $L_i = 1$, one has $P_i \equiv 1$, i.e., the distribution of P_i is a Dirac mass at 1, denoted $\delta_{\{1\}}$. If $L_i > 1$, one has $S_i \leq L_i M_i$, and hence $P_i \in [1/L_i, 1]$. Thus given $L_i = \ell$, a natural choice for the density of P_i would be defined, for all $p \in (0, 1)$, by

$$f_{(P_i|L_i=\ell)}(p) = \begin{cases} \delta_{\{1\}}(p) & \text{if } \ell = 1, \\ \mathcal{B}^*_{(1/\ell,1)}\left(p \mid \alpha_{\ell-1}, \beta_{\ell-1}\right) & \text{if } \ell \in \{2, 3, \ldots\}. \end{cases}$$

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. This approach, however, requires a model for the cluster length.

We propose a simpler solution in that we use the scaled Beta distribution $\mathcal{B}^*_{(\theta,1)}(p \mid \alpha, \beta)$, but make $\theta \in (0, 1)$ an additional parameter in the model. To account for clusters of length 1, we further inflate this scaled Beta distribution by placing mass $\omega \in (0, 1)$ at 1. The resulting 1-inflated scaled Beta density for P_i is then given, for all $p \in (0, 1]$, by

$$\mathcal{IB}(p \mid \omega, \theta, \alpha, \beta) = \omega\, \delta_{\{1\}}(p) + (1 - \omega)\, \mathcal{B}^*_{(\theta,1)}\left(p \mid \alpha, \beta\right). \tag{3}$$

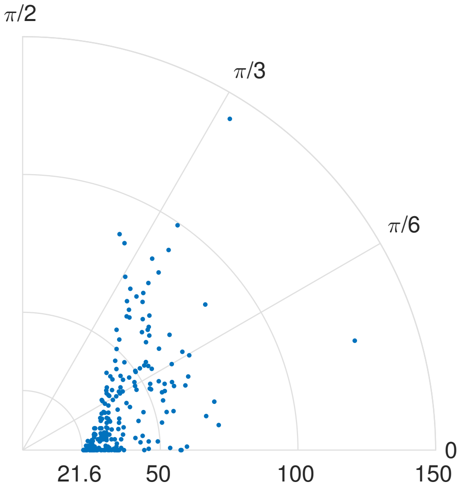

Fig. 6: Polar representation of the cluster sum as the radius and the angle between the horizontal axis corresponds to the proportion of the cluster maximum in the sum.

This proposal effectively pools together all clusters of length $\ell \geq 2$. As will be seen below, it is sufficiently rich to capture the key features of the Burlington Airport precipitation data.

4. Application to the Burlington Airport precipitation data

In this section, we apply the random scale model introduced in Section 3 to the precipitation series measured at Burlington (Vermont). We first examine the fit in Section 4.1 and then use it to compute the return period of the 2011 flood in Section 4.2.

4.1. Fitting the random scale model

The choice of threshold u and the resulting clusters of high precipitation remain as described in Section 2. As a preliminary step, the pairs (S_i, P_i) are visualized in Figure 6. Displayed are the points $(S_i \cos(\Theta_i), S_i \sin(\Theta_i))$, where $\Theta_i = \arccos(P_i)$ for every $i \in \{1, \ldots, 220\}$. When the cluster length $L_i = 1$, one has $\Theta_i = 0$ so that the point lies on the x-axis. In contrast, a large angle Θ_i corresponds to $M_i \ll S_i$. Because $P_i \geq 1/L_i$, a large value of Θ_i also indicates a large value of L_i. Such a cluster would thus typically include several days of heavy precipitation. In this data set, there were 48, 65, 44, 16 clusters of length 1, 2, 3, 4, respectively; the largest cluster was of size 14.

The left panel of Figure 7 shows the rankplot of the pairs $(M_1, P_1), \ldots, (M_{220}, P_{220})$. One can discern ties in the data, due in part to the fact that the angular component equals 1 for the 48 clusters of size 1. As no association is apparent, the assumption of independence between M_i and P_i seems appropriate at threshold level $u = 21.6$mm. This conclusion is further supported by a p-value of 0.45 based on the test of independence for data with ties proposed in Genest et al. (2017).

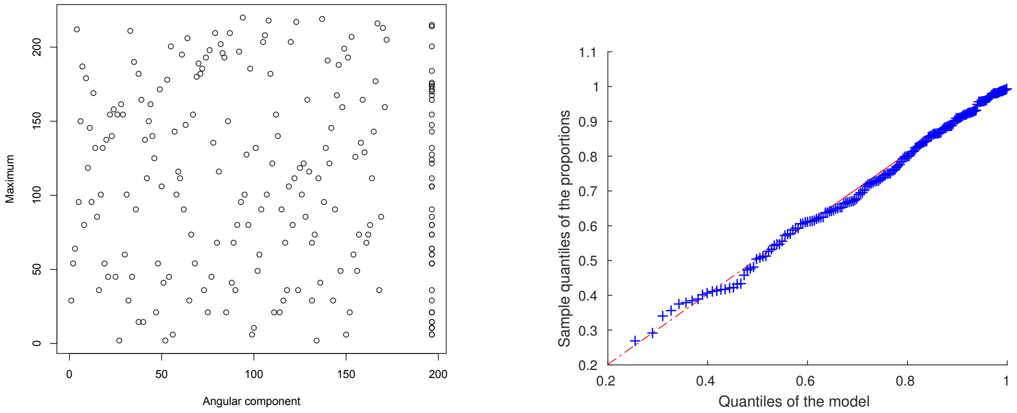

Fig. 7: Rank plot of the cluster maxima and scaling factors (left panel) and QQ-plot of the 1-inflated scaled Beta distribution fitted to $P_1, \ldots, P_{220}$ (right panel).

Next, the 1-inflated scaled Beta distribution given in Eq. (3) was fitted. To this end, it was first reparametrized by setting $\nu = \alpha/(\alpha + \beta)$ and $\gamma = \alpha + \beta$, so that the following non-informative priors can be used:

$$\begin{aligned} &f_\omega(\omega) \propto \omega^{-1}(1-\omega)^{-1}, \text{ for } \omega \in (0,1); \qquad && f_\theta(\theta) = 1, \text{ for } \theta \in (0,1);\\ &f_\nu(\nu) \propto 1, \text{ for } \nu \in (0,1); && f_\gamma(\gamma) \propto 1/\gamma, \text{ for } \gamma \in (0,\infty). \end{aligned}$$

The posterior of the lower bound θ is insensitive to this choice of prior (not shown). The QQ-plot of the fitted 1-inflated scaled Beta distribution is displayed in the right panel of Figure 7. It suggests a good fit, particularly in the lower tail. This is important because low values of P typically correspond to long clusters with several days of heavy rain.

4.2. Computation of the return period of the 2011 flood

In the Lake Champlain watershed, the spring accumulation of precipitation is the main contributing factor in flooding. As mentioned before and illustrated in Figure 3, the value of T observed in 2011 was very high: 510mm. Because of the presence of extreme rainfall, we propose to decompose T into the accumulation Z of non-extreme rainfall and the accumulation W of precipitation from the clusters of high precipitation, i.e., $T = Z + W$.

For any given year $k \in \{1, \ldots, 127\}$ between 1884 and 2010, the observed value Z_k is simply the total precipitation accumulation in year k minus the accumulation W_k of rain in clusters of high precipitation in that same year. Because Z_k is a sum of variables, none of which is extreme, and given that the entire series is stationary, it seems reasonable to assume that $Z_1, \ldots, Z_{127}$ form a random sample from the Gaussian distribution. This assumption was validated using a Shapiro–Wilks normality test (p-value ≈ 0.67). The predictive distribution of the accumulation Z of non-extreme rainfall was found to be $\mathcal{N}(\zeta, \rho^2)$ with

$\zeta = 233.698 \in (221.487, 245.91)$ and $\rho = 69.5384 \in (61.9103, 79.3274)$. These Bayesian estimates were obtained using Jeffreys' improper prior defined, for all $\rho > 0$, by $f_{(\zeta,\rho)} \propto \rho$.

Using the random scaling model, the distribution of W can be approximated as follows by a Monte Carlo simulation. First, the number N of clusters of high precipitation in a given spring is drawn from the predictive distribution of the Poisson point process with intensity λ obtained from the POT model in Section 2. The latter is given, for all $n \in \mathbb{N}$, by

$$f_{(N|\boldsymbol{Y}=y)}(n) = \int_0^\infty \mathcal{P}(n \mid 91\,\lambda) \times f_{(\lambda|\boldsymbol{Y}=\boldsymbol{y})}(\lambda)\; d\lambda. \tag{4}$$

This distribution models the number of cluster maxima in a period of 91 days, i.e., the months of April–June which constitute the spring season. Second, given the number $N = n$ of clusters of high precipitation, the cluster maxima $M_1, \ldots, M_n$ are drawn independently from the predictive distribution obtained from the POT model given, for all $z > 0$, by

$$f_{(M-u|\boldsymbol{Y}=\boldsymbol{y},N=n)}(z) = \int_{-\infty}^\infty \int_0^\infty \mathcal{GP}(z \mid \sigma, \xi) \times f_{[(\sigma,\xi)|\boldsymbol{Y}=\boldsymbol{y}]}(\sigma, \xi) d\sigma d\xi. \tag{5}$$

Third, the proportions $P_1, \ldots, P_n$ are drawn independently from the predictive distribution defined, for all $p > 0$, by

$$f_{(P|\boldsymbol{Y}=\boldsymbol{y},N=n)}(p) = \int_0^1 \int_0^1 \int_0^1 \int_0^\infty \mathcal{IB}(p \mid \omega, \theta, \nu, \gamma) \times f_{[(\omega,\theta,\nu,\gamma)|\boldsymbol{Y}=\boldsymbol{y}]}(\omega, \theta, \nu, \gamma) d\gamma d\nu d\theta d\omega. \tag{6}$$

Then $W = M_1/P_1 + \cdots + M_n/P_n$ is the total amount of rain within clusters of high precipitation. This procedure is summarized in Algorithm 1.

Algorithm 1 Generating a spring rainfall accumulation from clusters of high precipitation

```
1) Draw the number $N = n$ of clusters of high precipitation from distribution (4).
2) Draw the cluster maxima $M_1 - u, \ldots, M_n - u$ from distribution (5).
3) Draw the proportions $P_1, \ldots, P_n$ of from distribution (6).
4) Draw the accumulation of precipitation pertaining to clusters of high precipitation:
       $W = M_1/P_1 + \cdots + M_n/P_n$.
5) Draw the accumulation $Z$ of non-extreme rainfall from its predictive distribution.
6) Compute the total spring accumulation $T = Z + W$.
```

To estimate the probability that T surpasses the value observed in Spring 2011, viz.

$$\Pr(T > 510\text{mm}), \tag{7}$$

The predictive distribution of R is displayed in the left panel of Figure 8 and the corresponding one-sided 95% credible interval is $[231, \infty)$. Thus while the heavy rain of 69.6mm recorded on April 26, 2011 is not particularly unusual, as seen in Section 2, the total Spring 2011 rainfall accumulation does qualify as a rare event according to the random scale model.

Spring 2011 was also atypical in that 5 clusters of high precipitation were recorded and the total rain accumulation in these clusters was 318mm. Based on the random scale model, the

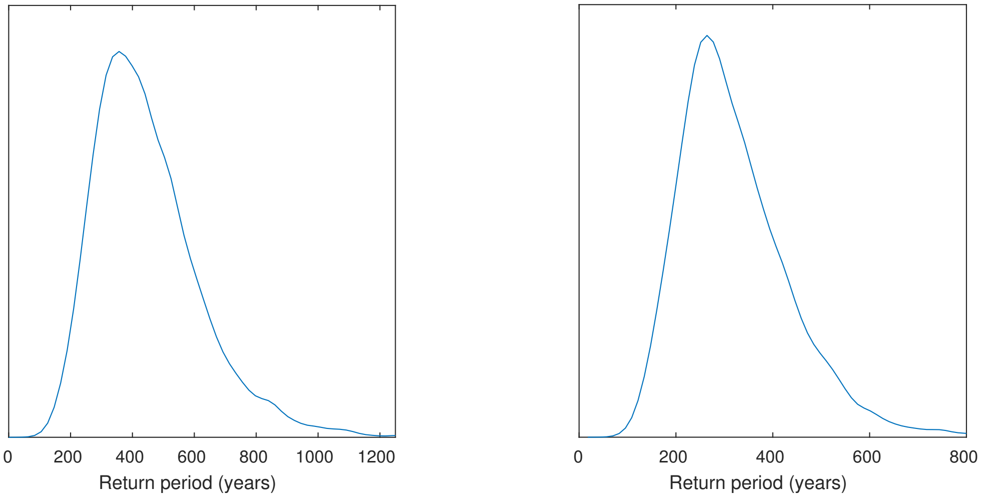

Fig. 8: Predictive distribution of the return period estimated with the precipitation data prior of 2011 (left panel) and with all the data (right panel).

probability of observing 5 or more clusters in a given spring is 3.62×10^{-2}; the corresponding Bayesian estimate of the return period is 33 years, which is not so high. However, $\Pr(W > 318\text{mm}) \approx 3.13 \times 10^{-3}$, which corresponds to a return period of 302 years.

Remark 2. Note that it is also possible to sample directly from the observed proportions $P_1, \ldots, P_n$ in Algorithm 1 instead of modelling them with a 1-inflated scaled Beta distribution, which may make particularly good sense when a large data set is available. In the present application, however, the parametric and nonparametric approaches lead to practically the same predictive distribution of the return period.

If we use all the data from 1884 to 2016, i.e., 133 years, the fit of the random scale model remains equally good. The probability that the spring accumulation of precipitation exceeds the 2011 level of 510mm is then of the order of 3.55×10^{-3}, based on a simulation using Algorithm 1. Thus even when the 2011 event is included, the return period of that year's spring accumulation is still very large, estimated at 320 years; the one-sided 95% credible interval is $[167, \infty)$. The predictive distribution of the return period estimated with all the data is shown in the right panel of Figure 8. The estimated return level of a spring accumulation corresponding to the 100-year return period is $461\text{mm} \in (451, 470)$.

5. Connection with existing models

In this section, we briefly review existing approaches for the modelling of clusters of extreme events and explain why they appear less suitable for the Burlington precipitation series than the random scale model advocated here. In particular, we focus on the M3-Dirichlet approach of Süveges and Davison (2012) in Section 5.1, and on the conditional exceedance model of Heffernan and Tawn (2004) in Section 5.2.

5.1. The M3-Dirichlet model

The article by Süveges and Davison (2012) studies a disastrous rainfall that occurred in coastal Venezuela in December 1999. Similar to the Burlington precipitation data, standard extremal models fail to account for this catastrophe because of the inadequate treatment of clusters of heavy precipitation. To model such clusters, Süveges and Davison (2012) propose to rely on the moving maximum process (M3) due to Smith and Weissman (1996).

Recall that a univariate stationary time series $(Y_i : i \in \mathbb{Z})$ is said to be an M3 process if, for each $i \in \mathbb{Z}$, one can write $Y_i = \max_{k\in\mathbb{Z}} \max_{\ell\in\mathbb{N}} a_{\ell,k} X_{\ell,i-k}$ in terms of mutually independent unit Fréchet random variables $(X_{\ell,k} : \ell \in \mathbb{N}, k \in \mathbb{Z})$ and a so-called filter matrix $A = (a_{\ell,k} : \ell \in \mathbb{N}, k \in \mathbb{Z})$ of non-negative constants summing up to 1. It is typically assumed that $a_{\ell,k} > 0$ only when $\ell \in \{1, \ldots, L\}$ and $k \in \{1, \ldots, K\}$. When normalized by the sum of its components, viz. $(c_{\ell,1}, \ldots, c_{\ell,K}) = (a_{\ell,1}, \ldots, a_{\ell,K})/(a_{\ell,1} + \cdots + a_{\ell,K})$, the ℓth row of A is referred to as the signature of the ℓth cluster type.

Süveges and Davison (2012) argue that when the threshold u is sufficiently high, any cluster $(Y_j : j \in \mathcal{C})$ of extremes, once normalized by the sum of its components, viz.

$$\boldsymbol{W} = (W_j : j \in \mathcal{C}) = \frac{1}{\sum_{k\in\mathcal{C}} Y_k} \times (Y_j : j \in \mathcal{C}), \tag{8}$$

corresponds to a noisy version of one of the signatures. This intuition is rooted in a result of Zhang and Smith (2004) that if $(Y_i : i \in \mathbb{Z})$ is an M3 process, then for each $\ell \in \{1, \ldots, L\}$,

$$\Pr\left\{ \frac{(Y_{t+1}, \ldots, Y_{t+K})}{Y_{t+1} + \cdots + Y_{t+K}} = (c_{\ell,1}, \ldots, c_{\ell,K}) \text{ infinitely often} \right\} = 1.$$

Therefore, Süveges and Davison (2012) propose (i) to transform the series so that its marginals are approximately unit Fréchet; (ii) to identify clusters of extremes of a fixed length K through an elaborate algorithm; and (iii) to model the normalized cluster profiles $\boldsymbol{W}$ with a finite Dirichlet mixture. The number of mixing components is larger or equal to L (a single signature could require more than one Dirichlet component) and an estimate of the filter matrix A is then obtained from the fitted Dirichlet parameters.

To contrast the M3-Dirichlet model with the random scale model proposed here, consider an arbitrary cluster $(Y_j : j \in \mathcal{C})$ of high precipitation and the corresponding normalized cluster profile $\boldsymbol{W}$ defined in Eq. (8). In the M3-Dirichlet approach, the entire vector $\boldsymbol{W}$ is modeled; this requires all clusters to have the same fixed length. In contrast, the random scale model allows for variable cluster length and focuses exclusively on the variable $P = \max(W_j : j \in \mathcal{C}) \in (0, 1]$, which is a lot easier to model than the vector $\boldsymbol{W}$. Moreover, in order to model the total spring rain accumulation, the M3-Dirichlet model would need to be extended to account for the cluster sum, and this does not seem straightforward.

When applying the M3-Dirichlet model to the Burlington precipitation data, the requirement of a fixed cluster length proved to be a serious obstacle. The algorithm from Süveges and Davison (2012, Section 2.3) identified unreasonably long clusters, often containing days with no rain (exact zeros); this phenomenon is a result of the fact that the cluster length is fixed and that, at the same time, overlaps between clusters must be avoided. Furthermore, the finite Dirichlet mixture did not fit the normalized profiles $\boldsymbol{W}$ well.

5.2. The conditional exceedance model

Another model that could be used for the Burlington precipitation data is the conditional exceedance model of Heffernan and Tawn (2004), along with the modifications later proposed by Keef et al. (2013a,b). Let $\boldsymbol{Y}$ be a d-dimensional random vector with Laplace margins and let $\boldsymbol{Y}_{-i}$ denote its $(d-1)$-dimensional margin obtained by leaving out the ith component. The conditional exceedance model accounts for the distribution of $\boldsymbol{Y}_{-i}$ given $Y_i > u$ for some high threshold u, and is meaningful even under asymptotic independence scenarios. Of particular relevance for the application considered here is the work of Keef et al. (2009), where this approach is used to model temporal dependence, i.e., the distribution of $Y_{t+\tau}$ conditionally on $Y_t > u$, for some integer lag τ. This model was recently applied in Winter and Tawn (2016) to simulate clusters of extreme values.

In the Burlington precipitation series, independence appears to hold at any lag $\tau > 1$ and asymptotically when $\tau = 1$ for the chosen threshold. We thus chose $\tau = 1$ and used the runs method with $r = 1$ to identify the clusters. Each cluster was further enlarged by one day at each end. These clusters were then used to fit the conditional exceedance model. The return period of a spring precipitation accumulation of 510mm was then computed by Monte Carlo using 100,000 simulated spring scenarios. As in Section 4.2, a normal distribution was used to model the sum of precipitation occurring outside of the clusters of extreme precipitation.

To simulate a cluster of extreme precipitation, the excess $Y_i - u$ was first simulated independently, and the conditional exceedance model was used to simulate the following day; the simulation continued until an observation dropped below u. In order to model the day preceding a cluster of extreme rainfall, we fitted a model with lag -1 only to the first exceedance of a cluster.

Although the combined model appears to capture the observed cluster lengths and totals, its extrapolation is conservative. The return period of a spring accumulation of 510mm is 1051 years, which is much longer than the estimate derived from the random scale model. One possible explanation for the conservative return level estimate using the conditional exceedance model is the fact that estimation of one of the model parameters still requires the assumption of Gaussian residuals via constrained maximum likelihood. For the Burlington precipitation data, the residuals are clearly not normal. In addition, as we are only using a lag of 1, this model does not have the flexibility required to simulate a cluster in which two threshold exceedances are separated by a day of moderate precipitation below the threshold.

6. Conclusion

In this paper, we used the precipitation recorded at Burlington, Vermont, to estimate the return period of the 2011 flood in the Lake Champlain watershed. For adequate estimation, clustering of high precipitation needed to be taken into account. To this end, we proposed an extension of the classical POT model called the random scale model, in which the cluster maximum is scaled up by an independent random factor. In particular, this allows to model spring precipitation accumulation. Although the approach is tailored here for precipitation data, it could be used in other situations where cluster totals are of interest.

The random scale model was seen to fit the Burlington precipitation data well. Through Monte Carlo simulations and using the whole observation period, it led to a high, yet realistic, estimate of 320 years for the return period of the 2011 spring accumulation of 510mm. Assuming stationarity of the precipitation series, the probability of such an event occurring again thus remains small. In fact, the estimated 100-year return level of a spring accumulation is 446mm, which is 70mm less than the value observed in 2011. The estimate of the return period of the 2011 flood, provided here for the first time, should help the International Joint Commission on the Lake Champlain and the Richelieu River in identifying the causes and impacts of flooding, and in developing appropriate mitigation solutions and recommendations.

Acknowledgements

The precipitation data used in this study were acquired from NOAA's National Climatic Data Center. Thanks are due to the NOAA for freely providing them. Funding in partial support of this work was provided by the Canada Research Chairs Program, the Natural Sciences and Engineering Research Council (RGPIN/39476–2011, RGPIN/06801–2015), the Canadian Statistical Sciences Institute, and the Fonds de recherche du Québec – Nature et technologies (2015–PR–183236), as well as by the MITACS Elevate Program.

References

Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J. (2004) *Statistics of Extremes: Theory and Applications.* New York: Wiley.

Coles, S. (2001) *An Introduction to Statistical Modeling of Extreme Values.* London: Springer.

Genest, C., Nešlehová, J. G., Rémillard, B. and Murphy, O. (2017) Test of independence for sparse frequency tables and beyond. *In preparation.*

Heffernan, J. E. and Tawn, J. A. (2004) A conditional approach for multivariate extreme values (with discussion). *J. Roy. Statist. Soc. Ser. B*, **66**, 497–546.

International Joint Commission (2013) The Identification of Measures to Mitigate Flooding and the Impacts of Flooding of Lake Champlain and Richelieu River. *Technical Report*, International Joint Commission, Ottawa, Canada and Washington, DC.

Jessen, A.H. and Mikosch, T. (2006) Regularly varying functions. *Publ. Inst. Math. (Beograd) (N.S.)*, **80 (94)**, 171–192.

Keef, C., Papastathopoulos, I. and Tawn, J.A. (2013a) Estimation of the conditional distribution of a multivariate variable given that one of its components is large: Additional constraints for the Heffernan and Tawn model. *J. Multivariate Anal.*, **115**, 396–404.

Keef, C., Svensson, C. and Tawn, J.A. (2009) Spatial dependence in extreme river flows and precipitation for Great Britain. *Journal of Hydrology*, **378**, 240–252.

Keef, C., Tawn, J.A. and Lamb, R. (2013b) Estimating the probability of widespread flood events. *Environmetrics*, **24**, 13–21.

Ledford, A.W. and Tawn, J.A. (1996) Statistics for near independence in multivariate extreme values. *Biometrika*, **83**, 169–187.

Northrop, P. and Attalides, N. (2016) Posterior propriety in bayesian extreme value analyses using reference priors. *Statistica Sinica*, **26**.

O'Brien, G.L. (1987) Extreme values for stationary and Markov sequences. *Ann. Probab.*, **15**, 281–291.

Resnick, S.I. (1987) *Extreme Values, Regular Variation and Point Processes.* New York: Springer.

Riboust, P. and Brissette, F. (2015) Climate Change Impacts and Uncertainties on Spring Flooding of Lake Champlain and the Richelieu River. *Journal of the American Water Resources Association*, **51**, 776–793.

— (2016) Analysis of Lake Champlain/Richelieu River's historical 2011 flood. *Canadian Water Resources Journal*, **41**, 174–185.

Smith, R.L. and Weissman, I. (1994) Estimating the extremal index. *J. Roy. Statist. Soc. Ser. B*, **56**, 515–528.

— (1996) Characterization and estimation of the multivariate extremal index.

Süveges, M. and Davison, A.C. (2012) A case study of a "dragon-king": The 1999 Venezuelan catastrophe. *The European Physical Journal Special Topics*, **205**, 131–146.

Winter, H.C. and Tawn, J.A. (2016) Modelling heatwaves in central France: A case-study in extremal dependence. *J. R. Stat. Soc. Ser. C Appl. Stat.*, **65**, 345–365.

Zhang, Z. and Smith, R.L. (2004) The behavior of multivariate maxima of moving maxima processes. *Journal of Applied Probability*, **41**, 1113–1123.