Validation of the Extremal Behaviour of Rainfall Generators

Author: **Lucas REYMOND**
Supervisors: **Linda MHALLA** and **Erwan KOCH**

UNIL | Université de Lausanne
Expertise Center for Climate Extremes (ECCE)

1. Motivation

Flooding is one of the primary climate-related threats to people's livelihoods and development opportunities globally. In 2022, 1.8 billion people (23% of the world population) were exposed to 1-in-100-year floods. The frequency and severity of extreme rainfall events are increasing due to climate change, presenting significant challenges for hydro-meteorological modeling.
Assessing the risk of extreme rainfall events is crucial for several reasons:

- **Public Safety:** Enables timely emergency responses to flash floods and landslides.
- **Infrastructure:** Informs the design of resilient infrastructure, reducing potential damages.
- **Economics:** Helps manage financial losses through improved insurance schemes and disaster funds.
- **Urban Planning:** Guides land use and prevents construction in high-risk areas.

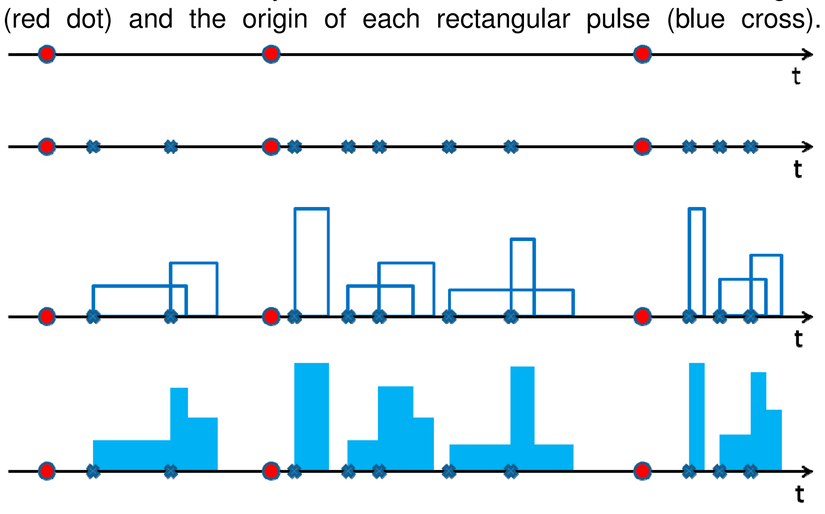
Observational data are essential for studying extreme events like floods, as they provide the most accurate climate information. However, due to the rarity of such events, meteorological records are often insufficient. This is one of the reasons why rainfall generators are used. They are stochastic models that can generate artificial time series of rain, that, hopefully, have the same statistical properties as the real observations. By enabling us to generate an unlimited number of time series with arbitrary length, they allow us to generate some rare extreme events. These simulations allow us to study rare extreme events and their impacts. A natural question to ask is: do the extreme events generated by a rainfall generator have the same statistical properties as the extreme events that might happen in nature?

This work aims to provide a methodological contribution for hydro-meteorologists, by bringing existing extreme-value theory tools to the field of hydrology to validate the extremal behaviour of rainfall generators.

2. Model and Data

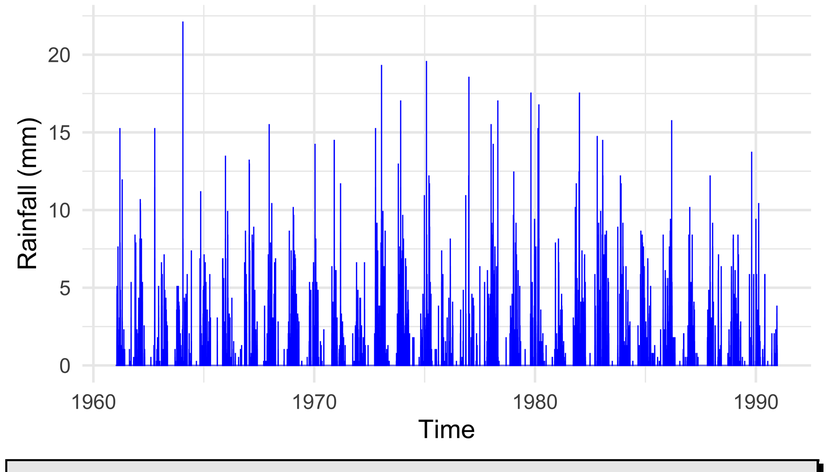
The Model

To validate the extremal behaviour of a rainfall generator, we require both observed and simulated data. We use the AWE-GEN (Advanced WEather GENerator) model to simulate rainfall events using a Poisson-cluster approach. A random number of rectangular pulses (empty blue rectangle on the following figure [1]) with random intensity and duration are associated with each storm, with the total rainfall (blue area) being the sum of active pulses. AWE-GEN employs the Neyman-Scott Rectangular Pulse (NSRP) model, which randomly determines the time between storm origins (red dot) and the origin of each rectangular pulse (blue cross).



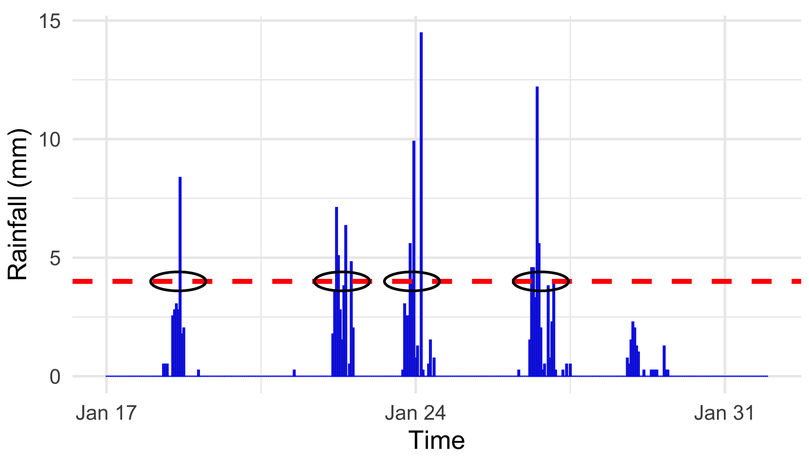
The Data

We calibrate AWE-GEN by estimating the NSRP model parameters from an observed dataset of hourly rainfall over 30 years (as displayed in the following figure) from San Francisco Airport in California, USA. We then run 100 simulations of hourly rainfall, each covering a 30-year period, that should replicate the statistical properties of the observed data. To ensure stationarity, we focus on meteorological winters (December 1 to February 29). Thus, our observational dataset includes hourly measurements for 29 complete winters, while our simulated dataset includes hourly data for 2900 winters.



3. Clustering of Extreme Values

Zooming in allows us to observe our dataset in detail. The figure below displays the observed dataset over a period of two weeks, highlighting clusters of extreme values.



Black ellipses indicate clusters of rainfall events exceeding $4\ mmh^{-1}$. These clusters are intuitively grouped rather than formally defined.

Indeed, in stationary data, storms can cause exceedances of a threshold for several hours, forming clusters. We study the mean cluster size using the extremal index θ, ranging from 0 to 1. An extremal index $\theta < 1$ indicates temporal dependence at asymptotically high levels. [2] interpreted the extremal index as the reciprocal of the mean cluster size, $\theta =$ (limiting mean cluster size)$^{-1}$, where limiting is interpreted in the sense of clusters of exceedances of increasingly high threshold.

To estimate the extremal index, we use the estimator from [3], defined as:

$$\hat{\theta}_n^*(u) = \frac{2\left\{\sum_{i=1}^{N-1}(T_i-1)\right\}^2}{(N-1)\sum_{i=1}^{N-1}(T_i-1)(T_i-2)}, \quad (1)$$

where n is the size of the time series, N is the number of exceedances of the threshold u, and T_i are the times separating two consecutive exceedances of u.

4. Currently Used Method

To study the extremal behaviour of a rainfall time series, the state-of-the-art method involves dividing the time series into equal-length blocks, fitting a generalized extreme value (GEV) distribution to the block maxima, and computing extreme quantiles.

GEV Approach

The GEV approach examines the distribution of block maxima by partitioning data into equal-length blocks and fitting the GEV distribution to each block's maximum. The block length choice is crucial, balancing bias and variance. Short blocks yield too many maxima, violating model assumptions and increasing bias. Long blocks result in a small number of maxima, providing insufficient data to accurately estimate the model. The GEV distribution is defined as:

$$G(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\},$$

on the set $\{z : 1+\xi(z-\mu)/\sigma > 0\}$, where the parameters satisfy $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$ and taking the limit $\xi \to 0$ if $\xi = 0$.

Return Levels

The extreme values of a sequence $\{X_i\}_{i\geq 1}$ of i.i.d. random variables are analyzed using return levels. In general, the return level, noted $r(T)$, is the value exceeded on average once over a period T, it satisfies the equation $\mathbb{E}\left(\sum_{i=1}^{T}\mathbb{1}\{X_i \geq r(T)\}\right) = 1$ and is given by $r(T) = F^{\leftarrow}(1-1/T)$, where F is the CDF of X_i. For convenience, let $p = 1/T$.
For the GEV distribution, the return level z_p associated with return period $1/p$, for $p \in]0;1[$, is:

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}[1-\{-\log(1-p)\}^{-\xi}], & \text{for } \xi \neq 0 \\ \mu - \sigma\log\{-\log(1-p)\}, & \text{for } \xi = 0. \end{cases}$$

Limitations

Modeling block maxima can be inefficient for extreme event analysis, potentially discarding significant data. Moreover, the choice of block length, which is crucial for balancing bias and variance, is arbitrary.

5. Developed Methods

To address these issues and make use of all available data, we employ the threshold method. For a sequence of i.i.d. random variables $\{X_i\}_{i\geq 1}$ and a high threshold u, the distribution function of $(X-u)$, conditional on $X > u$, is approximately:

$$H(z) = 1 - \left(1+\frac{\xi z}{\tilde{\sigma}}\right)^{-1/\xi},$$

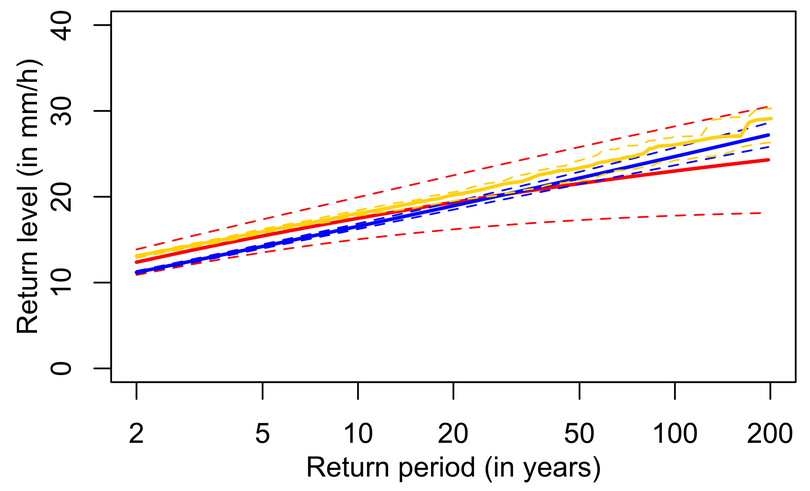
defined on $\{z : z > 0$ and $(1+\xi z/\tilde{\sigma}) > 0\}$, where $\tilde{\sigma} = \sigma + \xi(u-\mu)$. This defines the generalized Pareto distribution (GPD) [2].

Similar to the GEV distribution, GPD is easier to interpret through extreme quantiles. The value z_m exceeded on average once every m observations is: $z_m = u + \frac{\sigma}{\xi}\left[(m\zeta_u)^{\xi} - 1\right]$, where ζ_u is the probability that X exceeds u. For a time series with m observations per year, the return level z_p associated with the return period $1/p$ is:

$$z_p = \begin{cases} u + \frac{\sigma}{\xi}\left[(pm\zeta_u)^{\xi} - 1\right], & \text{for } \xi \neq 0 \\ u + \sigma\log(pm\zeta_u), & \text{for } \xi = 0. \end{cases}$$

GPD Applied to Declustered Time Series

To apply the generalized Pareto distribution, we need to determine thresholds. Some theoretical considerations lead us to choose $u_{obs} = 3.3$ (quantile 0.99 of observed data) and $u_{sim} = 9.5$ (quantile 0.999 of simulated data). Since the datasets are stationary but not independent, we decluster them to obtain independent variables. Declustering involves considering two hourly events as independent if separated by a sufficient time s, determined from the autocorrelation function of our datasets. We extract clusters of values above u separated by at least s hours and fit GPDs to the highest value in each cluster.
Using the estimated GPD parameters, we compute return levels for the observed (red) and simulated (blue) datasets. To validate, we compare with empirical quantiles of the simulated data (quantiles $1-1/(p\times m)$, with m the number of hours in a winter) (yellow).

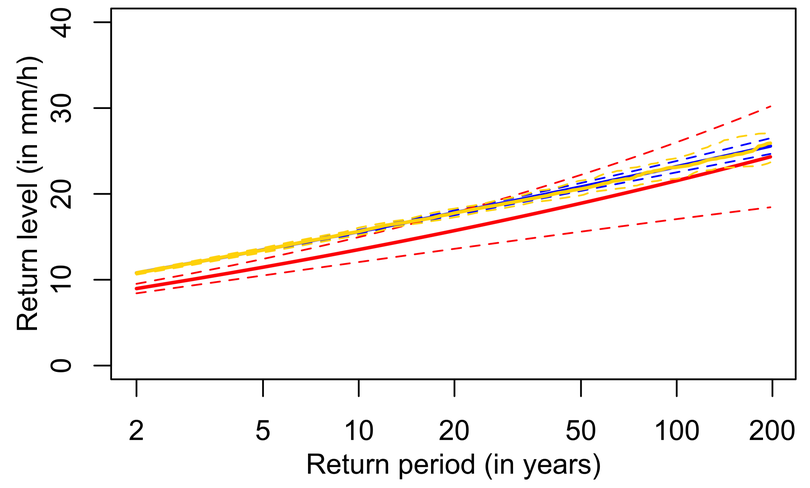


GPD Using the Extremal Index

Declustering discards threshold exceedances that are close in time. To account for time dependence, we use the extremal index θ. We estimate θ for each dataset using the estimator (1), fit a GPD to each peak-over-threshold series, and calculate return levels incorporating θ. The formula [4] used is:

$$z_p = \begin{cases} u + \frac{\sigma}{\xi}\left[(p\zeta\theta)^{\xi} - 1\right], & \text{for } \xi \neq 0 \\ u + \sigma\log(p\zeta\theta), & \text{for } \xi = 0. \end{cases}$$

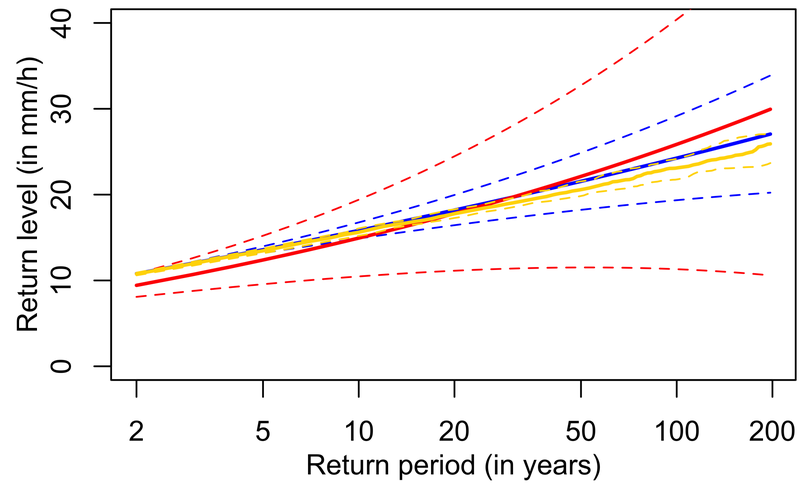
This formula adjusts return levels for a GPD fitted to non-independent series. Since θ represents the inverse of the mean cluster size, substituting $\zeta \times \theta$ for the exceedance probability ensures we account for only one exceedance per cluster, even though all exceedances are used to fit the GPD.



In this figure, empirical quantiles of the simulations are computed using θ (quantiles $1-1/(p\times m\times\theta_{sim})$). The return levels curve for the simulated dataset aligns perfectly with the empirical quantiles, indicating the high reliability of this method.

Hierarchical Model for Time Dependence

To estimate the GPD parameters, we use a hierarchical model designed to account for time dependence [5]. This approach embeds a latent process, that is a Markov chain, in the GPD parameters, capturing the dependence between exceedances over time. This method provides more accurate estimates of the parameters σ and ξ compared to a simple GPD. We fit this model to our datasets and use the estimated parameters σ and ξ to compute the return levels for a GPD, adjusted by θ.



References

[1] Taken from: D. L. De Luca & L. Galasso. *Calibration of NSRP models from extreme value distributions.* Hydrology. 2019.
[2] M.R. Leadbetter, G. Lindgren & H. Rootzén. *Extremes and related properties of random sequences and processes.* Springer, New York. 1983.
[3] C. Fero & J. Segers. *Inference for clusters of extreme values.* Journal of the Royal Statistical Society Series B. 2003.
[4] S. Coles *An Introduction to Statistical Modeling of Extreme Values.* Springer, New York. 2002. Equation (5.5).
[5] P. Bortot & C. Gaetan. *A latent process model for temporal extremes.* Scandinavian Journal of Statistics. 2014. Equation (4).