



HAL
open science

A parsimonious tail compliant multiscale statistical model for aggregated rainfall

Pierre Ailliot, Carlo Gaetan, Philippe Naveau

► **To cite this version:**

Pierre Ailliot, Carlo Gaetan, Philippe Naveau. A parsimonious tail compliant multiscale statistical model for aggregated rainfall. *Advances in Water Resources*, In press. hal-05457346

HAL Id: hal-05457346

<https://hal.science/hal-05457346v1>

Submitted on 14 Jan 2026

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

A parsimonious tail compliant multiscale statistical model for aggregated rainfall

Pierre Ailliot^a, Carlo Gaetan^b, Philippe Naveau^c

^a*Univ Brest, CNRS UMR 6205, Laboratoire de Mathématiques de Bretagne Atlantique, France*

^b*Dipartimento di Scienze Ambientali, Informatica e Statistica, Università Ca' Foscari di Venezia, Venice, Italy*

^c*Laboratoire des Sciences du Climat et l'Environnement, France*

Abstract

Modeling the probability distribution of rainfall intensities at different aggregation scales, say from sub-hourly to weekly, has always played a key role in most hydrological risk analysis, in particular in the computation of Intensity-Duration-Frequency (IDF) curves. Since any aggregation procedure involves accumulating rainfall over a prescribed time window, it naturally induces simple mathematical constraints related to summation. In particular, return levels inferred from a statistical model should be ordered across time scales, reflecting for example the fact that observed daily accumulations necessarily exceed those at sub-daily scales. From a statistical modeling perspective, each aggregation step combines information from shorter time scales without introducing additional data. Consequently, the number of model parameters should remain limited. Still, parsimonious aggregation models that describe the full distribution of rainfall intensities are sparse in the hydrological literature. In particular, most studies focus on extremes, e.g. by taking seasonal block maxima at different aggregation scales.

In this study, we propose a statistical framework that allows to model all rainfall intensities (low, medium and large) at different aggregation scales, while being parsimonious. To reach this goal, we use the extended generalized Pareto distribution (EGPD), which complies with extreme value theory for both low and high extremes and is flexible enough to capture the bulk of the distribution. We show a general result that explains how EGPD random variables behave under different types of aggregation procedures. Direct likelihood inference is difficult in our setting. However, by linking the EGPD class to Poisson compound sums, we can use the Panjer algorithm to quickly

and efficiently evaluate the composite likelihood of our proposed model. As a result, return levels can be obtained for any return period, particularly those below the annual and seasonal scales. In addition, our approach insures that return levels do not cross with aggregation.

To demonstrate the applicability of our method, we analyze sub-hourly time series from six gauging stations in France that have different climatological features. For each station, we only need a total of eight parameters to capture aggregation scales from six minutes to three days. IDF curves above and below the annual scale are provided.

Keywords: Rainfall Distribution, Aggregation, IDF curves, Extreme Value Theory, Extended Generalized Pareto Distribution, Compound Poisson Distribution.

1. Introduction

Rainfall aggregation over different time scales is a ubiquitous topic in hydrology. Describing changes in precipitation distributions from, say sub-hourly to higher scales such as daily or weekly periods, represents an important statistical task in any hydrological risk analysis. The archetypal example of this practical and scientific endeavor is the extensive literature dedicated to the generation of Intensity-Duration-Frequency (IDF) curves. These curves are fundamental tools in hydrology and water resources engineering, and illustrate the relationship between rainfall intensity, duration, and frequency (or return period). Historically, the geophysical analysis of IDF curves can be rooted back, at least, to more than half century ago. For example, [1] provided nationwide IDF curves for durations from 30 minutes to 24 hours and return periods up to 100 years. Today, one can find various flavors of IDF curves [see, e.g. 2], their associated maps for different regions of the world, and also well documented software packages for IDF computations [see, e.g. 3].

A common fundamental thread in all IDF curve studies is to determine what is the appropriate probability distribution of precipitation sums (equivalently averages), and how the features of such distributions vary when the chosen aggregation period varies, say from sub-hourly to weekly scales, see [4]. In this context, one main motivation of this work is to pinpoint a statistical paradox within extremes of rainfall aggregates, and to propose a model that reconciles contradictory interpretations of high return levels. To explain

such a paradox, we need to introduce the following notation. The capital letter Y corresponds to the random variable of interest at the smallest time scale of interest, say sub-hourly precipitation. Then, the simplest way to make rainfall aggregates is to sum these sub-hourly observations over a given period of interest of length d , by computing $Y_1 + \dots + Y_d$.

By construction, we always have $Y_i \leq Y_1 + \dots + Y_d$, for any $i = 1, \dots, d$. This constraint simply means that the sum of positive terms is always greater than any elements composing this sum¹. In terms of return levels and survival functions, this implies that for any $u \geq 0$

$$\Pr(Y_i > u) \leq \Pr(Y_1 + \dots + Y_d > u). \quad (1.1)$$

Therefore, the return levels² of Y_i should never intersect the return levels of the sum. It follows that any statistically coherent rainfall aggregation study should impose constraints (1.1) for all values of u , even when u is large. However, this is not always the case, as we will briefly illustrate.

To estimate high return levels, most hydrological analyses rely on extreme value theory (EVT) [see, e.g. 6, 7]. In particular, exceedances above a high threshold are classically modeled by the generalized Pareto (GP) cumulative distribution function defined by

$$H_\xi(y/\sigma) = 1 - \left(1 + \xi \frac{y}{\sigma}\right)_+^{-1/\xi}, \quad \sigma > 0, \xi \in \mathbb{R}, \quad (1.2)$$

with the notation $a_+ = \max(a, 0)$. The choice $\xi = 0$ corresponds to the exponential case. The scale parameter σ drives the spread among extreme exceedances, while the shape parameter ξ controls the upper tail behavior. Estimates of ξ are often positive for precipitation [see, e.g. 5] and we assume that $\xi > 0$ in this work.

In the context of IDF studies, it is natural to ask how the shape parameter ξ of the sum changes with the aggregation scale. In particular, is the shape parameter of the sum different from that of its components Y_i ?

¹Although simple, this statement is very robust. Even if the non-negative sample (Y_1, \dots, Y_d) corresponds to non-stationary data that may be strongly dependent, the inequality remains true as all $Y_i \geq 0$.

²The return level with respect to the return period T corresponds to the scalar u_T satisfying the equation $\Pr(Y > u_T) = 1/T$, i.e. it is the $1 - 1/T$ quantile of the random variable Y [see, e.g. 5].

To start answering this question, one can look at a simple case study and compare the empirical distributions obtained at different aggregation scales. For example, Figure 1 displays four histograms of rainfall rates recorded in Brest (France). The upper left panel begins with the aggregation scale of six minutes, while yearly aggregation scale is displayed in the bottom right panel. As anticipated from (1.1), the x-axis range (rainfall sums) increases with the increasing scale, but the yearly histogram appears closer to a bell shape curve (normal distribution). This suggests that the upper tail behaviors of precipitation are affected by increasing scales.

This phenomenon is well known by hydrologists working on extreme rainfall analysis. For example, [2] carefully modeled rainfall measurements from 81 weather stations in Switzerland with a minimum record length of 20 years. Figure 6 in that paper clearly indicates that the shape parameter ξ decreases from 0.3 (30-minute scale) to essentially zero (daily scale). Statistically, one could invoke the central limit theorem to explain this phenomenon. Sums or averages of random variables with finite variances should become closer to a Gaussian distribution as the aggregation scale increases. As the Gaussian distribution has a lighter tail than a Pareto tail [see 8, for instance], it seems to make sense that cumulative rainfall extremes appear to have a lighter tail than the ones at finer timescale. However, this reasoning contradicts inequality (1.1)!³

Interestingly, probability theory confirms this. For large u , the so-called Feller approximation [9] states

$$\Pr(Y_1 + \dots + Y_d > u) \approx \Pr(Y_1 > u) + \dots + \Pr(Y_d > u) \quad (1.3)$$

when Y_i are independently and identically distributed (i.i.d.) with a heavy tail behavior. Note that extensions with respect to the simple setting of i.i.d. random variables exist in the literature [see, e.g. 10]. Practically, this probability result means that aggregated precipitation extremes cannot, at least statistically, become lighter with increasing scales.

In this context, one can wonder why such a strong disagreement between the practical consensus and the theoretical aspect appears. One main objec-

³To see this, EVT with (1.2) tells us that, for large u , $\Pr(Y_1 > u) \approx \alpha u^{-1/\xi}$ for some constant α . Similarly, $\Pr(Y_1 + \dots + Y_d > u) \approx c_d u^{-1/\xi_d}$. If the shape parameter ξ_d decreases with d , i.e. $\xi_d < \xi$, then one can always find a large u such that $\alpha u^{-1/\xi} > c_d u^{-1/\xi_d}$. This implies that inequality (1.1) does not hold for large u , and unwelcome crossing of return levels will eventually occur for large extremes.

tive of this work is to address this issue and to propose a statistical model that can reconcile the theory with practical findings.

Intuitively, from a probability point of view, there are two opposite forces at play when summing. The central limit theorem “flattens” the upper tail behavior, but the Feller approximation imposes a constant value of the shape parameter, ξ , to avoid high return level crossings. Additionally, aggregation reduces sample sizes. Consequently, extremal exceedances at scale d are fewer than those of the original data. This makes the generalized Pareto distribution (GPD) approximation less reliable.

Hence, our strategy to avoid this problem is to bypass the threshold selection step. A fundamental aspect of our approach is to strongly control low and large tail behaviors at all aggregation scales, while keeping the bulk and the transition to tails very flexible. This will allow to model a wide range of shapes for the bulk; from Pareto type like in the top panels of Figure 1 to more Gaussian shapes like in the bottom right panel of Figure 1, while preserving constant tail shape parameters. To do so, we choose a specific distribution class for Y_i and study how the mathematical properties of this distribution change with aggregation.

In this work, we focus on the extended generalized Pareto distributions (EGPD) class [11, 12]. In addition to being a flexible class of distributions, it has been shown to be applicable to various hydrological setups. For example, the entire distribution of rainfall amounts was modeled using an EGPD in [13, 14]. The EGPD class had been integrated into a random forest scheme in order to improve the post-processing of forecasted rainfall [15, 16]. It has also been used to perform rainfall comparison [17, 18] and to produce regional clustering analysis [19]. Besides rainfall analysis with heavy tails, it is also possible to tailor the EGPD class to model lighter and even bounded variables, such that wind data [20], wave heights [21] or temperatures [13]. From an inferential point of view, EGPD regression approaches have been also studied in detail, [see, e.g. 22] for a Bayesian hierarchical scheme or [23] within a distributional regression framework. Still, we were not able to find articles that study the capability of the EGPD to model theoretically sum of EGPD distributed random variables, and to apply such distributions within a rainfall aggregation context.

Concerning IDF curves, various definitions exist [see, e.g. 24]. In most setups, they share two common threads: the computation of total rainfall accumulation sums (or averages) over a given time scale d [see, e.g. 2] and sometimes taking a block maximum, typically over a year or a season [see, e.g.

3]. Mathematically, such operations can be viewed as the transformation of the original rainfall measurement times series $\mathbf{Y} = (Y_1, Y_2, \dots)^\top$ into a single index. For example, a common form of such indices [see, e.g. 3] is

$$T(\mathbf{Y}) = \max(\bar{Y}_1, \bar{Y}_2, \dots) \quad (1.4)$$

where \bar{Y}_i represents any type of averages over d time steps and the maximum block size corresponds to the number of such averages over the full period of interest, classically one year.

In this work, our first task is to provide a general result, see Proposition 2.2, that details the conditions under which, given that the original time series \mathbf{Y} have identical EGPD marginals, the aggregated vector $T(\mathbf{Y})$ stays EGPD distributed. A key element of this result is to understand how the constraint (1.3) can be generalized to $T(\mathbf{Y})$. In addition, it will be important to determine how EGPD parameters change with the transformation $T(\mathbf{Y})$, to explain how to infer them. A limiting aspect of some existing IDF approaches is that they cannot produce curves for short return periods. This is particularly true for IDF methods based on block maxima. For example, taking annual maxima at various aggregation scales, like in [3], prevents the computation of any return period below the annual scale. To solve this issue, we avoid taking block maxima in our application and, in this paper, we will mainly focus on the following two additive aggregation types. Firstly the classical aggregation scheme [see, e.g. 2] is defined by

$$T(\mathbf{Y}) = Y_1 + \dots + Y_d. \quad (1.5)$$

The second corresponds to a stochastic representation based on the number of wet events over the duration d , say N and their corresponding intensities Y_i , in the following way

$$T(\mathbf{Y}) = \sum_{i=1}^N Y_i, \quad (1.6)$$

with the convention $T(\mathbf{Y}) = 0$ if $N = 0$. This stochastic representation is well-known in hydrology [see, e.g. 25, 26, 27, 28, 29, 30] and will be a key element in the statistical model proposed in Section 3.

Concerning multiscale rainfall modeling, stochastic rainfall generators offer a different avenue. For example, randomized Bartlett-Lewis pulse models [see, e.g. 31] aimed at reproducing and combining different storm variables (arrival time, duration, intensities, etc). Although efficient at reproducing

mean rainfall statistics, underestimation of extreme rainfall may appear. For this reason, we do not pursue this approach here as one of our main focus is to capture extremes distributions at all aggregation scales with statistical guaranties. Complementary scaling and multifractal approaches provide valuable descriptive insights into cross-scale behavior, but are not usually formulated as generative, likelihood-based models for the full distribution of aggregated rainfall intensities at each duration [32]

The organization of this article is as follows: Section 2 recalls the main features of the EGPD and contains Proposition 2.2 which explains how the EGPD parameters can change with aggregation scales. In addition, important links between EGPD and Poisson compound distributions are investigated. This leads to the definition of our statistical model for aggregated rainfall in Section 3. Inference and application are presented in Section 4. Conclusions are given in Section 5. Proofs of all propositions, details of a numerical algorithm for evaluating distribution functions, and results of a simulation experiment can be found in the Appendices.

2. Extended generalized Pareto distributions

The fundamental feature of $H_\xi(\cdot)$ defined in (1.2) is its stability under thresholding (up to a normalizing constant). Although mathematically sound for modeling heavy rainfall, there is no reason that such a GPD will fit well low and moderate precipitation. This limitation has been addressed, and there exists today a wide range of possible statistical options to model the probability distribution of the entire spectrum of precipitation [see, e.g. 33, 34].

Definition 2.1. Let U be a uniformly distributed random variable on $[0, 1]$. Let $B(\cdot)$ be a cumulative distribution function (cdf) on $[0, 1]$ with a continuous probability density function (pdf) $b(\cdot)$ on $(0, 1]$ and a possible non-zero mass $B(0)$ at 0. The non-negative random variable defined as

$$Y = \sigma H_\xi^{-1} \left((B^{-1}(U))^{1/\kappa} \right), \quad (2.1)$$

where σ, κ, ξ are three positive constants, is said to follow an extended generalized Pareto distribution (EGPD), denoted by $Y \sim EGPD(\sigma, \kappa, \xi, B)$, if

$$0 < b(0^+) < \infty \text{ and } 0 < b(1) < \infty, \quad (2.2)$$

where

$$b(0^+) = \lim_{u \rightarrow 0^+} \frac{B(u) - B(0)}{u} \quad (2.3)$$

The definition (2.1) is based on that provided by [11], but differs in terms of notation, as the cdf $G(u) = B(u^\kappa)$ was used in [11]. The notation with $G(\cdot)$ was slightly ambiguous because the parameter κ driving the lower tail was "hidden" in the cdf G itself, which was not the case for ξ . In contrast, the new definition, $EGPD(\sigma, \kappa, \xi, B)$, more precisely distinguishes the roles of κ in modeling the lower tail, $B(\cdot)$ as the transfer function from low to heavy intensities, and ξ in modeling the upper tail. Although this work does not explicitly model dry events, we recognize their important role in aggregated rainfall distributions. In this respect, Definition 2.1 is further different from the definition in [11]. It separates the probability of a dry event, $B(0)$, and the behavior of the distribution of Y for small but positive values near 0.

More formally, one can check that the lower tail satisfies

$$P(Y \leq y) = B(0) + b(0^+) \left(\frac{y}{\sigma}\right)^\kappa + o(y^\kappa) \quad \text{as } y \rightarrow 0^+, \quad (2.4)$$

and the upper tail is GP equivalent, i.e.

$$P(Y > y) = \kappa b(1) \overline{H}_\xi\left(\frac{y}{\sigma}\right) + o(y^{-\frac{1}{\xi}}) \quad \text{as } y \rightarrow +\infty \quad (2.5)$$

with $\overline{H}_\xi\left(\frac{y}{\sigma}\right) = 1 - H_\xi\left(\frac{y}{\sigma}\right)$. The pdf and cdf expressions the EGPD, as well as the proof of (2.4) and (2.5), can be found in [Appendix A](#).

The special case $B(u) = u$ is called Type 1, see the case $G(u) = u^\kappa$ with $\kappa \neq 0$ in [11]. It offers a flexible choice to model the full range of hourly and daily rainfall, extremes included, see [35, 14, 36]. When $\kappa = 1$, it corresponds to the classical generalized Pareto distribution. Another member of the EGPD class used in hydrology is the so-called Pareto-Burr-Feller proposed in [37].

[2] used a EGPD of Type 1 to derive an IDF for different aggregation scales. To add flexibility, they allowed the shape parameter ξ to vary with the aggregation scale. As mentioned earlier, this leads to a mathematical contradiction with (1.3) and allows for unwanted return level crossings in extremes. One option to be in compliance with (1.3) is to fix the shape parameter ξ but then, to fit the data adequately, the function $B(u)$ needs to vary with the aggregation scale. We will follow this modeling path that leads to the question of how to characterize mathematically $B(u)$ changes

with the aggregation scale. Before answering this complex question, we need to derive some properties of the EGPD.

2.1. Sums and other transformations of EGPD random variables

A fundamental question regarding IDF curves is to determine how the distributional features of a stationary time series, say $\mathbf{Y} = (Y_1, Y_2, \dots)^\top$, change according to a transformation $T(\cdot)$. When all Y_i have the same EGPD marginal, one can wonder what are the sufficient conditions to ensure that the real-valued transformed vector $T(\mathbf{Y})$ has also a EGPD distribution, not necessarily with the same parameters as projecting the data at hand into a single value index will likely change the bulk of the distribution. For example, the sum of two type 1 EGPD is not a type 1 EGPD, but we will show that it is still a EGPD. The following proposition provides the precise conditions to ensure this. Note that this result is broad enough to allow dependencies among the Y_i 's, and to avoid imposing a specific form on the transform $T(\cdot)$.

Proposition 2.2. *Let T be a non-negative random variable and $Y \sim \text{EGPD}(\sigma, \kappa, \xi, B)$ for $\xi > 0$, $\kappa > 0$. If there exist some positive and finite constants α , β and γ such that*

$$\lim_{y \rightarrow \infty} \frac{\Pr(T > y)}{\Pr(Y > y)} = \alpha, \quad (2.6)$$

$$\lim_{y \rightarrow 0^+} \frac{\Pr(0 < T \leq y)}{[\Pr(0 < Y \leq y)]^\gamma} = \beta, \quad (2.7)$$

then there exists a cdf B_T such that $T \sim \text{EGPD}(\sigma, \gamma\kappa, \xi, B_T)$ with

$$b_T(0^+) = \beta b(0^+)^\gamma \text{ and } b_T(1) = \alpha b(1)/\gamma. \quad (2.8)$$

Condition (2.6) holds for most applications when $T = T(\mathbf{Y})$ with $T(\cdot)$ one of the usual transforms like (1.4), (1.5) or (1.6) applied when working with aggregated data. It forces the upper tail of $T(\mathbf{Y})$ to be proportional to the marginal one. In cases like (1.5), it corresponds to Feller's lemma, see (1.3). For transforms like (1.6), it is related to Breiman's lemma [38]. It basically tells us that the upper tail behavior of a random sum of heavy tailed distributed random variables is driven by the largest tail index. [10] detailed different setups for dependent cases and [39] reviews existing results for a variety of transformations $T(\cdot)$, including the case $T(\mathbf{Y}) = \max(Y_1, \dots, Y_d)$. Condition (2.7) does the same but for the lower tail. Checking the validity of

condition (2.7) is also possible for simple setups. For example, we detail the i.i.d. case in [Appendix C](#) with five different setups. The case $\gamma > 1$ appears when $B(0) = 0$ (only wet events), while $\gamma = 1$ otherwise. In the remainder of the paper, this later case is assumed, meaning that the behavior of the lower tail of $T(\mathbf{Y})$ is proportional to that of Y_i .

An interesting practical outcome of this proposition is that the shape parameters, κ and ξ , which are related to the lower and upper tails of the distribution, remain unchanged when the transform $T(\cdot)$ is applied. Equation (2.8) also provides the values of $b_T(u)$ for $u = 0^+$ and $u = 1$. It would be of great interest to also deduce, for all u in $(0, 1)$, the function $b_T(u)$ in function of $B(u)$ for any d for the simple aggregate defined by (1.5). If possible, then the practitioner could just focus on modeling the finest time scale to infer $B(\cdot)$. But, this task is mathematically challenging. For example, suppose that at the finest timescale (six minutes in our application), the non-zero observations simply follow an i.i.d. GPD sample. Even in this simple case, we were unable to determine the parametric form of $b_T(u)$ for any $u \in (0, 1)$. Consequently, the pdf of $T(\mathbf{Y}) = Y_1 + \dots + Y_d$ is not explicit.

In this work, compound sums (1.6) are used as a model for aggregated sum (1.5). This modeling choice is motivated by interpretability, theoretical and computational reasons, which are detailed in the next section.

2.2. Compound Poisson-EGPD distributions

A classical modeling approach in hydrology considers that rainfall time series can be decomposed into a succession of rainfall events, which arise as a point process with rainfall amounts associated to each of these events [40]. A natural stochastic representation of aggregated rainfall associated to such representation was recalled in the introduction via the compound sum (1.6) [see, e.g. 25, 26, 27, 28, 29, 30]. To make the link with the EGPD class, we introduce the following definition.

Definition 2.3. Let $\mathbf{Y}^* = (Y_1^*, Y_2^*, \dots)^\top$ be a sequence of i.i.d. EGPD random variables, $Y_i^* \sim \text{EGPD}(\sigma, \kappa, \xi, B)$, with $B(0) = 0$, $\kappa > 0$ and $\xi > 0$. Let N be a Poisson random variable with mean λ independent of \mathbf{Y}^* . The distribution of the random sum defined by $\sum_{i=1}^N Y_i^*$, with the convention that $\sum_{i=1}^0 Y_i^* = 0$, is called a compound Poisson-EGPD.

The following proposition shows that Compound Poisson-EGPD is a sub-family of EGPD. Combined with Proposition 2.2, it implies that compound

Poisson-EGPD have the same tails that aggregated sums (1.5) of EGPD random variables. For mathematical completeness, Proposition 2.4 characterizes more precisely Compound Poisson-EGPD as the sub-family of EGPD which are infinitely divisible, i.e. that can be expressed as the probability distribution of aggregated sums (1.5) of any arbitrary number d of i.i.d. random variables.

Proposition 2.4. *A compound-EGPD random variable is an EGPD random variable. More precisely, with the notations of Definition 2.3, we have*

$$\sum_{i=1}^N Y_i^* \sim EGPD(\sigma, \kappa, \xi, B_\lambda)$$

with $B_\lambda(0) = \exp(-\lambda)$, $b_\lambda(0^+) = \lambda \exp(-\lambda)b(0^+)$ and $b_\lambda(1) = \lambda b(1)$.

Conversely, any infinitely divisible EGPD(σ, κ, ξ, B) random variable with $B(0) > 0$ is a compound Poisson-EGPD random variable.

At this stage, let us highlight that the random variables Y_i^* in Definition 2.3 can be very general as the cdf B has only the requirements (2.2) and (2.3). To balance between parameters parsimony and model versatility, we fix $B(u) = u$ and let λ vary according to the aggregation scale. This leads to the following definition.

Definition 2.5. In the particular case when $B(u) = u$, the positive⁴ part of the distribution of the random sum $\sum_{i=1}^N Y_i^*$ in Definition 2.3 is denoted $EGPD(\sigma, \kappa, \xi, \lambda)$.

Proposition 2.4 implies that if $Y \sim EGPD(\sigma, \kappa, \xi, \lambda)$ then there exists a cdf B_λ such that $Y \sim EGPD(\sigma, \kappa, \xi, B_\lambda)$ with explicit expressions for $b_\lambda(0^+)$ and $b_\lambda(1)$. Remark that $B_\lambda(0) = 0$ since only the positive part of the compound Poisson-EGPD is kept in Definition 2.5. However the bulk of the cdf B_λ and the pdf of Y are unknown. This issue is addressed by observing that the pdf of compound sums of the form (1.6) can be numerically approximated with a low computational cost (see e.g. [41] and references therein). The numerical results given in this paper were obtained using the Panjer's algorithm [42] detailed in Appendix F.

When all parameters except for the Poisson mean λ are fixed, λ is the sole factor that drives the transition from the lower to the upper tail of the $EGPD(\sigma, \kappa, \xi, B_\lambda)$. The right panels of Figure 2 shows the function $b_\lambda(\cdot)$ for $\lambda \in \{0.01, 1, 3, 10\}$ (from top to bottom panels). In these plots, the other

parameters are equal to $\kappa = 0.3$, $\sigma = 1$ and $\xi = 0.25$. These are typical values for the rainfall data considered in this study (the shapes of the pdf plotted on Figures 1 and 2 share similarities). The corresponding pdfs are displayed in the left panels. To emphasize the log-linear behaviors in the tails, the middle panels show these pdfs in log scale for the x- and y-axis, these slopes being related to the parameters κ and ξ , see Proposition 2.4.

In the top row, when λ is close to zero, we can recognize a type 1 EGPD as $b(u) \approx 1$. This can be retrieved formally by computing the limit of $\Pr(N = n | N > 0)$ when $\lambda \rightarrow 0$ for a Poisson random variable N , which is equal to 1 for $n = 1$ and 0 for $n \geq 2$. For the largest values of the parameter λ (see the last row in blue), the pdf of the EGPD is bimodal, with a sharp mode at 0 and a second mode in the bulk of the distribution, i.e. the central limit theorem appears here.

3. A statistical model for aggregated rainfall

Let D be the largest scale of aggregation under study. For any integer $d = 1, \dots, D$, we denote

$$A_d = (Y_1 + \dots + Y_d)^+ \quad (3.1)$$

the random variable which describes the positive⁴ precipitation amount at aggregation scale d , where zero values are truncated after computing aggregated rainfall $Y_1 + \dots + Y_d$.

In the preceding section, it was pointed that the compound EGPD can serve as an appropriate statistical model for the distribution of aggregated rainfall data. In the remaining of the work, we thus assume that for any $d = 1, \dots, D$

$$A_d \sim EGPD(\sigma_d, \kappa, \xi, \lambda_d). \quad (3.2)$$

It should be noted that the scale parameter, σ_d , and the mean, λ_d , are allowed to vary with d , while the parameters κ and ξ are assumed to be

⁴Both the compound-EGPD and the rainfall distributions have a point mass at zero. However, numerical experiments have shown that additional challenges arise when attempting to simultaneously model dry and wet conditions. For example, an excess of zeros in the rainfall distribution necessitates zero-inflated distributions. Additionally, different temporal dynamics in dry and wet events affect the aggregation properties. For this study, only the positive part of the rainfall distributions is modeled.

constants. In this study, we hypothesize that the following parametrization is applicable:

$$\log \sigma_d = \sum_{i=0}^p s_i (\log d)^i \quad , \quad \log \lambda_d = \sum_{j=0}^q l_j (\log d)^j . \quad (3.3)$$

Such parametrization share similarities with classical models used in the literature on IDF curves. For example, log-linearly changes in the scale parameter were present in the GEV approach based on annual block maxima [see, e.g. 43], or the GPD approach when dealing with threshold exceedances [see, e.g. 44]. A significant difference between our EGPD modeling and the one in [2] is that our shape parameters κ and ξ remain constant across different aggregation scales, to be consistent with Propositions 2.2 and 2.4.

The coefficients s_i and l_j are not variation free. Proposition 3.1 implies that the non-crossing return level condition (1.1), which is an important motivation for this work, is true when both functions σ_d and λ_d increase with d . These monotonicity conditions will be used as constraints in the estimation procedure.

Proposition 3.1. *Let Y_1, Y_2, \dots be a stationary sequence of non-negative random variables such that assumption (3.2) is satisfied with $\sigma_d \geq \sigma_1$ and $\lambda_d \geq \lambda_1$. Then (1.1) holds true.*

4. Inference and application

4.1. Parameter estimation

The model introduced in Section 3 has $p+q+4$ parameters encapsulated in the vector $\theta = (s_1, \dots, s_p, \kappa, \xi, l_1, \dots, l_q)$. Although precipitation data are generally not independent in time and across aggregation scales, we estimate these parameters by maximizing the following composite likelihood function $L(\theta)$, constructed under working independence assumptions [see, e.g. 45],

$$L(\theta) = \prod_{d=1}^D \prod_{i=1}^{n_d} p(a_{d,i}; \sigma_d, \kappa, \xi, \lambda_d) \quad (4.1)$$

where $(a_{d,1}, \dots, a_{d,n_d})$ denotes the sample of positive rainfall amounts available at aggregation scale d (see (3.1)), $p(\cdot; \sigma, \kappa, \xi, \lambda)$ the pdf of the $EGPD(\sigma, \kappa, \xi, \lambda)$ (see Appendix F for more details) and σ_d and λ_d are function of θ , see (3.3).

The computational cost of maximizing (4.1) can be significant if we consider a large number of aggregation scales, D . In such cases, an estimate of θ can be obtained using a smaller subset of $\{1, \dots, D\}$ without a significant loss of efficiency. In this paper, we maximize (4.1) with the subset $d \in \{1, 2, \dots, 10, 20, \dots, 240, 270, \dots, 720\}$, i.e. instead of focusing on 720 durations, we focus on 42 representative durations. These durations correspond to all available sub-hourly durations and multiples of hourly durations up to the daily scale and on multiples of three hourly durations up to three days.

Regarding confidence bands around our estimates from (4.1), we follow [2] by using block bootstrap. All the numerical results presented in this study were obtained using a block size of two weeks length and by repeating the optimization procedure on 500 bootstrap samples. We validated the estimation procedure using simulations in an idealized i.i.d. setting. The results can be found in Appendix G.

4.2. Rainfall analysis of French precipitation intensities

In this study, we examine rainfall data recorded by Météo-France at six meteorological stations in France, see Figure 3, that represent different climate. Data are available at the url <https://meteo.data.gouv.fr/datasets/donnees-climatologiques-de-base-6-minutes/> with a 6 min time step from 2006 until 2023. To remove the seasonal component, the focus is on August and September, when intense convective precipitation events typically occur. Summary statistics of 6-minute rainfall data by station are given in Table 4.1. All data were obtained using tipping bucket gauges with 0.2 mm precision. To compute the composite likelihood (4.1), the same 0.2 mm discretization is applied to the distribution of the seed before running the Panjer recursions, see Appendix F.

The left panel of Figure 4 displays the estimate of the upper tail parameter at the different stations considered in this study. The point estimates of ξ seem physically plausible and range from approximately 0.15 in Brest, close to the Atlantic Ocean, to 0.35 in Lyon where more intense convective events are observed. The estimates of the lower tail parameter κ are all in the interval (0.2, 0.45) (see right panel of Figure 4), corresponding to distributions with a relatively sharp mode at 0^+ . Interestingly, the estimate of κ seems to be related to the percentage of measurement greater than 0.2 in the wet measurements (see the stars superimposed to the boxplot). This is consistent with the interpretation that κ describes the lower tail of the distributions and

Station	Altitude (m)	Percentage of missing values	Percentage of positive data	Mean of positive data (mm)	St. dev. of positive data (mm)
BREST	92	0.12	3.04	0.33	0.37
NANCY	212	0.00	2.34	0.38	0.52
LILLE	47	0.50	2.25	0.38	0.61
BORDEAUX	47	0.03	1.59	0.44	0.71
VILLACOUBLAY	174	1.28	1.71	0.38	0.52
LYON	235	0.01	2.09	0.44	0.71

Table 4.1: Summary statistics of 6-minutes rainfall data by station. Results for August-September based on 18 years of data.

should be smaller at stations where light rain conditions are more frequent, as it is the case in Brest for example.

Except the differences in the values of κ and ξ discussed above, similar fitting results were generally obtained at the different stations. We therefore choose to focus on one of these stations, namely Brest. Figure 5 shows the evolution of the parameters σ_d and λ_d with the aggregation scale d . As expected, the functions are increasing, since the constraint is imposed in the estimation procedure to ensure the non-crossing condition (1.1) (see discussion before Proposition 3.1). This is also consistent with the physical interpretation of the model, with λ_d representing the mean number of rainfall events on the aggregation scale d and σ_d describing the scale of their intensities.

Figure 6 compares the empirical distributions of rainfall with those given by the fitted parametric model at different representative aggregation scales, as illustrated by QQ plots (quantile-quantile plots). The model generally fits well for aggregation scales between 30 minutes and three days. However, it slightly underestimates the probability of extreme events at finer scales. Figure 7 provides another representation of the quantiles at different aggregation scales, which illustrates this as well. The figure focuses on high-order quantiles associated with return periods ranging from 15 days to 100 months. For the six stations examined in this study, the empirical quantiles generally align closely with the theoretical quantiles predicted by the fitted model for aggregation durations ranging from 30 minutes to one day. However, there is a consistent underestimation of larger quantiles for shorter durations and, conversely, a slight overestimation at larger aggregation scales for certain

stations. More flexible models need to be investigated to handle such durations with the proposed methodology. Once again, it is worth noting that, unlike the models usually used to construct IDF curves, the fitted model describes the entire distribution and can therefore be used to compute quantiles associated with short return periods.

5. Conclusions

A new statistical model has been proposed for the distribution of positive precipitation on various aggregation scales. The model uses the compound Poisson-EGPD as a key ingredient. This distribution was chosen for its physical interpretability and computational advantages. Theoretically, it is also shown to provide a tail-compliant model for the distribution of aggregated data. The proposed model is parsimonious, with only eight parameters, and can be fitted to data at low computational cost using the Panjer algorithm. The model was fitted to 6-minute rainfall data from six stations in France with varying climates. The entire distribution of positive rainfall data was found to be generally well described by the proposed model at aggregation scales ranging from 30 minutes to three days.

These results are encouraging and outline possible avenues for future research. Firstly, systematic validation across multiple stations and seasons could be considered. Secondly, longer observation periods could be used to take into account possible climate changes. From a model definition point of view, introducing a time-dependent parameterization to the model (3.2) does not present any particular estimation problems. However, it is worth noting that this model is justified on the basis of the results of Propositions 2.2 and 2.4 which are valid for identically distributed variables. Extending these results to non-identically distributed variables is an interesting theoretical problem. Finally, in the case of bounded-tail distributions, an extension of the EGPD framework is possible (see [21]), but adapting the theoretical results developed in this paper to that setting would require substantial additional work, as the key asymptotic arguments used here no longer apply directly.

Acknowledgments

The authors acknowledge the support of the SHARE PEPR Maths-Vives project (France 2030 ANR-24-EXMA-0008).

Part of Gaetan's research work took place within the framework of the DoE 2023-2027 (MUR, AIS.DIP.ECCELLENZA2023_27.FF project).

Part of Naveau's research work was supported by the French Agence Nationale de la Recherche: EXSTA, the PEPR TRACCS programme under grant number (PC4 EXTENDING, ANR-22-EXTR-0005), and the PEPR IRIMONT (France 2030 ANR-22-EXIR-0003). He has also benefited from the Geolearning research chair, a joint initiative of Mines Paris and the French National Institute for Agricultural Research (INRAE).

Appendix A. Basic EGPD properties

The cdf of a $EGPD(\sigma, \kappa, \xi, B)$ can be expressed as

$$F(y) = B(H_\xi^\kappa(y/\sigma)), \text{ for any } y \geq 0 \quad (\text{A.1})$$

and the pdf of a $EGPD(\sigma, \kappa, \xi, B)$ can be written as

$$\frac{\kappa}{\sigma} h_\xi(y/\sigma) H_\xi^{\kappa-1}(y/\sigma) b(H_\xi^\kappa(y/\sigma)), \text{ for } y > 0, \quad (\text{A.2})$$

where $h_\xi(\cdot)$ corresponds to the pdf of a Generalized Pareto random variable.

Since $H_\xi(y) = y + o(y)$ for y near zero, $h_\xi(0) = 1$ and $H_\xi(0) = 0$, we have

$$\lim_{y \rightarrow 0^+} \frac{F(y) - F(0)}{y^\kappa} = \lim_{y \rightarrow 0^+} \frac{f(y)}{\kappa y^{\kappa-1}} = \frac{b(0^+)}{\sigma^\kappa}. \quad (\text{A.3})$$

where the first relation was derived via L'Hôpital's rule. This implies (2.4).

Concerning the upper tail behavior L'Hôpital's rule and (A.2) give

$$\lim_{y \rightarrow +\infty} \frac{1 - F(y)}{1 - H_\xi(y/\sigma)} = \lim_{y \rightarrow +\infty} \frac{f(y)}{h_\xi(y/\sigma)/\sigma} = \kappa \cdot b(1). \quad (\text{A.4})$$

This implies (2.5).

We also highlight that following identifiability issue. The two distributions $EGPD(\sigma, \kappa, \xi, B)$ and $EGPD(\tilde{\sigma}, \kappa, \xi, \tilde{B})$ are equal if $B(0) = \tilde{B}(0)$ and

$$\tilde{B}(u^\kappa) = B(s^\kappa(u; \nu, \xi)),$$

with $\nu = \tilde{\sigma}/\sigma$ and $s(u; \nu, \xi) = H_\xi(\nu H_\xi^{-1}(u))$. Therefore, B and σ cannot be let completely free in practice.

Appendix B. Proof of Proposition 2.2

To simplify expressions, let us assume that $\sigma = 1$; the general case $\sigma \neq 1$ can easily be deduced from this particular case. Let F_T denote the cdf of T . It is always possible to define the function

$$B_T(u) := F_T(H_\xi^{-1}(u^{1/\kappa_T}))$$

for $\kappa_T > 0$. Note that $B_T(0) = F_T(0) = \Pr(T(\mathbf{Y}) = 0)$.

As a composition of non-decreasing functions, $B_T(u)$ is also non-decreasing for u on $[0, 1]$ and, by definition, we have $F_T(y) = B_T(H_\xi(y)^{\kappa_T})$.

According to Definition 2.1, we need to show that the quantities

$$b_T(0^+) := \lim_{u \rightarrow 0^+} \frac{B_T(u) - B_T(0)}{u} \text{ and } b_T(1) := \lim_{u \rightarrow 1^+} \frac{B_T(u) - B_T(1)}{u - 1}$$

are finite and non-null. By changing u into $y = H_\xi^{-1}(u^{1/\kappa_T})$ and using the Taylor expansion

$$\Pr(Y_1 > y) = 1 - (1 - \overline{H}_\xi(y))^{\kappa_T} \approx \kappa_T \overline{H}_\xi(y)$$

for large y , we can then deduce from condition (2.6)

$$\lim_{u \rightarrow 1^-} \frac{B_T(u) - B_T(1)}{u - 1} = \lim_{y \rightarrow \infty} \frac{\Pr(T > y)}{\kappa_T \overline{H}_\xi(y)} = \alpha \frac{\kappa}{\kappa_T} \lim_{y \rightarrow \infty} \frac{\Pr(Y_1 > y)}{\kappa \overline{H}_\xi(y)} = \alpha \frac{\kappa}{\kappa_T} b(1) = b_T(1).$$

Setting $\kappa_T = \kappa\gamma$, we obtain the second equation in (2.8).

Concerning the lower tail, we have for $y \geq 0$

$$\Pr(T \leq y) = \Pr(T = 0) + \Pr(0 < T \leq y).$$

In addition, we note that as $H_\xi(y) = y + o(1)$ for $y \rightarrow 0^+$ zero and consequently, we have $u = H_\xi(y^{\kappa_T}) = y^{\kappa_T} + o(1)$ when u is near zero. Therefore,

$$\begin{aligned} \lim_{u \rightarrow 0^+} \frac{B_T(u) - B_T(0)}{u} &= \lim_{y \rightarrow 0^+} \frac{\Pr(T \leq y) - \Pr(T = 0)}{y^{\kappa_T}}, \\ &= \lim_{y \rightarrow 0^+} \frac{\Pr(0 < T \leq y)}{y^{\kappa_T}}, \\ &= \beta \lim_{y \rightarrow 0^+} \frac{[\Pr(0 < Y_1 \leq y)]^\gamma}{y^{\kappa_T}}, \text{ by (2.7),} \\ &= \beta \lim_{y \rightarrow 0^+} \frac{(b(0^+)y^\kappa)^\gamma}{y^{\kappa_T}}, \text{ as } Y_1 \text{ follows an } \text{EGPD}(1, \kappa, \xi, B), \\ &= \beta b(0^+)^\gamma, \text{ when } \kappa_T = \gamma \kappa. \end{aligned}$$

□

Appendix C. Checking (2.7) in the i.i.d. case

Let $\mathbf{Y} = (Y_1, Y_2, \dots)^\top$ be a sequence of positive i.i.d. EGPD random variables, $Y_i \sim \text{EGPD}(\sigma, \kappa, \xi, B)$ with $\kappa > 0$ and $\xi > 0$.

Case 1. $T(\mathbf{Y}) = \max(Y_1, \dots, Y_d)$ and $B(0) = 0$.

In this case, we can easily check that $\Pr(Y_1 = 0) = \Pr(T(\mathbf{Y}) = 0) = 0$ and $\Pr(T(\mathbf{Y}) \leq y) = \Pr(Y_1 \leq y)^d$ for $y > 0$. This implies that

$$\frac{\Pr(0 < T(\mathbf{Y}) \leq y)}{\Pr(0 < Y_1 \leq y)^d} = \frac{\Pr(T(\mathbf{Y}) \leq y)}{\Pr(Y_1 \leq y)^d} = \frac{\Pr(Y_1 \leq y)^d}{\Pr(Y_1 \leq y)^d} = 1.$$

It shows that (2.7) is satisfied with $\gamma = d$ and $\beta = 1$.

Case 2. $T(\mathbf{Y}) = \max(Y_1, \dots, Y_d)$ and $B(0) > 0$.

For $y > 0$

$$\begin{aligned} \frac{\Pr(0 < T(\mathbf{Y}) \leq y)}{\Pr(0 < Y_1 \leq y)} &= \frac{\Pr(T(\mathbf{Y}) \leq y) - \Pr(T(\mathbf{Y}) = 0)}{\Pr(0 < Y_1 \leq y)} \\ &= \frac{\Pr(Y_1 \leq y)^d - \Pr(Y_1 = 0)^d}{\Pr(0 < Y_1 \leq y)} \\ &= \frac{(\Pr(Y_1 = 0) + \Pr(0 < Y_1 \leq y))^d - \Pr(Y_1 = 0)^d}{\Pr(0 < Y_1 \leq y)} \\ &\rightarrow d \Pr(Y_1 = 0)^{d-1}, \text{ when } y \rightarrow 0^+ \end{aligned}$$

It shows that (2.7) is satisfied with $\gamma = 1$ and $\beta = dB(0)^{d-1}$.

Case 3. $T(\mathbf{Y}) = \sum_{i=1}^N Y_i$, with $B(0) = 0$, and N is an integer value random variable, independent of \mathbf{Y} , such that $\Pr(N = 1) > 0$.

For $y > 0$, we have

$$\Pr(0 < T(\mathbf{Y}) \leq y) = \sum_{n \geq 1} \Pr\left(\sum_{i=1}^n Y_i \leq y\right) \Pr(N = n)$$

and thus

$$\frac{\Pr(0 < T(\mathbf{Y}) \leq y)}{\Pr(Y_1 < y)} = \Pr(N = 1) + \sum_{n=2}^{\infty} \frac{\Pr(\sum_{i=1}^n Y_i < y)}{\Pr(Y_1 < y)} \Pr(N = n).$$

Using the upper bound

$$\Pr\left(\sum_{i=1}^n Y_i \leq y\right) \leq \prod_{i=1}^n \Pr(Y_i \leq y) = \Pr(Y_1 \leq y)^n$$

for $n \geq 2$, it follows that

$$\begin{aligned}
\sum_{n=2}^{\infty} \frac{\Pr(\sum_{i=1}^n Y_i \leq y)}{\Pr(Y_1 \leq y)} \Pr(N = n) &\leq \sum_{n=2}^{\infty} \Pr(Y_1 \leq y)^{n-1} \Pr(N = n) \\
&\leq \Pr(Y_1 \leq y) \sum_{n=2}^{\infty} \Pr(N = n), \\
&\leq \Pr(Y_1 \leq y) \\
&\rightarrow 0, \text{ when } y \rightarrow 0^+.
\end{aligned}$$

Finally, we deduce that

$$\lim_{y \rightarrow 0^+} \frac{\Pr(0 < T(\mathbf{Y}) \leq y)}{\Pr(0 < Y_1 \leq y)} = \Pr(N = 1). \quad (\text{C.1})$$

Hence (2.7) holds true with $\gamma = 1$ and $\beta = \Pr(N = 1)$. \square

Case 4. Assume that $T(\mathbf{Y}) = \sum_{i=1}^d Y_i$ with $\Pr(Y_i = 0) = B(0) > 0$. Let $\mathbf{Y}^+ = (Y_1^+, Y_2^+, \dots)^\top$ denote an i.i.d. sequence such that $Y_i^+ \sim \text{EGPD}(\sigma, \kappa, \xi, B^+)$ with

$$B^+(u) = \frac{B(u) - B(0)}{1 - B(0)}$$

(corresponding to the distribution of Y_i truncated on $(0, +\infty)$). Let $\mathbf{O} = (O_1, O_2, \dots)^\top$ be an i.i.d. sequence of Bernoulli random variable independent of \mathbf{Y} with $\Pr(O_i = 0) = B(0)$. Then it can be checked that $Y_i \stackrel{\mathcal{D}}{=} O_i Y_i^+$ and then

$$T(\mathbf{Y}) \stackrel{\mathcal{D}}{=} \sum_{i=1}^d O_i Y_i^+ \stackrel{\mathcal{D}}{=} \sum_{i=1}^N Y_i^+$$

with N the random number of non-null components in (O_1, \dots, O_d) . N is a binomial random variables independent of \mathbf{Y}^+ and thus (C.1) applies

$$\lim_{y \rightarrow 0^+} \frac{\Pr(0 < T(\mathbf{Y}) \leq y)}{\Pr(0 < Y_1^+ \leq y)} = \Pr(N = 1).$$

Using $\Pr(0 < Y_1^+ \leq y) = \Pr(0 < Y_1 \leq y)/(1 - B(0))$ and $\Pr(N = 1) = dB(0)(1 - B(0))^{d-1}$ we finally deduce that (2.7) holds true with $\gamma = 1$ and $\beta = dB(0)(1 - B(0))^{d-2}$. \square

Case 5. Assume that $T(\mathbf{Y}) = \sum_{i=1}^d Y_i$ with $\Pr(Y_i = 0) = B(0) = 0$.

The proof is based on the following lemma which characterizes the lower tail of the sum of two independent EGPD.

Lemma Appendix C.1. *Let X_1 and X_2 be two independent continuous positive random variables with pdf f_{X_i} . Assume that for $i \in \{1, 2\}$*

$$\lim_{x \rightarrow 0^+} \frac{f_{X_i}(x)}{x^{\kappa_i-1}} = K_i$$

with $K_i > 0$ and $\kappa_i > 0$. Then

$$\lim_{x \rightarrow 0^+} \frac{f_{X_1+X_2}(x)}{x^{\kappa_1+\kappa_2-1}} = K_1 K_2 \mathcal{B}(\kappa_1, \kappa_2)$$

where $f_{X_1+X_2}$ denotes the pdf of $X_1 + X_2$ and $\mathcal{B}(a, b)$ is the beta function $\mathcal{B}(a, b) = \int_0^1 v^{a-1}(1-v)^{b-1} dv$.

Proof of Lemma Appendix C.1. We need to study the limit in 0^+ of the ratio

$$\begin{aligned} \frac{f_{X_1+X_2}(x)}{x^{\kappa_1+\kappa_2-1}} &= \frac{1}{x^{\kappa_1+\kappa_2-1}} \int_0^x f_{X_1}(u) f_{X_2}(x-u) du \\ &= \frac{1}{x^{\kappa_1+\kappa_2-2}} \int_0^1 f_{X_1}(xv) f_{X_2}(x(1-v)) dv \text{ with } u = vx, \\ &= \int_0^1 \frac{f_{X_1}(xv)}{(xv)^{\kappa_1-1}} \frac{f_{X_2}(x(1-v))}{(x(1-v))^{\kappa_2-1}} v^{\kappa_1-1} (1-v)^{\kappa_2-1} dv, \end{aligned}$$

The dominated convergence is then applied to obtain

$$\begin{aligned} \lim_{x \rightarrow 0^+} \frac{f_{X_1+X_2}(x)}{x^{\kappa_1+\kappa_2-1}} &= \int_0^1 \lim_{x \rightarrow 0^+} \frac{f_{X_1}(xv)}{(xv)^{\kappa_1-1}} \lim_{x \rightarrow 0^+} \frac{f_{X_2}(x(1-v))}{(x(1-v))^{\kappa_2-1}} v^{\kappa_1-1} (1-v)^{\kappa_2-1} dv \\ &= K_1 K_2 \int_0^1 v^{\kappa_1-1} (1-v)^{\kappa_2-1} dv \\ &= K_1 K_2 \mathcal{B}(\kappa_1, \kappa_2) \end{aligned}$$

□

According to (A.3), the pdf f of Y_i satisfies

$$\lim_{y \rightarrow 0^+} \frac{f(y)}{y^{\kappa-1}} = \kappa \frac{b(0^+)}{\sigma^\kappa}.$$

Then using Lemma [Appendix C.1](#) and reasoning by recurrence, we deduce that the pdf f_d of $T(\mathbf{Y}) = \sum_{i=1}^d Y_i$ is such that

$$\lim_{y \rightarrow 0^+} \frac{f_d(y)}{y^{d\kappa-1}} = \frac{b(0^+)^d \Gamma(\kappa+1)^d}{\sigma^{\kappa d} \Gamma(d\kappa)}.$$

Using L'Hôpital rule, we deduce that

$$\begin{aligned} \lim_{y \rightarrow 0^+} \frac{\Pr(Y_1 + \dots + Y_d \leq y)}{y^{d\kappa}} &= \lim_{y \rightarrow 0^+} \frac{f_d(y)}{(d\kappa)y^{d\kappa-1}} \\ &= \frac{b(0^+)^d \Gamma(\kappa+1)^d}{\sigma^{\kappa d} \Gamma(d\kappa+1)}. \end{aligned}$$

Using again [\(A.3\)](#), we obtain the following result

$$\begin{aligned} \lim_{y \rightarrow 0^+} \frac{\Pr(Y_1 + \dots + Y_d \leq y)}{\Pr(Y_1 < y)^d} &= \lim_{y \rightarrow 0^+} \frac{\Pr(Y_1 + \dots + Y_d \leq y)}{y^{d\kappa}} \frac{y^{d\kappa}}{\Pr(Y_1 < y)^d} \\ &= \frac{\Gamma(\kappa+1)^d}{\Gamma(d\kappa+1)}. \end{aligned}$$

This proves that [\(2.7\)](#) is satisfied with $\gamma = d$ and $\beta = \Gamma(\kappa+1)^d / \Gamma(d\kappa+1)$.

Appendix D. Proof of Proposition [2.4](#)

Let $T = \sum_{i=1}^N Y_i^*$ be a compound Poisson-EGPD random variable with the notations and assumptions of [Definition 2.3](#). Then T satisfies the conditions of [Proposition \(2.2\)](#); see [\[38\]](#) for the upper tail condition [2.6](#), which holds true with $\alpha = E[N] = \lambda$, and [\(C.1\)](#) for the lower tail condition [\(2.7\)](#), which holds true with $\gamma = 1$ and $\beta = \Pr(N = 1) = \lambda \exp(-\lambda)$. This implies that $T \sim EGPD(\sigma, \kappa, \xi, B_\lambda)$ with $B_\lambda(0) = \exp(-\lambda)$, $b_\lambda(0^+) = \lambda \exp(-\lambda)b(0^+)$ and $b_\lambda(1) = \lambda b(1)$.

Conversely, let $T \sim EGPD(\sigma, \kappa, \xi, B_T)$ be an infinitely divisible random variable. Then according to [\[46, Theorem 3.2\]](#), T can be written as a compound Poisson distribution, $T = \sum_{i=1}^N Y_i^*$, with N a Poisson random variable and $\mathbf{Y}^* = (Y_1^*, Y_2^*, \dots)$ an i.i.d. sequence of positive random variable independent of N . It remains to prove that Y_i^* is EGPD. According to [\[39, Lemma 3.7.\]](#),

$$\lim_{y \rightarrow \infty} \frac{\Pr(T > y)}{\Pr(Y_1^* > y)} = E[N]$$

and (C.1) implies that the lower tail of T and Y_i^* are also equivalent

$$\lim_{y \rightarrow 0^+} \frac{\Pr(0 < T \leq y)}{\Pr(0 < Y_1^* \leq y)} = \Pr(N = 1) \exp(-\lambda).$$

Using Proposition 2.2 we deduce that $Y_i^* \sim EGPD(\sigma, \kappa, \xi, B)$ (remark that the role of T and Y_i are symmetric in Proposition 2.2). \square

Appendix E. Proof of Proposition 3.1

Let Y_1, Y_2, \dots be a non-negative positive random and let $u > 0$. Remark that

$$\Pr(Y_1 \geq u) = \Pr(A_1 \geq u) \Pr(Y_1 > 0) \quad (\text{E.1})$$

where $A_1 = Y_1^+$ denotes the positive part of Y_1 . Under assumption (3.2) we have

$$A_1 \stackrel{\mathcal{D}}{=} \sigma_1 \sum_{i=1}^{N_1^+} X_i$$

where $\mathbf{X} = (X_1, X_2, \dots)^\top$ denotes a sequence of positive i.i.d. random variables with $X_i \sim EGPD(1, \kappa, \xi, B)$ and $B(u) = u$ and N_1^+ the positive part of a Poisson random variable N_1 with mean λ_1 independent of \mathbf{X} .

Similarly, we have

$$\Pr(Y_d \geq u) = \Pr(A_d \geq u) \Pr(Y_1 + \dots + Y_d > 0) \quad (\text{E.2})$$

with

$$A_d \stackrel{\mathcal{D}}{=} \sigma_d \sum_{i=1}^{N_d^+} X_i$$

where N_d is a Poisson random variable with mean λ_d independent of \mathbf{X} .

Under the assumptions of Proposition 3.1, we have $\lambda_1 \leq \lambda_d$ and $\sigma_1 \leq \sigma_d$. It implies that N_1 is smaller than N_d in the likelihood ratio order, i.e. the ratio $\Pr(N_d = n)/\Pr(N_1 = n)$ is increasing in n . Using [47, Theorem 1.C.6.] we deduce that N_1^+ is smaller than N_d^+ in the likelihood ratio order and thus also in the usual stochastic order. Then [47, Theorem 1.A.4] implies that A_1 is smaller than A_d in the usual stochastic order, meaning that for any $u > 0$

$$\Pr(A_1 \geq u) \leq \Pr(A_d \geq u). \quad (\text{E.3})$$

Combining the inequality

$$\Pr(Y_1 > 0) \leq \Pr(Y_1 + \dots + Y_d > 0)$$

with equations (E.1), (E.2), (E.3), we deduce that (1.1) holds true.

Appendix F. Some properties of $EGPD(\sigma, \kappa, \xi, \lambda)$

Let $\mathbf{Y}^* = (Y_1^*, Y_2^*, \dots)^\top$ be a sequence of i.i.d. EGPD random variables, $Y_i^* \sim EGPD(\sigma, \kappa, \xi, B)$, with $B(u) = u$, $\kappa > 0$ and $\xi > 0$. Let N be a Poisson random variable with mean λ independent of \mathbf{Y}^* . We denote $Y = \sum_{i=1}^N Y_i^*$ and $A = Y^+$ the positive part of this distribution. According to Definition 2.5 we have $A \sim EGPD(\sigma, \kappa, \xi, \lambda)$.

Remark that for $y > 0$,

$$\begin{aligned} P(A \leq y) &= P(Y \leq y | Y > 0) \\ &= \frac{P(Y \leq y)}{P(Y > 0)}, \end{aligned}$$

and thus

$$P(A \leq y) = \frac{P(Y \leq y)}{1 - \exp(-\lambda)}. \quad (\text{F.1})$$

Properties of A , such as its first two moments for example, can then be deduced from the general properties of compound Poisson distributions [see, e.g. 48]. The pdf of A is of particular interest for this work since it appears in the definition of composite likelihood function 4.1. Several methods have been proposed in the literature to compute numerical approximations of the pdf of compound Poisson distributions [41]. In this work, we use the Panjer recursions [42], which takes a discrete distribution as input.

We thus first replace Y_i^* with its discrete version $E_i^*(h)$ concentrated on $\{h, 2h, \dots, \}$ where $h = 0.2$ is the built-in precision of existing precipitation gauges in Section 4. Using the general expression (A.1) of the cdf of the EGPD distribution with $B(u) = u$, we obtain the probability function of $E_i^*(h)$

$$f_j = \Pr(E_i^*(h) = jh) = H_\xi^\kappa\left(\frac{jh}{\sigma}\right) - H_\xi^\kappa\left(\frac{(j-1)h}{\sigma}\right) \quad (\text{F.2})$$

for $j \in \{1, 2, \dots\}$. For compound Poisson distribution, the recursive Panjer formula for $p_a = \Pr(\sum_{i=1}^{N_a} E_i^*(h) = ah)$, $a \in \{0, 2, \dots\}$, reduces to

$$p_a = \begin{cases} \exp(-\lambda) & \text{if } a = 0 \\ (\lambda/a) \sum_{j=1}^a j f_j p_{a-j} & \text{if } a \geq 1 \end{cases} \quad (\text{F.3})$$

Appendix G. Simulation results

In this Appendix, synthetic rainfall data at the finer time scale is simulated as an i.i.d. sequence of a compound Poisson-EGPD, see Definition 2.3

with $B(u) = u$. In this idealized setting, it can be checked that the positive aggregated data defined by (3.1) satisfy (3.2) and (3.3) where $\sigma_d = \sigma$ and $\lambda_d = d\lambda$ are log-polynomials of order $p = 0$ and $q = 1$. To mimic our application setup, the parameters are fixed to $\kappa = .3$, $\xi = .25$, $\sigma = 1$, $\lambda = .01$, the sample size is equivalent to 36 months of 6 minute rainfall data and we let $p = q = 3$ in the estimation procedure described.

Figure G.8 displays our estimates of ξ , κ , σ_d and λ_d in the top left, top right, bottom left and bottom right panels, respectively. Overall, the parameters are well estimated.

References

- [1] D. Hershfield, Rainfall frequency atlas of the United States, Technical Report TP-40, U.S. Weather Bureau (1961).
- [2] A. Haruna, J. Blanchet, A.-C. Favre, Modeling intensity-duration-frequency curves for the whole range of non-zero precipitation: a comparison of models, *Water Resources Research* 59 (6) (2023) e2022WR033362.
- [3] J. Ulrich, O. E. Jurado, M. Peter, M. Scheibel, H. W. Rust, Estimating IDF curves consistently over durations with spatial covariates, *Water* 12 (11:3119) (2020).
- [4] D. Koutsoyiannis, [Stochastics of Hydroclimatic Extremes: A Cool Look at Risk](#), 4th Edition, Kallipos Open Academic Editions, Athens, Greece, 2024.
URL <https://www.itia.ntua.gr/en/docinfo/2000/>
- [5] R. Katz, M. Parlange, P. Naveau, Statistics of extremes in hydrology, *Advances in Water Resources* 25 (2002) 1287–1304.
- [6] S. Coles, *An Introduction to Statistical Modeling of Extreme Values*, London: Springer, 2001.
- [7] J. Beirlant, Y. Goegebeur, J. Teugels, J. Segers, *Statistics of Extremes*, Wiley, Chichester, 2004.
- [8] J. Nair, A. Wierman, B. Zwart, *The Fundamentals of Heavy Tails: Properties, Emergence, and Estimation*, Vol. 53, Cambridge University Press, 2022.
- [9] W. Feller, The asymptotic distribution of the range of sums of independent random variables, *Annals of Mathematical Statistics* 22 (1951) 427–432.
- [10] A.-L. Fougères, C. Mercadier, Risk measures and multivariate extensions of Breiman’s theorem, *Journal of Applied Probability* 49 (2) (2012) 364–384.

- [11] P. Naveau, R. Huser, P. Ribereau, A. Hannart, Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection, *Water Resources Research* 52 (4) (2016) 2753–2769.
- [12] I. Papastathopoulos, J. A. Tawn, Extended generalised pareto models for tail estimation, *Journal of Statistical Planning and Inference* 143 (1) (2013) 131–143.
- [13] P. Gamet, J. Jalbert, A flexible extended generalized Pareto distribution for tail estimation, *Environmetrics* 33 (6) (2022) e2744.
- [14] G. Evin, A.-C. Favre, B. Hingray, Stochastic generation of multi-site daily precipitation focusing on extreme events, *Hydrology and Earth System Sciences* 22 (1) (2018) 655–672.
- [15] M. Taillardat, A.-L. Fougères, P. Naveau, O. Mestre, Forest-based and semiparametric methods for the postprocessing of rainfall ensemble forecasting, *Weather and Forecasting* 34 (3) (2019) 617–634.
- [16] M. Taillardat, O. Mestre, From research to applications—examples of operational ensemble post-processing in France using machine learning, *Nonlinear Processes in Geophysics* 27 (2) (2020) 329–347.
- [17] P. Rivoire, O. Martius, P. Naveau, A comparison of moderate and extreme ERA-5 daily precipitation with two observational data sets, *Earth and Space Science* 8 (4) (2021) e2020EA001633.
- [18] P. Rivoire, P. Le Gall, A.-C. Favre, P. Naveau, O. Martius, High return level estimates of daily ERA-5 precipitation in Europe estimated using regionalized extreme value distributions, *Weather and Climate Extremes* 38 (2022) 100500.
- [19] P. Le Gall, A.-C. Favre, P. Naveau, C. Prieur, Improved regional frequency analysis of rainfall data, *Weather and Climate Extremes* (2022) 100456.
- [20] M. A. A. Turkman, K. F. Turkman, P. de Zea Bermudez, S. Pereira, P. Pereira, M. de Carvalho, Calibration of the bulk and extremes of spatial data, *REVSTAT-Statistical Journal* 19 (3) (2021) 309–325.

- [21] J. Legrand, P. Ailliot, P. Naveau, N. Raillard, Joint stochastic simulation of extreme coastal and offshore significant wave heights, *Annals of Applied Statistics* 17 (4) (2023) 3363–3383.
- [22] M. de Carvalho, S. Pereira, P. Pereira, P. de Zea Bermudez, An extreme value Bayesian Lasso for the conditional left and right tails, *Journal of Agricultural, Biological and Environmental Statistics* 27 (2) (2022) 222–239.
- [23] N. L. Carrer, C. Gaetan, Distributional regression models for Extended Generalized Pareto distributions (2022). [arXiv:2209.04660](https://arxiv.org/abs/2209.04660).
- [24] D. Koutsoyiannis, D. Kozonis, A. Manetas, A mathematical framework for studying rainfall intensity-duration-frequency relationships, *Journal of Hydrology* 206 (1-2) (1998) 118–135.
- [25] K. Revfeim, An initial model of the relationship between rainfall events and daily rainfalls, *Journal of Hydrology* 75 (1-4) (1984) 357–364.
- [26] C. Thompson, Homogeneity analysis of rainfall series: an application of the use of a realistic rainfall model, *Journal of Climatology* 4 (6) (1984) 609–619.
- [27] P. K. Dunn, Occurrence and quantity of precipitation can be modelled simultaneously, *International Journal of Climatology* 24 (10) (2004) 1231–1239.
- [28] M. M. Hasan, P. K. Dunn, Two Tweedie distributions that are near-optimal for modelling monthly rainfall in Australia, *International Journal of Climatology* 31 (9) (2011) 1389–1397.
- [29] R. M. Yunus, M. M. Hasan, N. A. Razak, Y. Z. Zubairi, P. K. Dunn, Modelling daily rainfall with climatological predictors: Poisson-gamma generalized linear modelling approach, *International Journal of Climatology* 37 (3) (2017) 1391–1399.
- [30] N. C. Dzipire, P. Ngare, L. Odongo, A Poisson-Gamma model for zero inflated rainfall data, *Journal of Probability and Statistics* 2018 (1) (2018) 1012647.

- [31] J. Park, D. Cross, C. Onof, Y. Chen, D. Kim, A simple scheme to adjust Poisson cluster rectangular pulse rainfall models for improved performance at sub-hourly timescales, *Journal of Hydrology* 598 (2021) 126296.
- [32] D. Koutsoyiannis, Climate change, the hurst phenomenon, and hydrological statistics, *Hydrological Sciences Journal* 48 (2003) 3–24.
- [33] A. MacDonald, C. Scarrott, D. Lee, B. Darlow, M. Reale, G. Russell, A flexible extreme value mixture model, *Computational Statistics & Data Analysis* 55 (2011) 2137–2157.
- [34] M. Boutigny, P. Ailliot, P. Naveau, B. Saussol, A. Chaubet, A meta-gaussian distribution for sub-hourly rainfall, *Stochastic Environmental Research and Risk Assessment* (2023) 3915–3927.
- [35] J. Blanchet, E. Paquet, P. Ayar, D. Penot, An objective cross-validation framework for mapping rainfall hazard based on rain gauge data, *Hydrology and Earth System Sciences Discussions* 23 (2018) 829–849.
- [36] C. Li, V. P. Singh, A. K. Mishra, Simulation of the entire range of daily precipitation using a hybrid probability distribution, *Water Resources Research* 48 (3) (2012).
- [37] D. Koutsoyiannis, P. Dimitriadis, F. Lombardo, S. Stevens, *From Fractals to Stochastics: Seeking Theoretical Consistency in Analysis of Geophysical Data*, Springer International Publishing, 2017, p. 237–278.
- [38] L. Breiman, On some limit theorems similar to the arc-sin law, *Theory of Probability & Its Applications* 10 (1965) 323–331.
- [39] H. A. Jessen, T. Mikosch, Regularly varying functions, *Publications de L’institut Mathématique* 80 (94) (2006) 171–192.
- [40] I. Rodriguez-Iturbe, D. R. Cox, V. Isham, Some models for rainfall based on stochastic point processes, *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences* 410 (1839) (1987) 269–288.
- [41] P. Embrechts, M. Frei, Panjer recursion versus fft for compound distributions, *Mathematical Methods of Operations Research* 69 (2009) 497–508.

- [42] H. H. Panjer, Recursive evaluation of a family of compound distributions, *ASTIN Bulletin* 12 (1) (1981) 22–26.
- [43] J. Ulrich, F. S. Fauer, H. W. Rust, Modeling seasonal variations of extreme rainfall on different timescales in Germany, *Hydrology and Earth System Sciences* 25 (12) (2021) 6133–6149.
- [44] H. Van de Vyver, G. R. Demarée, Construction of Intensity–Duration–Frequency (IDF) curves for precipitation at Lubumbashi, Congo, under the hypothesis of inadequate data, *Hydrological Sciences Journal–Journal des Sciences Hydrologiques* 55 (4) (2010) 555–564.
- [45] R. E. Chandler, S. Bate, Inference for clustered data using the independence loglikelihood, *Biometrika* 94 (1) (2007) 167–183.
- [46] F. W. Steutel, K. Van Harn, *Infinite Divisibility of Probability Distributions on the Real Line*, CRC Press, 2003.
- [47] M. Shaked, J. G. Shanthikumar, *Stochastic Orders*, Springer, 2007.
- [48] S. M. Ross, *Introduction to Probability Models*, Academic Press, 2014.

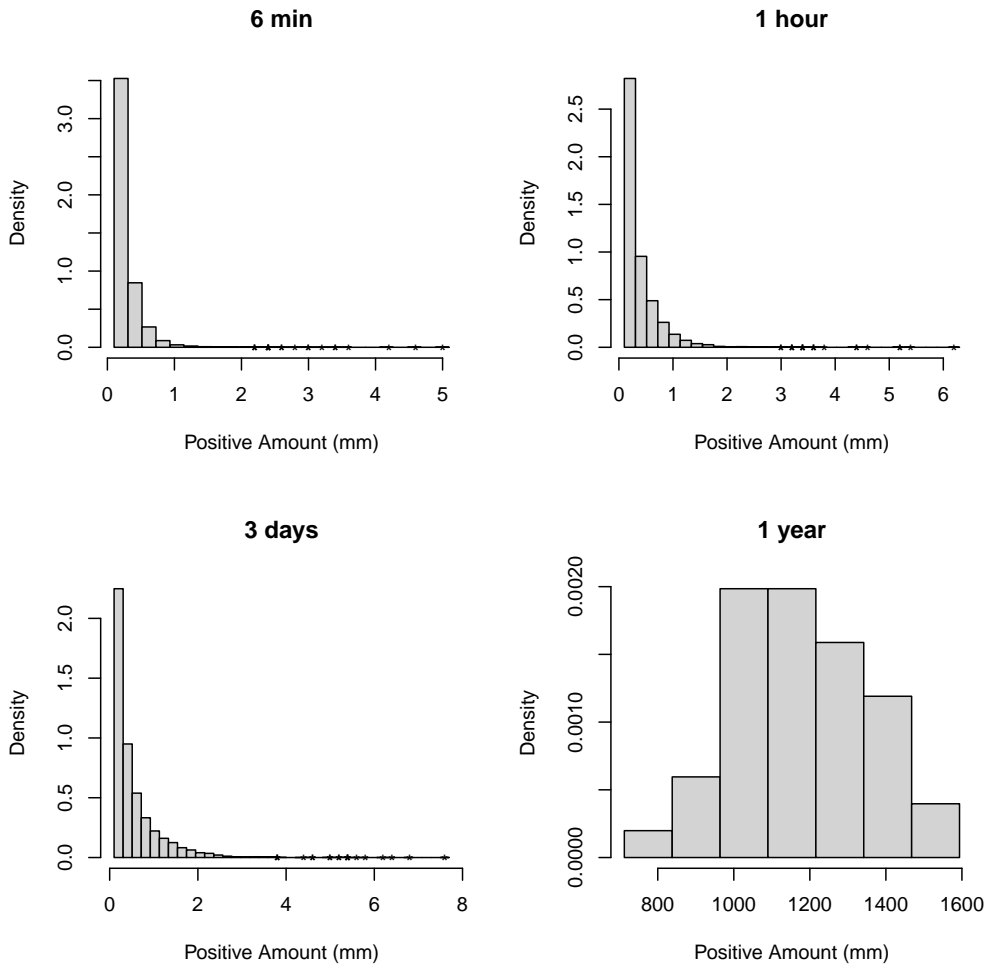


Figure 1: Histogram of positive rainfall (i.e. after removing dry) aggregated over different time scales in Brest. The stars on the x-axis correspond to the 20 more extremes observations. The yearly histogram is plotted for illustration purpose, the modeling of such aggregation scale is not discussed in the paper. The results at the annual scale were obtained using 80 years of daily data downloaded from <https://www.ecad.eu/>, whereas the other histograms are based on 18 years of 6-minute data for August-September described in Section 4.2.

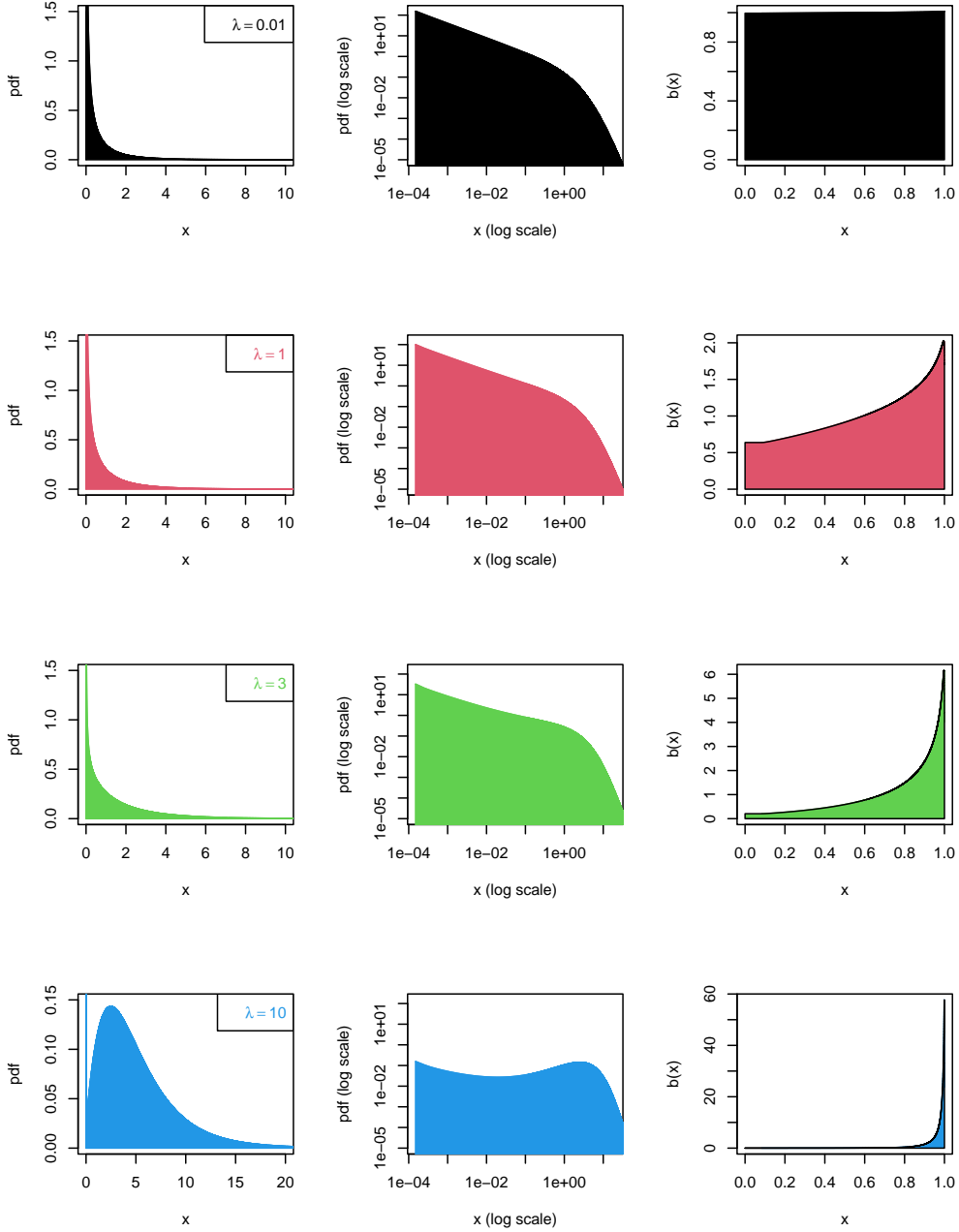


Figure 2: Left panels: pdf of the $EGPD(\sigma, \kappa, \xi, \lambda)$, see Definition 2.5, with $\kappa = 0.3$, $\sigma = 1$, $\xi = 0.25$ and various values of the mean λ given in the legend. Middle panels: same as the left panels with log scale on both axis. Right panels: corresponding pdfs $b_\lambda(\cdot)$. All the pdfs are numerically approximated using the Panjer recursions, see Appendix F

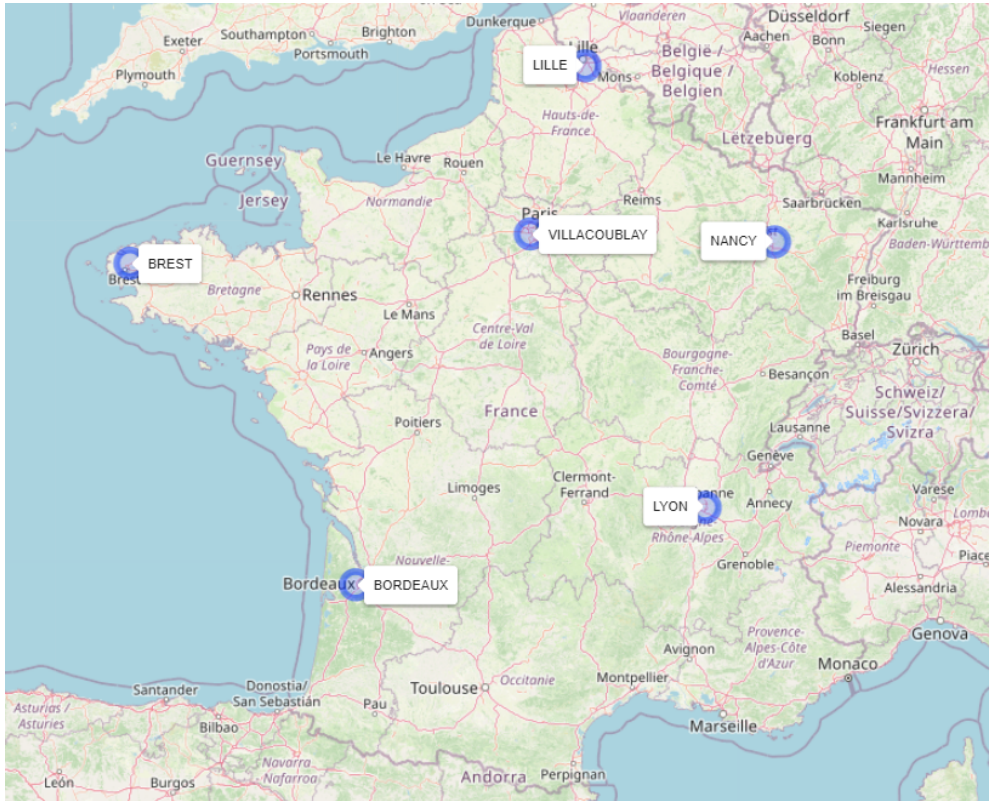


Figure 3: Locations of the meteorological stations considered in this study. Météo-France recorded 6-minute rainfall time series for the period 2006-2023 (with a tipping bucket precision of 0.2 mm).

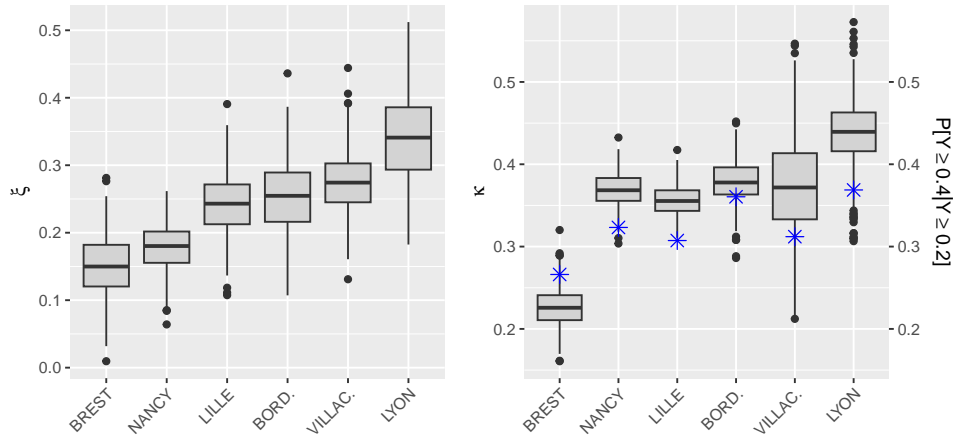


Figure 4: Estimation of ξ (left panel) and κ (right panel) of model (3.2) applied to the eight locations shown in Figure 3. The x-axis is ordered with respect to the estimated value of ξ . The boxplots are computed using block bootstrap. The blue stars on the right panel represent the empirical estimate of $P[Y \geq 0.4] = 1 - P[Y = 0.2]$ where Y denotes the 6-minutes rainfall data. Results for August-September based on 18 years of six-minutes data.

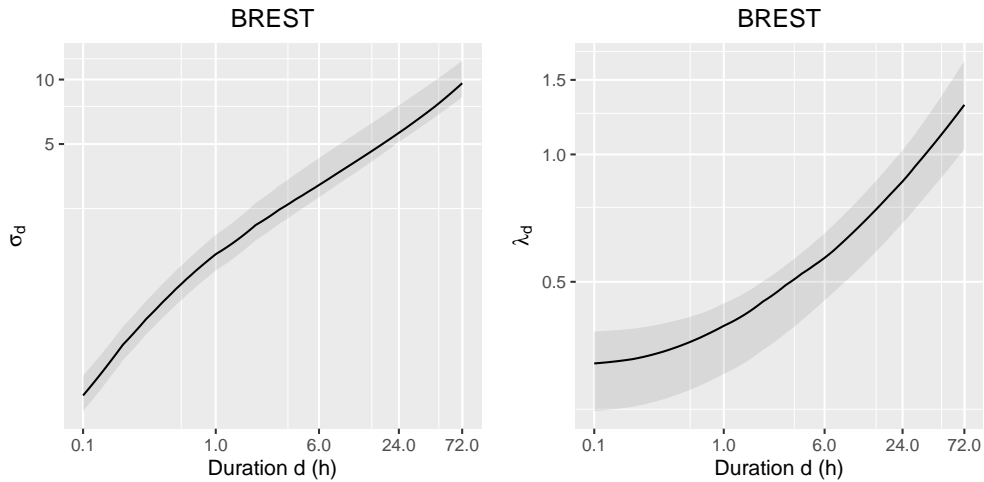


Figure 5: Estimation of σ_d (left panel) and λ_d (right panel) as a function of the duration of aggregation d in Brest. The 95% confidence bands were computed using block bootstrap. Results for August-September based on 18 years of 6-minute data.

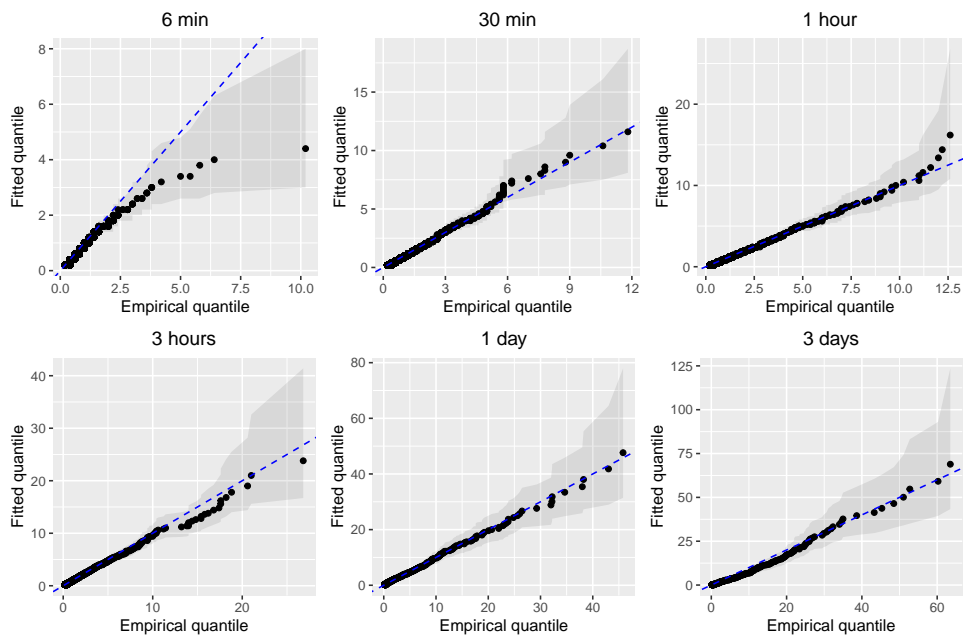


Figure 6: Quantile-quantile plots of the fitted model at different aggregation time in Brest. The 95% confidence bands were computed using block bootstrap. Results for August-September based on 18 years of 6-minute data.

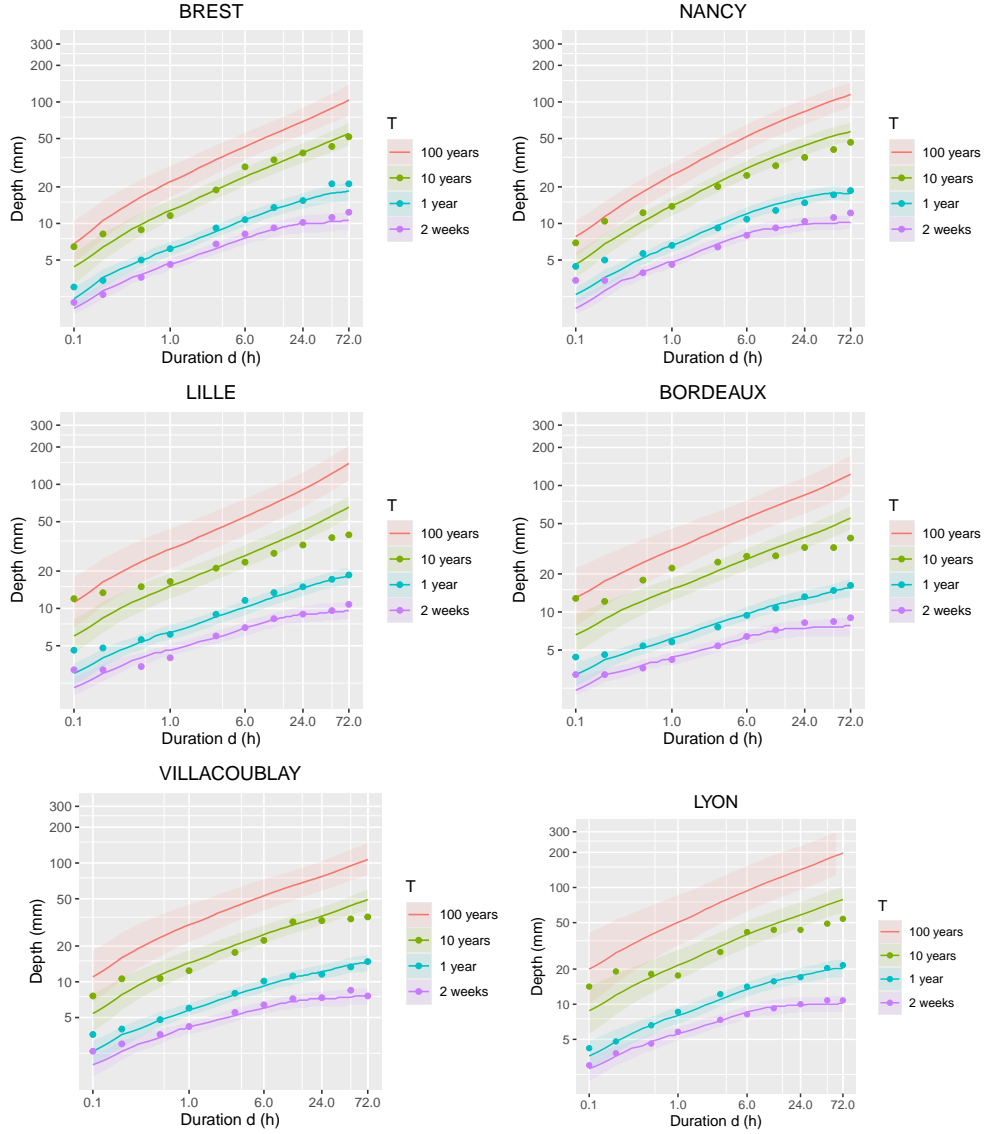


Figure 7: Quantiles of order $1 - \frac{1}{n_d T}$ as a function of d with n_d the number of observations available per month at the duration of aggregation d . The 95% confidence bands were computed using block bootstrap. The points represent the corresponding empirical quantiles (only for $T \leq 10$ years). Results for August-September based on 18 years of 6-minute data.

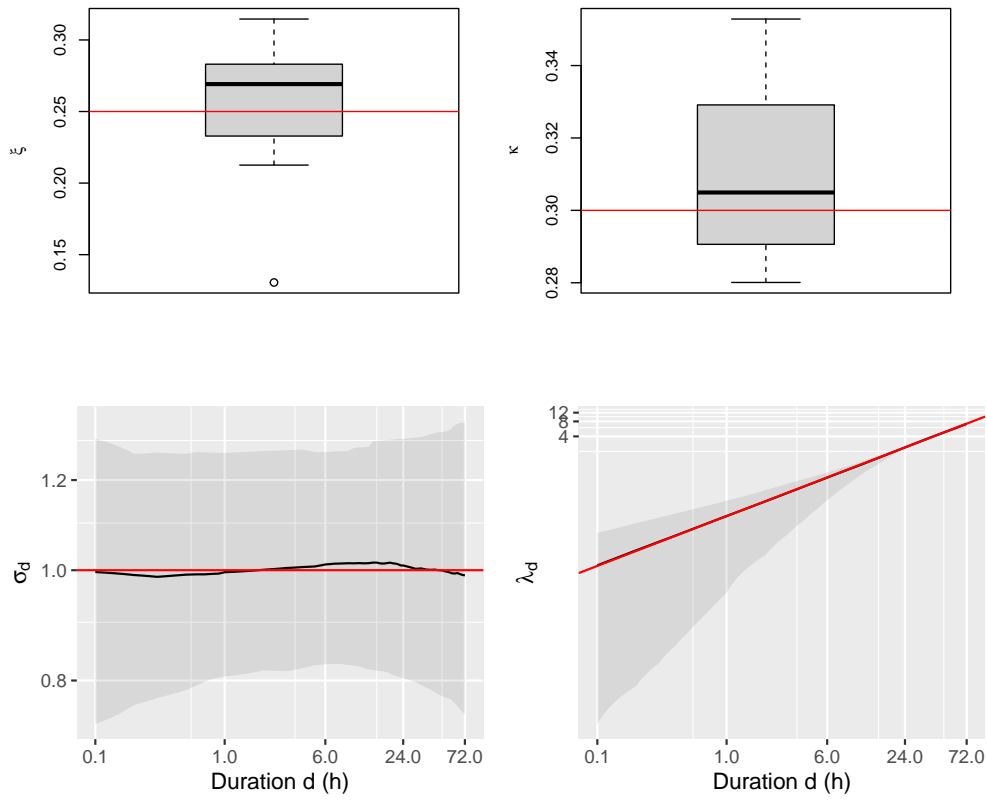


Figure G.8: Empirical distribution of the parameter estimates obtained using the simulation setup described in [Appendix G](#). The red lines correspond to the truth. On the bottom plots, the black lines correspond to the median, grey areas to 95% fluctuations interval computed by repeating the experiment 500 time.