mathematics

MDPI

Article

A New Two-Component Hybrid Model for Highly Right-Skewed Data: Estimation Algorithm and Application to Rainfall Data from South Tyrol, Italy

Patrick Osatohanmwen

Faculty of Economics and Management, Free University of Bolzano, 39100 Bolzano, Italy; profpato2014@gmail.com or patrick.osatohanmwen@unibz.it

Abstract

In many real-life processes, data with high positive skewness are very common. Moreover, these data tend to exhibit heterogeneous characteristics in such a manner that using one parametric univariate probability distribution becomes inadequate to model such data. When the heterogeneity of such data can be appropriately separated into two components—the main innovation component, where the bulk of data is centered, and the tail component which contains some few extreme observations—in such a way, and without a loss in generality, that the data possesses high skewness to the right, the use of hybrid models to model the data becomes very viable. In this paper, a new two-component hybrid model which combines the half-normal distribution for the main innovation of a positive and highly right-skewed data with the generalized Pareto distribution (GPD) for the observations in the data above a certain threshold is proposed. To enhance efficiency in the estimation of the parameters of the hybrid model, an unsupervised iterative algorithm (UIA) is adopted. The hybrid model is applied to model the intensity of rainfall which triggered some debris flow events in the South Tyrol region of Italy. Results from Monte Carlo simulations, as well as from the model's application to the real data, clearly show how the UIA enhances the estimation of the free parameters of the hybrid model to offer good fits to positive and highly right-skewed data. Application results of the hybrid model are also compared with the results of other two-component hybrid models and graphical threshold selection methodologies in extreme value theory.

Keywords: estimation algorithm; generalized Pareto distribution; half-normal distribution; hybrid model; rainfall data

MSC: 60G70; 62E20; 62F35; 62P05; 62P10; 65D15; 68W4

JEL Classification: C02

check for updates

Academic Editors: David Carfì and Heng Lian

Received: 20 June 2025
Revised: 22 August 2025
Accepted: 8 September 2025
Published: 16 September 2025

Citation: Osatohanmwen, P. A New Two-Component Hybrid Model for Highly Right-Skewed Data: Estimation Algorithm and Application to Rainfall Data from South Tyrol, Italy. *Mathematics* **2025**, *13*, 2987. https://doi.org/10.3390/math13182987

Copyright: © 2025 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Skewed distributions are common in real-world processes, particularly when extreme values or outliers significantly influence standard statistical approaches. The analysis of skewed data is critical in various fields, including finance and insurance [1–4], communication and signal processing [5–9], and environmental science [10–14]. In many instances, data characterized by high positive skewness can exhibit heterogeneity, where a single parametric univariate probability distribution falls short in effectively modeling such data.

When dealing with highly right-skewed data, it is necessary to understand the asymmetry of observed data, which can be decomposed into two or more distinct components [15]. For the case of two components, the first component is the main innovation, representing where the bulk of the data is centered, while the second is the tail component, which captures the few extreme observations above a certain threshold that contribute to the skewness. Failing to account for this heterogeneity could lead to significant loss in efficiency when modeling the data. Moreover, the modeling of this type of data has received widespread attention in recent years, and many models/methods have been put forth, including non-parametric models [16] and models following the Peak over Threshold (PoT) methodology [11,17]. While non-parametric models usually offer good fits to the data, they typically fail to account for a few outlying observations in the tail. On the other hand, the PoT methodology only focuses on the extreme observations beyond a certain threshold while ignoring the remaining observations in the data, and thus, it does not make use of the entire distribution. Given these challenges, hybrid models have emerged as effective solutions for modeling complex distributions. A hybrid model combines two or more probability distributions to adequately fit the characteristics of observed data. In [15], a general framework for generating multi-component hybrid models was developed. Moreover, several families of two-component hybrid models have been defined and studied in the literature [3,13,18–24].

In this paper, a new two-component hybrid model for data exhibiting high skewness to the right is introduced. The model links a half-normal distribution for the bulk of the data distribution and a GPD for the tail after a certain threshold point determined when imposing a condition of class $\mathcal{C}^1$ [3,25]. The choice of using the half-normal distribution as the bulk distribution, in contrast to using distributions such as the normal, gamma, and lognormal ones, is motivated by the fact that its support on $[0, \infty)$ naturally aligns with strictly positive data with zeroes, common in many real-world measurements, avoiding the conceptual and practical issues of using a normal distribution that extends into negative values. The half-normal distribution also captures moderate skewness with a light, exponentially decaying tail, creating a clear boundary between the bulk and the GPD tail component, a situation crucial for reliable threshold selection and accurate tail index estimation. Compared to the gamma or lognormal distributions, the half-normal distribution is more parsimonious, having only a single scale parameter, which reduces overfitting risk and minimizes parameter identifiability conflicts with the tail component. Its relatively fast tail decay prevents 'tail leakage', where bulk observations intrude into the extremes, and its closed-form expressions for key properties simplify estimation and inference. Furthermore, its interpretability as the distribution of the absolute value of a normal variate makes it particularly appealing in cases where data represents magnitudes of symmetric underlying processes. These qualities, together, make the half-normal model a robust, stable, and theoretically well-matched partner to the GPD model in hybrid modeling frameworks. Furthermore, the two components of the hybrid half-normal–GPD model are weighted non-uniformly, and an unsupervised iterative and convergent estimation scheme based on the Levenberg–Marquardt (L-M) algorithm [26,27] is adopted to estimate the threshold point and other free parameters of the two-component model. In [15] a maximum likelihood-based algorithm was used to estimate the parameters of multi-component hybrid models, and the challenge of selecting initial values for the parameters before the maximization of the log-likelihood function was bypassed by the use of the unsupervised iterative and convergent estimation scheme presented in this paper, since the scheme is self-calibrating. Also, in standard PoT methodology, the threshold point is usually estimated graphically, whereas in this framework, it is a parameter to be estimated in the hybrid model. This allows us to determine the point beyond which the extremes are observed algorithmically.

In Section 2, the specification of the two-component non-uniform weights hybrid model framework and the half-normal–GPD model is carried out. A description of the UIA for the estimation of the hybrid model's free parameters is presented in Section 3. Results from numerical studies based on Monte Carlo simulations conducted to assess the UIA's efficiency in estimating the hybrid model's parameters are reported in Section 4. In Section 5, application of the new hybrid model to fit data on the intensity of rainfall that triggered some debris flow events in the South Tyrol region of Italy is performed. The paper closes with discussions and a conclusion in Section 6.

2. Two-Component Non-Uniform Weight Hybrid Model

Suppose we have data which can be decomposed into two components and the components represent specific behavior of the dichotomized data, while the goal is to use a smooth piecewise probability density function (pdf) to model the data. Assume the data is continuous and follows a non-degenerate distribution. Let f_1 and f_2 be two pdfs, each with parameter vector Θ_1 and Θ_2, in such a manner that each pdf is suitable for modeling specific components of the data and without a loss in generality. f_1 and f_2 are suitable for modeling the first and second components of the data, respectively. Suppose F_1 and F_2 are the respective cumulative distribution functions (cdfs) corresponding to f_1 and f_2, with respective corresponding quantile functions $Q_1(p;\Theta_1)$ and $Q_2(p;\Theta_2)$, where $Q_i(p;\Theta_i) = \inf\{w; F_i(w;\Theta_i) > p\}, 0 < p < 1$. The general two-component non-uniform weight hybrid model for the data can be specified by the pdf of the form

$$f(x;\Theta) = \begin{cases} w_1 f_1(x;\Theta_1), & -\infty < x \leq u, \\ w_2 f_2(x;\Theta_2), & u \leq x < \infty, \end{cases} \tag{1}$$

where Θ is a vector of free parameters in the model, w_1 and w_2 $(w_1 \neq w_2)$ are weights associated with the respective component of the pdf in Equation (1), and u is a junction point or threshold indicating the point of transition from one component or behavior of the data to another.

In modeling complex data distributions, particularly those with heterogeneous characteristics across the range, a hybrid approach offers greater flexibility and accuracy. The two-component model in Equation (1) assigns non-uniform weights to each component, as it is specifically designed to capture distinct features of the data, and hence, $w_1 \neq w_2$. To further justify the rationale behind the assignment of non-uniform weights in the hybrid model in Equation (1), we argue that the first component is weighted more heavily, as it targets the bulk or central mass of the distribution, and the second component is given less weight and is focused on modeling the extreme tail behavior, where outlier events occur. This non-uniform weighting reflects the natural asymmetry in many real-world data sets, where most observations cluster around a central trend and a smaller proportion of data lies in the tail. In contrast, a uniform weighting framework treats both components as contributing equally to the overall model regardless of the relative frequency or importance of the patterns they are meant to capture. While seemingly fair or neutral, this approach introduces several limitations, including under-representation of the central structure of the data if a tail component is given too much influence, ignoring context-specific relevance by treating all observations as equally informative and overall model over-sensitivity to outliers.

Suppose that, in the pdf in Equation (1), the transition from one component to another is smooth; then, the following assumptions are made:

(i) The pdf f is positive and satisfies

$$\int_{\mathbb{R}} f(x;\Theta)dx = 1,$$

implying that

$$w_1 F_1(u;\Theta_1) + w_2[1 - F_2(u;\Theta_2)] = 1.$$

(ii) The distribution of the data has a heavy right tail belonging to the Fréchet maximum domain of attraction.

(iii) The pdf f is continuous and differentiable at the threshold u; in addition, it is smooth and $\mathcal{C}^1$-regular, implying that

$$w_1 f_1(u;\Theta_1) = w_2 f_2(u;\Theta_2),$$
$$w_1 f_1^{'}(u;\Theta_1) = w_2 f_2^{'}(u;\Theta_2).$$

Given these assumptions, we obtain

$$\begin{cases} w_1 = w_2 \frac{f_2(u;\Theta_2)}{f_1(u;\Theta_1)}; \\ w_2 = \left\{ \frac{F_1(u;\Theta_1) f_2(u;\Theta_2)}{f_1(u;\Theta_1)} + 1 - F_2(u;\Theta_2) \right\}^{-1}. \end{cases} \tag{2}$$

The pdf in Equation (1) has cdf and corresponding quantile function given by the equations below:

$$F(x;\Theta) = \begin{cases} w_1 F_1(x;\Theta_1), & -\infty < x \leq u, \\ 1 - w_2[1 - F_2(x;\Theta_2)], & u \leq x < \infty, \end{cases} \tag{3}$$

$$Q(p;\Theta) = \begin{cases} Q_1\left(\frac{p}{w_1};\Theta_1\right), & \text{if} \quad p \leq w_1, \\ Q_2\left(\frac{p-(1-w_2)}{w_2};\Theta_2\right), & \text{if} \quad p \geq 1 - w_2. \end{cases} \tag{4}$$

Remark 1. *One can simulate random samples from the model in Equation (1) using Equation (4) by simply replacing p with the random variable U, where U is uniform on $(0,1)$.*

To define a two-component half-normal–GPD hybrid model, f_1 is taken to be the half-normal distribution with pdf, cdf, and quantile function expressed by the following respective equations:

$$f_1(x;\sigma) = \frac{\sqrt{2}}{\sigma\sqrt{\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \quad x \in \{0\} \cup \mathbb{R}_+ \quad \sigma \in \mathbb{R}_+,$$
$$F_1(x;\sigma) = \operatorname{erf}\left(\frac{x}{\sigma\sqrt{2}}\right) \quad x \in \{0\} \cup \mathbb{R}_+ \quad \sigma \in \mathbb{R}_+,$$
$$Q_1(p;\sigma) = \sigma\sqrt{2}\operatorname{erf}^{-1}(p) \quad \sigma \in \mathbb{R}_+ \quad 0 < p < 1,$$

where σ is a scale parameter, $\operatorname{erf}(\cdot)$ is the error function and $\operatorname{erf}^{-1}(\cdot)$ is its inverse. Furthermore, take f_2 as the GPD with pdf, cdf, and quantile function expressed by the following respective equations :

$$f_2(x-u;\beta,\gamma) = \frac{1}{\beta}\left(1+\gamma\frac{x-u}{\beta}\right)^{-1-\frac{1}{\gamma}}, \quad \beta\in\mathbb{R}_+ \quad \gamma\in\mathbb{R},$$

$$F_2(x-u;\beta,\gamma) = 1-\left(1+\gamma\frac{x-u}{\beta}\right)^{-\frac{1}{\gamma}}, \quad \beta\in\mathbb{R}_+ \quad \gamma\in\mathbb{R},$$

$$Q_2(p;u,\beta,\gamma) = \frac{\beta}{\gamma}\left[(1-p)^{-\gamma}-1\right]+u, \quad \beta\in\mathbb{R}_+ \quad \gamma\in\mathbb{R} \quad 0<p<1,$$

$$\forall x\geq u\in Z(\beta,\gamma), \quad Z(\beta,\gamma) = \begin{cases}[0,\infty) & \text{if} \quad \gamma\geq 0\\ [0,-\beta/\gamma) & \text{if} \quad \gamma<0,\end{cases}$$

where β is a scale parameter and γ is the tail index parameter which controls the shape of the GPD. Using assumptions (i)–(iii), we obtain the following relations for some of the parameters of the distribution:

$$\begin{cases} w_1 = \frac{w_2}{\beta f_1(u;\sigma)};\\ w_2 = \left\{1+\frac{F_1(u;\sigma)}{\beta f_1(u;\sigma)}\right\}^{-1};\\ \beta = -(1+\gamma)\frac{f_1(u;\sigma)}{f_1'(u;\sigma)}. \end{cases} \tag{5}$$

It follows that the vector of parameter Θ will contain only the free parameters including the threshold u. Thus, $\Theta = [\sigma, u, \gamma]$. These would be the parameters to be estimated using the proposed UIA. Once Θ has been estimated, the estimates of the parameters w_1, w_2, and β can easily be realized from Equation (5).

Remark 2. *Observe that in the half-normal–GPD hybrid model, there were six parameters whose values we needed to estimate. However, with assumptions (i)–(iii), the number of free parameters to be estimated was reduced to three. The graphs of the density and hazard functions for selected values of free parameters are given by Figure 1.*

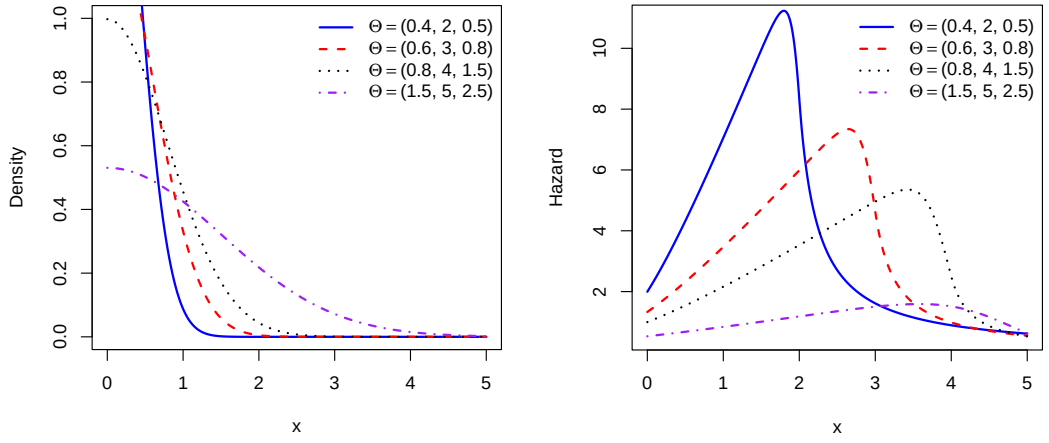


Figure 1. The half-normal–GPD densities for some selected values of free parameters (**left**) and the corresponding hazards (**right**). The densities are all skewed to the right, with heavy tails to the right. The hazards display reversed bathtub shapes.

Remark 3. *Given that $w_1 \neq w_2$, it follows that the threshold u can be any quantile of the half-normal–GPD model, and thus, the distribution is free of any constraint which is usually imposed for the case when $w_1 = w_2$ [3,25]. Also, since the half-normal–GPD model is positively skewed, the mode of f is equal to zero and less than the median M. This implies that $0 \leq \frac{w_1}{2} \leq F(M;\Theta) = \frac{1}{2}$,*

and consequently, $0 \leq w_1 \leq 1$. *Furthermore,* $0 \leq w_2 \leq 1$ *and* $1 - w_1 \leq w_2$. *Lastly,* w_2 *can be interpreted as the probability of exceeding the threshold* u, *while* w_1 *is a normalization parameter ensuring that the density* f *integrates to unity.*

3. Unsupervised Iterative Estimation Algorithm

In this section, a description of the UIA employed in estimating the vector of free parameter Θ is presented. The model described in Section 2 is taken to belong to the Fréchet maximum domain of attraction (i.e., $\gamma > 0$). For each iteration in the UIA, the UIA breaks down the problem of estimating the vector of free parameters Θ into a double-nested sub-problem, namely the parameter $\theta = [\sigma, u]$ and γ. These are estimated successively. For each iteration of the UIA, the parameter vector θ is first estimated by minimizing the Squared Distance (SD) between the empirical cdf based on some sample and the theoretical one based on the value of γ, and then, θ is replaced by its estimate obtained in this iteration to estimate γ in the next step of the algorithm using a similar procedure. Furthermore, θ is estimated again by minimizing the SD between the empirical cdf based on some sample and the theoretical one based on the estimate of γ obtained from the previous iteration, and then, θ is replaced by its estimate obtained in this iteration to estimate γ in the next step of the algorithm using a similar procedure. Evidently, the algorithm begins with an initial value for γ obtained by minimizing the SD between the empirical cdf based on some sample and the theoretical one based on the initial value chosen for θ and only ends when a stop condition is realized.

Consider a sample $X = (X_j)_{j\in\{1,2,\ldots,m\}}$ from the half-normal–GPD model, and let $\boldsymbol{x} = (x_i)_{i\in\{1,2,\ldots,n\}}$ be a given realization. Suppose $\widetilde{\alpha}^{(0)}$ and $\widetilde{\alpha}^{(k)}$ are the initial value and the estimate of the parameter α at the k^{th} iteration, respectively. To proceed with the UIA for the two-component hybrid model in 1, we start with the initial value $\widetilde{\theta}^{(0)} = \left[\widetilde{\sigma}^{(0)}, \widetilde{u}^{(0)}\right]$, rather than beginning by specifying an initial value for γ, because the only information we have about γ is that it is positive, given that we are dealing with positive data with positive skewness. Moreover, $\widetilde{\sigma}^{(0)} = q_{z\%}$ is chosen, where $q_{z\%}$ represents an empirical quantile of order $z\%$ associated with F. Also, $\widetilde{u}^{(0)} = q_{\rho\%}$ is chosen (as we fit a GPD above u). Moreover, $z < \rho$ since we naturally expect $\sigma < u$, the initial value of $\widetilde{\theta}^{(0)}$ is used to estimate $\widetilde{\gamma}^{(0)}$ by minimizing the SD between the hybrid cdf given $\theta = \widetilde{\theta}^{(0)}$ (fixed) and the empirical cdf F_n associated with the sample $X = (X_i)_{i\in\{1,2,\ldots,n\}}$ of size n, defined, for all $t \in \mathbb{R}$, by $F_n(t) = \sum_{i=1}^{n} \mathbb{1}_{(X_i \leq t)}/n$. Moreover, the SD is not evaluated only on the realizations x_i (because there might turn out to be just a few realizations in the tail), but also on a sequence of generated synthetic data with an increasing property $\boldsymbol{y} = (y_j)_{j\in\{1,2,\ldots,m\}}$ of size m (which may not be the same as n), with a logarithmic step. The synthetic data is added to increase the number of realizations above the tail threshold u. In particular, for any $i \in \{1, 2, \ldots, n\}$, y_j is defined by

$$y_j = \min_{i\in\{1,2,\ldots,n\}}(x_i) + \left(\max_{i\in\{1,2,\ldots,n\}}(x_i) - \min_{i\in\{1,2,\ldots,n\}}(x_i)\right)\log_{10}\left(1 + \frac{9(j-1)}{m-1}\right). \tag{6}$$

Remark 4. *The introduction of new points between the observations of* X *only has an impact on* F *by aiding its evaluation on more points, with no impact on the step function* F_n.

To obtain $\widetilde{\gamma}^{(0)}$, we solve the following minimization problem using the L-M algorithm [26,27]:

$$\widetilde{\gamma}^{(0)} \leftarrow \underset{\gamma>0}{argmin}\left\|F(\boldsymbol{y}; \Theta \mid \widetilde{\theta}^{(0)}) - F_n(\boldsymbol{y})\right\|_2^2,$$

where $\Theta \mid \widetilde{\theta}^{(0)}$ stands for Θ for $\theta = \widetilde{\theta}^{(0)}$, and $\|.\|_2$ denotes the L_2-norm. After realizing $\widetilde{\gamma}^{(0)}$, we proceed with the iterations. Now, $\forall k \geq 1$, the k^{th} iteration, is divided into two separate minimization problems, which are resolved successively, as described in the following steps.

***Step* 1:** Determine $\widetilde{\theta}^{(k)} = \left[\widetilde{\sigma}^{(k)}, \widetilde{u}^{(k)}\right]$ by minimizing the SD between the hybrid cdf, given $\widetilde{\gamma}^{(k-1)}$, and the empirical one as follows:

$$\widetilde{\theta}^{(k)} \leftarrow \underset{(\sigma,u)\in\mathbb{R}_+\times\mathbb{R}_+}{argmin} \left\| F(\boldsymbol{y};\, \Theta \mid \widetilde{\gamma}^{(k-1)}) - F_n(\boldsymbol{y}) \right\|_2^2$$

where $\Theta \mid \widetilde{\gamma}^{(k-1)}$ denotes Θ for $\gamma = \widetilde{\gamma}^{(k-1)}$ (fixed). The L-M algorithm is employed to numerically solve this minimization problem.

***Step* 2:** Determine $\widetilde{\gamma}^{(k)}$ by minimizing the SD between the hybrid cdf, given $\widetilde{\theta}^{(k)}$, and the empirical one by solving the following minimization problem using the L-M algorithm:

$$\widetilde{\gamma}^{(k)} \leftarrow \underset{\gamma>0}{argmin} \left\| F(\boldsymbol{y};\, \Theta \mid \widetilde{\theta}^{(k)}) - F_n(\boldsymbol{y}) \right\|_2^2,$$

where $\Theta \mid \widetilde{\theta}^{(k)}$ represents Θ for $\theta = \widetilde{\theta}^{(k)}$ (fixed).

Stop condition: The iterations continue till the following stop conditions are satisfied:

$$\left(\underbrace{d\Big(F(\boldsymbol{y};\Theta^{(k)}), F_n(\boldsymbol{y})\Big) < \varepsilon}_{\text{Condition 1}} \quad \text{and} \quad \underbrace{d\Big(F(\boldsymbol{y}_{q_\delta};\Theta^{(k)}), F_n(\boldsymbol{y}_{q_\delta})\Big) < \varepsilon}_{\text{Condition 2}} \right) \quad \text{or} \quad \underbrace{k = k_{max}}_{\text{Condition 3}}$$

where ε is a small positive real number; $\boldsymbol{y}_{q_\delta}$ stands for the observations above a fixed quantile q_δ of a given order δ, which is associated with the cdf F; and $d(x, y)$ denotes the distance between x and y. The distance $d(x, y)$ is chosen in this paper to be the Mean Squared Error (MSE), and it can be further interpreted as the Cramér-von-Mises test of goodness of fit.

To guarantee a good fit for all of the data points and not just the data points lying in the area where the bulk of the distribution lies, but also for the tail, the UIA is forced to stop only when the MSE between the hybrid cdf and the empirical one is small enough, and this implies satisfaction of Conditions 1 and 2; otherwise, satisfaction is implied when a fixed maximum number of iterations k_{max} is attained (Condition 3).

Remark 5. *While the hybrid model considered in this paper is assumed to belong to the Fréchet maximum domain of attraction, this algorithm can be extended to the case when the tail index of the GPD is free of any constraints. Also, though the maximum likelihood estimation method appears as a natural estimation method for the parameters of the hybrid model, in practice, this can be very challenging to execute, especially when the number of parameters to be estimated is high. Estimating many free parameters at once can be very challenging, and the challenge starts when selecting initial values for these parameters, as well as obtaining the expression of the gradient function from the likelihood or log-likelihood function. This has informed the choice of the UIA as an alternative method to estimate the parameters of the hybrid model. Nevertheless, when the number of free parameters is relatively small, the maximum likelihood estimation method can be used and is equally robust.*

Pseudo-code of the algorithm is given by Algorithm 1.

To study the convergence of the algorithm, it is necessary to provide proof of the existence of stationary points. Moreover, the algorithm, which consists of a sequence of minimizations, does not rely on the optimization of a cost function by seeking a path to

reach an extremum of an error surface. For a given realization $\boldsymbol{x} = (x_i)_{i\in\{1,2,\ldots,n\}}$ and given parameters $\zeta, \omega \in \{\theta, \gamma\}$ with $\zeta \neq \omega$, we define Ω_ζ and Ω_ω as the respective domain of ζ and ω. Now, consider the function

$$\begin{array}{rcl} \vartheta_\zeta : \Omega_\omega & \to & \Omega_\zeta \\ \omega & \mapsto & \vartheta_\zeta(\omega;\boldsymbol{x}) = \underset{\zeta\in\Omega_\zeta}{argmin}\ \mathcal{L}_\omega(\zeta;\boldsymbol{x}), \end{array}$$

where for $\Theta \mid \omega$, $\mathcal{L}_\omega$ is defined by

$$\begin{array}{rcl} \mathcal{L}_\omega : \Omega_\zeta & \to & \mathbb{R} \\ \zeta & \mapsto & \mathcal{L}_\omega(\zeta;\boldsymbol{x}) = \sum_{i=1}^{n} \big(F(x_i;\Theta \mid \omega) - F_n(x_i)\big)^2. \end{array}$$

To verify that ϑ_ζ is a mapping, it is sufficient to show that $\mathcal{L}_\omega$ admits a minimum which is unique, for any $\omega \in \{\theta, \gamma\}$, with $\theta = [\sigma, u] \in \mathbb{R}_+ \times \mathbb{R}_+$ and $\gamma > 0$. To do so, ω is fixed at $\omega = \theta$ (the result would be the same for $\omega = \gamma$), and we show that $\mathcal{L}_\theta(\gamma;\boldsymbol{x})$ is a strongly quasiconvex function with respect to γ, which is deduced from the strict convexity with respect to γ of the function $F(.;\Theta \mid \theta)$, denoted by $F(.;\gamma)$ to simply express it. Thus, we can write $\forall \gamma_1, \gamma_2 \in \Omega_\gamma$, with $\gamma_1 < \gamma_2$, $0 \leq \nu \leq 1$, and $\forall i \in \{1, \cdots, n\}$: $F(x_i; \nu\gamma_1 + (1-\nu)\gamma_2) < F(x_i;\gamma_2)$, which implies

$$\begin{aligned} \mathcal{L}_\theta(\nu\gamma_1 + (1-\nu)\gamma_2;\boldsymbol{x}) = & \sum_{i=1}^{n} \big(F(x_i;\gamma_1 + (1-\nu)\gamma_2) - F_n(x_i)\big)^2 < \\ & \sum_{i=1}^{n} \big(F(x_i;\gamma_2) - F_n(x_i)\big)^2 = \mathcal{L}_\theta(\gamma_2;\boldsymbol{x}). \end{aligned}$$

It follows that $\forall \gamma_1, \gamma_2 \in \Omega_\gamma$, $\gamma_1 < \gamma_2$, and $0 \leq \nu \leq 1$, $\mathcal{L}_\theta(\nu\gamma_1 + (1-\nu)\gamma_2;\boldsymbol{x}) < \max(\mathcal{L}_\theta(\gamma_1;\boldsymbol{x}), \mathcal{L}_\theta(\gamma_2;\boldsymbol{x}))$. Thus, $\mathcal{L}_\theta(\gamma;\boldsymbol{x})$ is strongly quasiconvex on $\mathcal{L}_\gamma$, a compact of $\mathbb{R}$, which ensures the presence of a unique minimum. The strict convexity with respect to γ of F further shows that the second derivative of F with respect to γ is positive.

Algorithm 1 UIA for the hybrid half-normal–GPD parameter estimation

1: Initialization: Decide start values for $\widetilde{\theta}^{(0)}$ and δ as well as values for ε and k_{max}. Proceed to obtain $\widetilde{\gamma}^{(0)}$ from:

$$\widetilde{\gamma}^{(0)} \leftarrow \underset{\gamma>0}{argmin} \left\| F(\boldsymbol{y};\Theta \mid \widetilde{\theta}^{(0)}) - F_n(\boldsymbol{y}) \right\|_2^2.$$

2: Iterative process: For

- $k \leftarrow 1$
  Step 1: Estimate $\widetilde{\theta}^{(k)}$ from:

$$\widetilde{\theta}^{(k)} \leftarrow \underset{(\sigma,u)\in\mathbb{R}_+\times\mathbb{R}_+}{argmin} \left\| F(\boldsymbol{y};\Theta \mid \widetilde{\gamma}^{(k-1)}) - F_n(\boldsymbol{y}) \right\|_2^2.$$

  Step 2: Estimate $\widetilde{\gamma}^{(k)}$ from:

$$\widetilde{\gamma}^{(k)} \leftarrow \underset{\gamma>0}{argmin} \left\| F(\boldsymbol{y};\Theta \mid \widetilde{\theta}^{(k)}) - F_n(\boldsymbol{y}) \right\|_2^2.$$

- $k \leftarrow k+1$
  until $\Big(d(F(\boldsymbol{y};\Theta^{(k)}), F_n(\boldsymbol{y})) < \varepsilon \text{ and } d(F(\boldsymbol{y}_{q_\delta};\Theta^{(k)}), F_n(\boldsymbol{y}_{q_\delta})) < \varepsilon\Big)$ or $\big(k = k_{max}\big)$.

3: Output: Return $\Theta^{(k)} = \big[\widetilde{\sigma}^{(k)}, \widetilde{u}^{(k)}, \widetilde{\gamma}^{(k)}\big]$.

The two steps of the first iteration of Algorithm 1 can be expressed using $\vartheta_\gamma : \mathcal{L}_\theta \to \mathcal{L}_\gamma$ and $\vartheta_\theta : \mathcal{L}_\gamma \to \mathcal{L}_\theta$, for a fixed $\widetilde{\gamma}^{(0)}$ via the following relations:

$$\begin{cases} \widetilde{\theta}^{(1)} = \vartheta_\theta(\widetilde{\gamma}^{(0)}; \boldsymbol{x}), \\ \widetilde{\gamma}^{(1)} = \vartheta_\gamma(\widetilde{\theta}^{(1)}; \boldsymbol{x}) = \vartheta_\gamma(\vartheta_\theta(\widetilde{\gamma}^{(0)}; \boldsymbol{x}); \boldsymbol{x}). \end{cases}$$

In general, for any $k \geq 1$, one can write

$$\widetilde{\gamma}^{(k)} = \iota(\widetilde{\gamma}^{(k-1)}; \boldsymbol{x}), \tag{7}$$

where the function $\iota(\cdot)$ is defined from Ω_γ to Ω_γ by: $\iota(\gamma; \boldsymbol{x}) = \vartheta_\gamma(\vartheta_\theta(\gamma; \boldsymbol{x}); \boldsymbol{x})$. As a consequence, Algorithm 1 can be expressed alternatively as Algorithm 2.

Algorithm 2 Alternative UIA for the hybrid half-normal–GPD parameter estimation

1: Initialization: Decide start values for $\widetilde{\theta}^{(0)}$ and δ as well as values for ε and k_{max} and proceed to obtain $\widetilde{\gamma}^{(0)}$
2: Iterative process: $k \leftarrow 1$

$$\widetilde{\gamma}^{(k)} = \iota(\widetilde{\gamma}^{(k-1)}; \boldsymbol{x})$$

$k \leftarrow k + 1$
until $(|\widetilde{\gamma}^{(k)} - \widetilde{\gamma}^{(k-1)}| < \varepsilon)$ or $(k = k_{max})$.
3: Output: Return $\widetilde{\gamma}^{(k)}$.

To show that stationary points exist for Algorithm 2, it is necessary to prove the existence of fixed points of the function ι. To realize this, we proceed from the fixed-point theorem due to Brouwer (see [28]), which means that *every continuous function from a closed ball of a Euclidean space into itself has a fixed point*. This implies that the functional ι admits at least one fixed point given the satisfaction of the following two conditions:

Condition 1: Ω_γ is a closed ball of a Euclidean space.

Condition 2: ι is continuous on Ω_γ.

Condition 1 is readily satisfied because, for a given realization $\boldsymbol{x}$, $\Omega_\gamma = [0, k_0], k_0 < \infty$, is a closed ball of $\mathbb{R}$ that is a Euclidean space. **Condition 1** can be established by proving that ϑ_γ and ϑ_θ are both continuous on their domains (recall that ι is the composite function: $\iota = \vartheta_\gamma \circ \vartheta_\theta$) using the Heine–Cantor theorem and Ramsay et al.'s one (see [29]). Using Ramsay et al.'s theorem, it is required to verify that Ω_θ is a compact set and that $\mathcal{L}_\gamma$ is uniformly continuous on Ω_θ, in order to conclude the continuity of ϑ_θ. The first condition, Ω_θ, is a compact set of $\mathbb{R}^2$, which is satisfied, given that we are working with a half-normal density, with finite mean and variance; hence, it is bounded.

Now, as Ω_θ is a compact set, it is sufficient to show that $\mathcal{L}_\gamma$ is continuous on Ω_θ to obtain, by the Heine–Cantor theorem, its uniform continuity. For this, it is necessary to study the continuity of F with respect to θ to determine the continuity of $\mathcal{L}_\gamma$ with respect to θ, since $\mathcal{L}_\gamma(\theta; \boldsymbol{x}) = \sum_{i=1}^{n} \left(F(x_i; \theta \mid \gamma) - F_n(x_i)\right)^2$. Recall that, by construction, F is continuous with respect to x and not to its parameters. Hence, its continuity according to θ needs to be established. Since we know that, from Equation (3), F is composed of two functions, the half-normal cdf and that of the GPD, it suffices to study the continuity of each of them with respect to θ. The continuity of the half-normal cdf F as a function of θ is obvious since it means to look at the continuity of its likelihood with respect to $\theta = [\sigma, u]$. Now, for the GPD, its parameter β is expressed as a function of θ: $\beta(\theta) = -(1+\gamma)\dfrac{f_1(\theta)}{f_1'(\theta)}$. The GPD parameter γ can equally be expressed as a function of θ: $\gamma(\theta) = \beta(\theta)\dfrac{f_1'(\theta)}{f_1(\theta)} + 1$, and hence, both are continuous in θ. It follows that the GPD is continuous in θ as the composition of continuous functions with respect to θ. The continuity of the function $\mathcal{L}_\gamma$ on Ω_θ as a composition, the sum and products of continuous functions on Ω_θ can be inferred, and from which, one can conclude the continuity of ϑ_θ on Ω_γ.

In sum, since a fixed point exists for the functional ι, the algorithm admits at least one stationary point. Moreover, since the method does not follow a path on an error surface, it is free from local minima traps, as found in standard gradient search-based methods. It can also be verified numerically, as performed in the next section, that the algorithm converges to a unique stationary point regardless of its initialization.

4. Numerical Studies

To study the performance of the UIA described in Section 3, Monte Carlo simulations are carried out. Through simulations, an attempt is made to test the efficiency of the UIA as applied in estimating the parameters of the hybrid half-normal–GPD model.

We proceed with the simulations as follows: We consider N (=100 in this paper) training sets $\{\boldsymbol{x^q} = (x_p^q)_{p\in\{1,2,\ldots,n\}}\}_{q\in\{1,2,\ldots,N\}}$ of size n and N test sets $\{\boldsymbol{y^q} = (y_p^q)_{p\in\{1,2,\ldots,l\}}\}_{q\in\{1,2,\ldots,N\}}$ of size l, simulated from the hybrid half-normal–GPD model with a fixed parameters vector Θ. Using each training set $\boldsymbol{x^q}$, $q \in \{1,2,\ldots,N\}$, Θ is estimated to be $\widetilde{\Theta}^q = [\widetilde{\sigma}^q, \widetilde{u}^q, \widetilde{\gamma}^q]$, using the UIA. We denote by $\widetilde{\alpha}^q$ the estimate of the parameter α relative to the q^{th} training set. Furthermore, the empirical mean and variance of $\widetilde{\alpha}^q$ over the N training sets is computed, namely $\tilde{\alpha} = \sum_{q=1}^{N} \widetilde{\alpha}^q/N$ and $\tilde{S}_N^{\alpha} = \sum_{q=1}^{N} (\widetilde{\alpha}^q - \widetilde{\alpha})^2/(N-1)$, respectively. The significance of $\tilde{\alpha}$ is determined based on two criteria: the MSE and test of hypothesis on α. The MSE is expressed for any parameter α as $\mathrm{MSE}_{\alpha} = \sum_{q=1}^{N} (\widetilde{\alpha}^q - \alpha)^2/N$. A small MSE value indicates the efficiency of the UIA in estimating the parameter α. To test $\tilde{\alpha}$ (with unknown variance), we set up the hypothesis

$$\begin{aligned} H_0 &: \tilde{\alpha} = \alpha \\ H_1 &: \tilde{\alpha} \neq \alpha. \end{aligned}$$

Because N is large, a z-test (instead of a t-test) of size κ, with a rejection region of H_0 at risk κ% described by $\left(|T_{\tilde{\alpha},N}| > \Phi^{-1}(1-\kappa/2)\right)$ is used, where the statistics $T_{\tilde{\alpha},N}$ are given by $T_{\tilde{\alpha},N} = (\tilde{\alpha} - \alpha)/\sqrt{\tilde{S}_N^{\alpha}}$, and $\Phi^{-1}(1-\kappa/2)$ denotes the quantile of order $1-\kappa/2$ of the standard normal distribution Φ. Additionally, we compute the p-value $p_{T_{\tilde{\alpha},N}} = 2(1-\Phi(|T_{\tilde{\alpha},N}|))$, with respect to the parameter α, which we will compare to κ. Whenever this p-value is higher than κ, we fail to reject H_0. For any n and any parameter $\alpha \in \{\sigma, u, \gamma\}$, we obtain $|T_{\tilde{\alpha},N}| < \Phi^{-1}(0.975) = 1.96$ and $p_{T_{\tilde{\alpha},N}} > \kappa = 5\%$ to confirm high acceptance (at the 5% level of significance) of H_0 ($\tilde{\alpha} = \alpha$), that is, a very high level of similarity between the values obtained through the UIA and the fixed ones. Lastly, the hybrid half-normal GPD pdf f, given Θ, is compared with the pdf $\widetilde{f}$ estimated on each test set $\boldsymbol{y^q}$, given $\widetilde{\Theta}^q$. For this, the average of the log-likelihood ratio $\mathcal{D}$ of $f(\boldsymbol{y^q};\Theta)$ by $\widetilde{f}(\boldsymbol{y^q};\widetilde{\Theta}^q)$ over the N simulations is computed:

$$\mathcal{D} = \frac{1}{Nl} \sum_{q=1}^{N} \sum_{p=1}^{l} \log\left(\frac{f(y_p^q;\Theta)}{\widetilde{f}(y_p^q;\widetilde{\Theta}^q)}\right). \tag{8}$$

A small value of $\mathcal{D}$ indicates an efficient estimation of the parameters of the hybrid half-normal–GPD model using the UIA.

We performed many Monte Carlo simulations by varying Θ and n in order to ascertain the robustness of the UIA for different values of the parameters and sample sizes. We also set $l = n$, $z = 20$, $\kappa = 5\%$, $\delta = 0.7$, $\varepsilon = 10^{-8}$, and $\rho = 40$. To conserve space, the results of three such simulations are reported in Tables 1–3, and the other unreported simulations follow a fashion

similar to the ones reported here. The efficiency of the UIA, in terms of goodness of fit, is shown through the two criteria described above and the average of the log-likelihood ratio $\mathcal{D}$.

We observe from the results in Tables 1–3 that, as the sample size increases, the MSE becomes smaller for all parameters. The variance of σ is also observed to be smaller than the variances of u and γ for all parameter combinations and sample sizes. We also observe that $\mathcal{D}$ is very small for all parameter combinations and sample sizes. This highlights the accuracy and efficiency of the estimation of the parameters using the UIA. Lastly, it can be observed that as the sample size increases, the average execution time of the UIA increases. It should be noted that the estimation algorithm was implemented using the R programming language. For faster programming languages, the average execution time of the estimation algorithm could be significantly reduced.

Table 1. Monte Carlo simulation results for $\Theta = [1, 3, 1.5]$.

		$n = 10^2$	$n = 10^3$	$n = 10^4$	$n = 10^5$
$\sigma = 1$	$\tilde{\sigma}$	0.9658	0.9898	0.9978	0.9999
	$\tilde{S}_N^{\sigma}$	$9.30\ 10^{-3}$	$1.02\ 10^{-3}$	$1.15\ 10^{-4}$	$1.56\ 10^{-5}$
	MSE_{σ}	$1.04\ 10^{-2}$	$1.11\ 10^{-3}$	$1.15\ 10^{-4}$	$1.54\ 10^{-5}$
	$T_{\tilde{\sigma},N}$	-0.3541	-0.3200	-0.2083	-0.0534
	$p_{T_{\tilde{\sigma},N}}$	0.7232	0.7490	0.8350	0.9574
$u = 3$	$\tilde{u}$	2.6249	3.3181	3.1098	3.0008
	$\tilde{S}_N^{u}$	5.3375	2.0292	$4.49\ 10^{-1}$	$1.52\ 10^{-2}$
	MSE_{u}	5.1201	2.1101	$4.49\ 10^{-1}$	$1.59\ 10^{-2}$
	$T_{\tilde{u},N}$	-0.6457	0.2233	0.1652	0.0061
	$p_{T_{\tilde{u},N}}$	0.5185	0.8233	0.8688	0.9951
$\gamma = 1.5$	$\tilde{\gamma}$	1.9583	1.7265	1.6312	1.4912
	$\tilde{S}_N^{\gamma}$	7.0064	$6.89\ 10^{-1}$	$3.12\ 10^{-2}$	$9.37\ 10^{-3}$
	MSE_{γ}	7.1464	$6.92\ 10^{-1}$	$3.26\ 10^{-2}$	$9.24\ 10^{-3}$
	$T_{\tilde{\gamma},N}$	0.1731	0.2728	0.1206	-0.0922
	$p_{T_{\tilde{\gamma},N}}$	0.8625	0.7850	0.8627	0.9265
Execution time (seconds)		82.59	237.50	1885.94	19,884.85
Maximum number of iterations		50	50	50	50
$\mathcal{D}$		$4.32\ 10^{-3}$	$1.62\ 10^{-3}$	$1.13\ 10^{-4}$	$1.12\ 10^{-6}$

Table 2. Monte Carlo simulation results for $\Theta = [2.5, 5, 1.5]$.

		$n = 10^2$	$n = 10^3$	$n = 10^4$	$n = 10^5$
$\sigma = 2.5$	$\tilde{\sigma}$	2.4569	2.4945	2.4981	2.4997
	$\tilde{S}_N^{\sigma}$	$1.57\ 10^{-1}$	$1.59\ 10^{-2}$	$1.35\ 10^{-3}$	$1.45\ 10^{-4}$
	MSE_{σ}	$1.57\ 10^{-1}$	$1.58\ 10^{-2}$	$1.32\ 10^{-3}$	$1.43\ 10^{-4}$
	$T_{\tilde{\sigma},N}$	-0.1089	-0.0435	-0.0517	-0.0250
	$p_{T_{\tilde{\sigma},N}}$	0.91327	0.9653	0.9586	0.9801
$u = 5$	$\tilde{u}$	5.7881	4.9973	4.9986	5.0002
	$\tilde{S}_N^{u}$	11.47158	$6.45\ 10^{-1}$	$9.71\ 10^{-2}$	$4.49\ 10^{-3}$
	MSE_{u}	11.978	$6.39\ 10^{-1}$	$10.01\ 10^{-2}$	$4.41\ 10^{-3}$
	$T_{\tilde{u},N}$	0.2327	-0.0034	-0.0045	0.0039
	$p_{T_{\tilde{u},N}}$	0.81601	0.9973	0.9964	0.9976
$\gamma = 1.5$	$\tilde{\gamma}$	2.4567	1.5536	1.5217	1.4985
	$\tilde{S}_N^{\gamma}$	6.9643	$3.60\ 10^{-1}$	$4.23\ 10^{-2}$	$2.24\ 10^{-3}$
	MSE_{γ}	7.8099	$3.59\ 10^{-1}$	$4.37\ 10^{-2}$	$2.22\ 10^{-3}$
	$T_{\tilde{\gamma},N}$	0.3625	0.0892	0.1055	-0.0317
	$p_{T_{\tilde{\gamma},N}}$	0.7167	0.9289	0.9160	0.9747
Execution time (seconds)		61.50	257.01	1925.2	20,184.8
Maximum number of iterations		50	50	50	50
$\mathcal{D}$		$8.61\ 10^{-3}$	$5.67\ 10^{-5}$	$1.26\ 10^{-5}$	$6.94\ 10^{-8}$

Table 3. Monte Carlo simulation results for $\Theta = [1.5, 3.5, 2]$.

		$n = 10^2$	$n = 10^3$	$n = 10^4$	$n = 10^5$
$\sigma = 1.5$	$\tilde{\sigma}$	1.4611	1.4980	1.4987	1.4991
	$\tilde{S}_N^{\sigma}$	$2.75\ 10^{-2}$	$5.34\ 10^{-3}$	$3.67\ 10^{-4}$	$8.85\ 10^{-5}$
	MSE_{σ}	$2.87\ 10^{-2}$	$5.29\ 10^{-3}$	$3.73\ 10^{-4}$	$9.07\ 10^{-5}$
	$T_{\tilde{\sigma},N}$	-0.2344	-0.0279	-0.0679	-0.0956
	$p_{T_{\tilde{\sigma},N}}$	0.8147	0.9778	0.9459	0.9238
$u = 3.5$	$\tilde{u}$	3.7692	3.5337	3.4790	3.4912
	$\tilde{S}_N^{u}$	2.5431	$4.65\ 10^{-1}$	$8.60\ 10^{-2}$	$4.38\ 10^{-3}$
	MSE_{u}	2.5902	$4.61\ 10^{-1}$	$9.01\ 10^{-2}$	$1.42\ 10^{-3}$
	$T_{\tilde{u},N}$	0.1688	0.0495	-0.0689	-0.1612
	$p_{T_{\tilde{u},N}}$	0.8659	0.9606	0.9451	0.8943
$\gamma = 2$	$\tilde{\gamma}$	3.7622	2.4352	1.9005	1.9601
	$\tilde{S}_N^{\gamma}$	24.776	3.1017	$2.19\ 10^{-1}$	$6.13\ 10^{-2}$
	MSE_{γ}	27.634	3.2601	$2.27\ 10^{-1}$	$9.42\ 10^{-2}$
	$T_{\tilde{\gamma},N}$	0.3540	0.2471	-0.2127	-0.1612
	$p_{T_{\tilde{\gamma},N}}$	0.7233	0.8048	0.8316	0.8719
Execution time (seconds)		68.64	217.13	1819.1	19,163.9
Maximum number of iterations		50	50	50	50
$\mathcal{D}$		$5.43\ 10^{-3}$	$1.03\ 10^{-3}$	$5.35\ 10^{-5}$	$7.85\ 10^{-6}$

5. Application

Here, the hybrid half-normal–GPD (HNGPD) model with parameter vector $\Theta = [\sigma, u.\gamma]$ alongside the hybrid lognormal–GPD (LNGPD), gamma–GPD (GGPD), and the normal GPD (NGPD) models [30,31] are used to model the intensity, in millimeters per hour, of rainfall which triggered some 785 debris flow events in the South Tyrol region of Italy between 1987 and 2022. The data set was provided by the Agency for Civil Protection of the Autonomous Province of Bozen-Bolzano, Italy. The skewness and excess kurtosis of the data are 1.8466 and 3.2528, respectively, which clearly shows that the data set is highly skewed to the right with a heavy tail. A violin plot of the data with a superimposed box plot is given in Figure 2.

The hybrid LNGPD, GGPD, and the NGPD models belong to the class of hybrid models with a cdf with the general form

$$F(x;\Theta) = \begin{cases} F_1(x;\Theta_1), & -\infty < x \leq u, \\ F_1(u;\Theta_1) + [1 - F_1(u;\Theta_1)]F_2(x;\Theta_2), & u \leq x < \infty. \end{cases} \tag{9}$$

For the LNGPD model, F_1 is the cdf of the lognormal distribution with parameter vector $\Theta_1 = [\mu, \sigma]$, where μ and σ are the logarithm of the mean and standard deviation, respectively, and F_2 is the cdf of the GPD distribution with parameter $\Theta_2 = [u, \beta, \gamma]$. Thus, the vector of all the parameters in the model is $\Theta = [\mu, \sigma, u, \beta, \gamma]$. For the GGPD model, F_1 is the cdf of the gamma distribution with parameter vector $\Theta_1 = [\alpha, \sigma]$, where α and σ are shape and scale parameters, respectively, and F_2 is the cdf of the GPD distribution with parameter $\Theta_2 = [u, \beta, \gamma]$. Again, the vector of all the parameters in the model is $\Theta = [\alpha, \sigma, u, \beta, \gamma]$. Lastly, for the NGPD model, F_1 is the cdf of the normal distribution with parameter vector $\Theta_1 = [\mu, \sigma]$, where μ and σ are the mean and standard deviation, respectively, and F_2 is the cdf of the GPD distribution with parameter $\Theta_2 = [u, \beta, \gamma]$, and the vector of all the parameters in the model is $\Theta = [\mu, \sigma, u, \beta, \gamma]$.

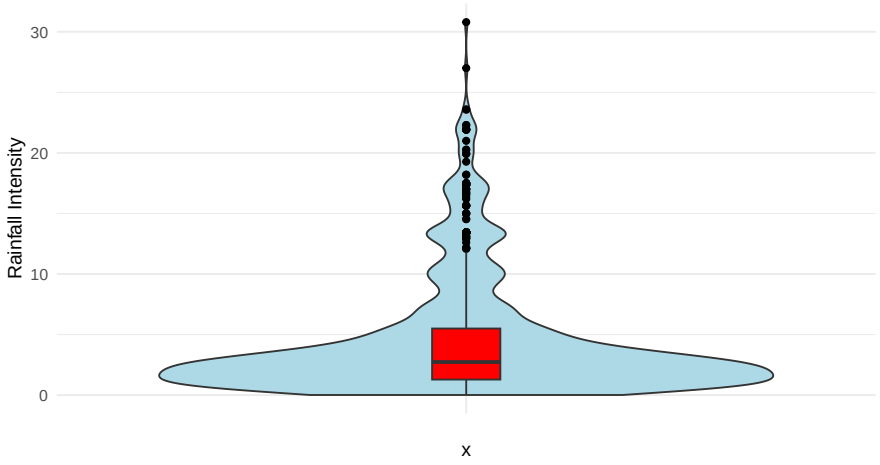


Figure 2. Intensity, in millimeters per hour, of rainfall which triggered some 785 debris flow events in South Tyrol, Italy.

Reported in Table 4 are the results obtained from using the HNGPD model, estimated using the UIA, in fitting the rainfall intensity data. Table 4 also contains results obtained from using the LNGPD, GGPD, and the NGPD models, all estimated using the maximum likelihood estimation (MLE) method, to fit the data. The results include estimates of the parameter vector Θ for all the models, the Akaike Information Criterion (AIC) for all the models, and the Kolmogorov–Smirnov (K-S) statistic, as well as its corresponding *p*-value, for all the models.

Table 4. Results of the fits of the HNGPD, LNGPD, GGPD, and NGPD models to the rainfall intensity data.

Model (Estimator)	Estimate of $\hat{\Theta}$	AIC	K-S	*p*-Value
HNGPD (UIA)	$\hat{\sigma} = 3.2855$ $\hat{u} = 4.4253$ $\hat{\gamma} = 0.6329$	3988.6	0.0388	0.1877
LNGPD (MLE)	$\hat{\mu} = 1.2167$ $\hat{\sigma} = 1.3291$ $\hat{u} = 2.6762$ $\hat{\beta} = 4.6878$ $\hat{\gamma} = 0.1318$	3994.1	0.1045	$7.3\ 10^{-8}$
GGPD (MLE)	$\hat{\alpha} = 1.3317$ $\hat{\sigma} = 2.6946$ $\hat{u} = 2.0525$ $\hat{\beta} = 3.9659$ $\hat{\gamma} = 0.1895$	3961.7	0.0563	$1.4\ 10^{-2}$
NGPD (MLE)	$\hat{\mu} = 2.3398$ $\hat{\sigma} = 1.2385$ $\hat{u} = 1.8156$ $\hat{\beta} = 3.3453$ $\hat{\gamma} = 0.2381$	4065.1	0.0646	$2.8\ 10^{-3}$

Figures 3–6 show the graph of the fitted density over the histogram of the data and Q-Q plot (on a log-scale) for the HNGPD, LNGPD, GGPD, and NGPD models, respectively. Results from Table 4 and Figures 3–6 reveal that the HNGPD offered the best fit to that data because it has the highest *p*-value for the K-S statistic.

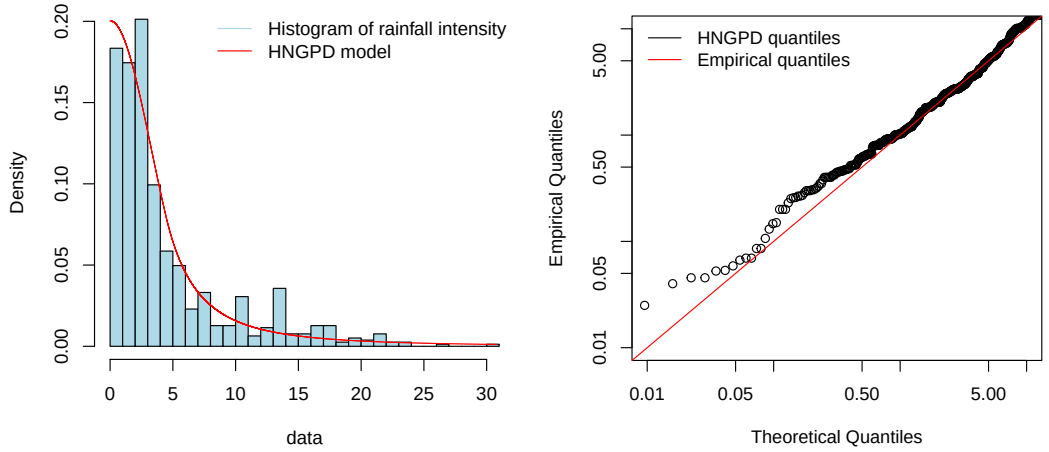


Figure 3. Rainfall intensity: Fitted HNGPD density over the histogram of the data (**left**) and a Q-Q plot showing the empirical quantiles against the theoretical ones of the HNGPD model (**right**).

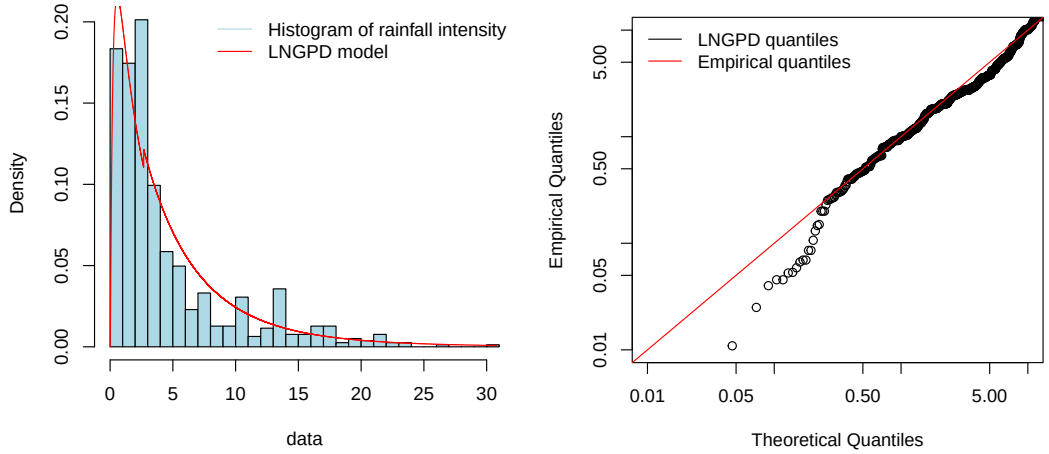


Figure 4. Rainfall intensity: Fitted LNGPD density over the histogram of the data (**left**) and a Q-Q plot showing the empirical quantiles against the theoretical ones of the LNGPD model (**right**).

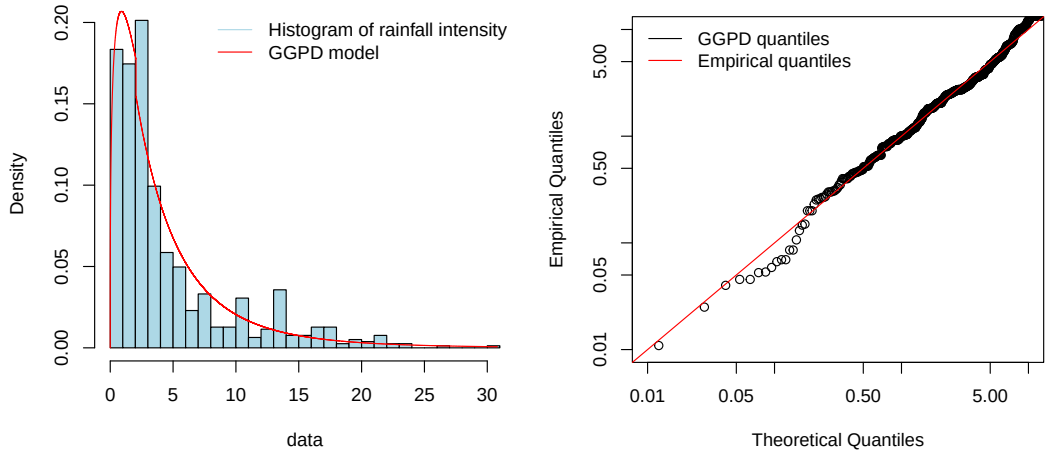


Figure 5. Rainfall intensity: Fitted GGPD density over the histogram of the data (**left**) and a Q-Q plot showing the empirical quantiles against the theoretical ones of the GGPD model (**right**).

We also compare the algorithmically estimated threshold $\hat{u}$ using the UIA with some graphical threshold selection approaches, namely the mean excess plot (MEP) [2], the Hill plot [32], and the stability plot [33]. These graphical threshold selection approaches are used to estimate u from the rainfall intensity data, and by using the method of probability-weighted moments (PWMs) [34], a GPD is fitted to the observations above the u selected

by each approach. Results regarding the estimated GPD parameters, the selected u, and the number of observations above each selected u are contained in Table 5. A comparison between the empirical cdf and the one estimated using the UIA for the HNGPD model is given by Figure 7. Also, A comparison between the empirical cdf and the one estimated using model parameters obtained from the MEP, Hill plot, and stability plot for the HNGPD model is given in Figures 8, 9 and 10, respectively. Lastly, a comparison between the empirical upper quantiles and the ones estimated using the UIA and the graphical approaches is given in Figure 11.

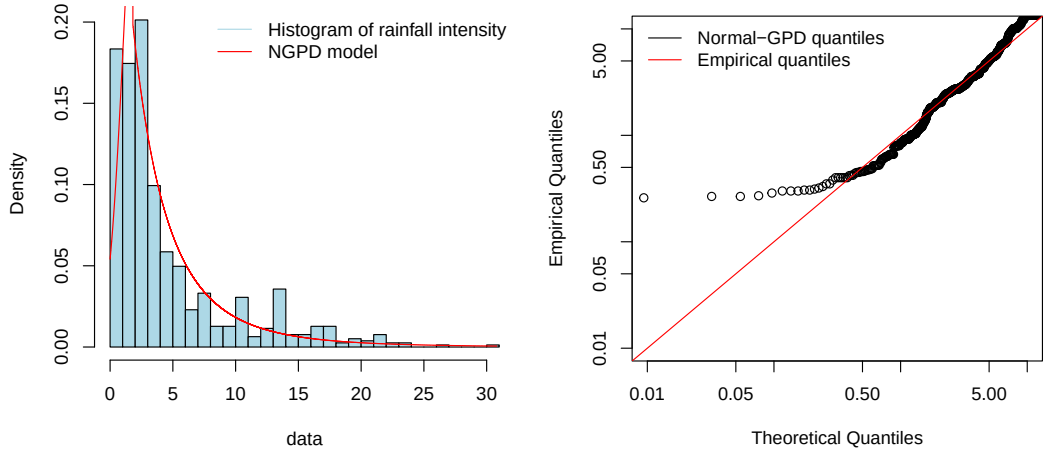


Figure 6. Rainfall intensity: Fitted NGPD density over the histogram of the data (**left**) and a Q-Q plot showing the empirical quantiles against the theoretical ones of the NGPD model (**right**).

Table 5. Comparison between the UIA approach and the MEP, Hill, and stability plot approaches for threshold selection of the upper tail of the rainfall intensity data, with N_u being the number of observations above the selected threshold $\hat{u}$. The parameters γ and β of the GPD are estimated from the exceedance observations using the PWM method, given the threshold $\hat{u}$ selected by the MEP, Hill, and stability plots.

Model	$\hat{\gamma}$	$\hat{\beta}$	$\hat{u}$	N_u
HNGPD (UIA)	0.6329	3.9830	$4.4253 = q_{68.79\%}$	245
GPD (MEP)	0.1141	4.7841	$2.9010 = q_{53.89\%}$	362
GPD (Hill)	0.1137	4.7822	$2.8609 = q_{53.50\%}$	365
GPD (Stability)	0.2151	3.9486	$2.7333 = q_{48.92\%}$	401

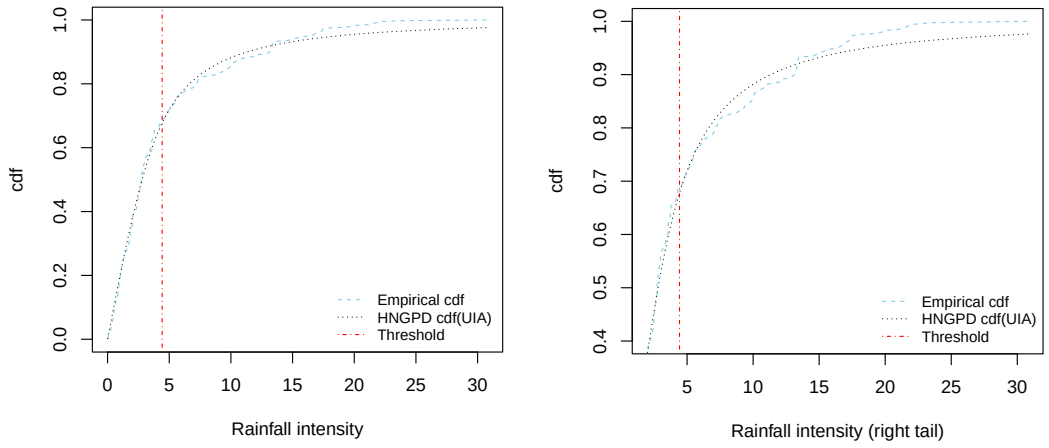


Figure 7. Rainfall intensity: Comparison between the empirical cdf and the one estimated using the UIA (**left**). The upper tail of the cdfs is shown in the graph (**right**).

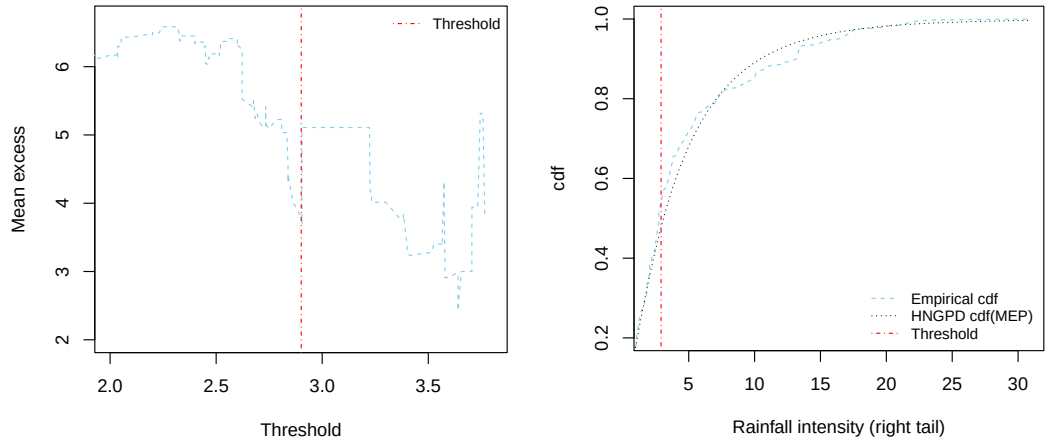


Figure 8. Rainfall intensity: MEP indicative of the threshold point (**left**) and a comparison between the empirical cdf and the one estimated using the threshold and tail index estimates obtained from the MEP (**right**).

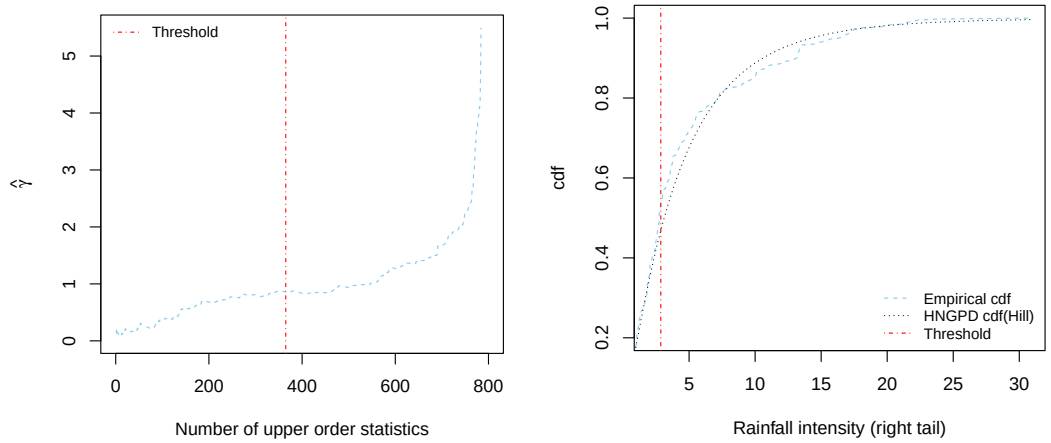


Figure 9. Rainfall intensity: Hill plot indicative of the threshold point (**left**) and a comparison between the empirical cdf and the one estimated using the threshold and tail index estimates obtained from the Hill plot (**right**).

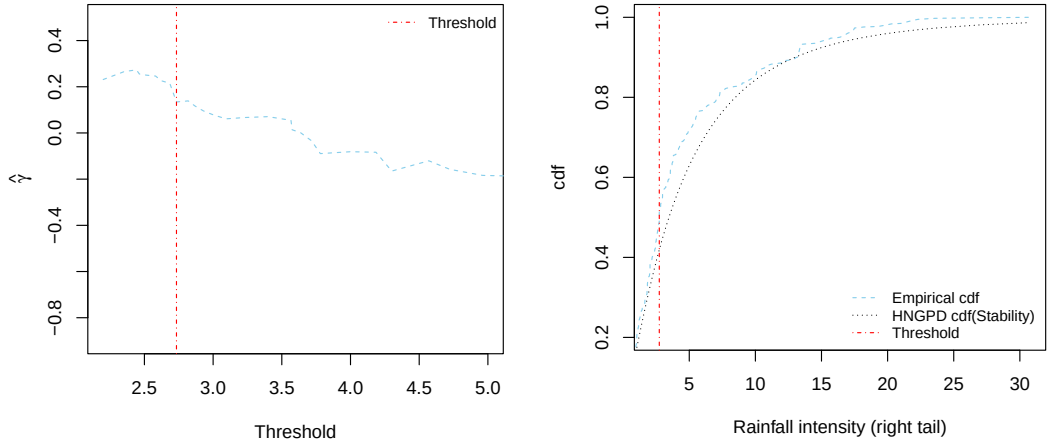


Figure 10. Rainfall intensity: Stability plot indicative of the threshold point (**left**) and a comparison between the empirical cdf and the one estimated using the threshold and tail index estimates obtained from the stability plot (**right**).

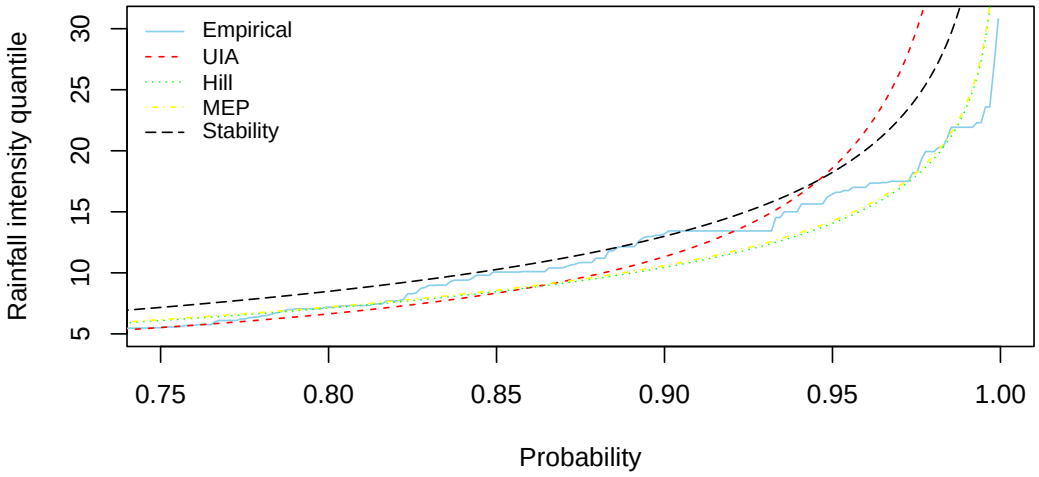


Figure 11. Rainfall intensity: Comparison between the empirical upper quantiles and the ones estimated using the UIA and the graphical approaches.

6. Discussion and Conclusions

In Section 5, it was established from the results derived from the application of the model to rainfall intensity data that the new HNGPD hybrid model, estimated using the unsupervised iterative estimation algorithm discussed in Section 3, offers the best fit to the data. Moreover, it can be observed that the LNGPD, GGPD, and the NGPD models all underestimated the value of the threshold u, beyond which a GPD is fitted to the tail of the distribution of the data. The implication of this is that all three models are suggesting the presence of more outliers or extreme observations in the data set than observations around the mean (bulk of the distribution). This can be very misleading from a practical point of view, where the bulk of the distribution is supposed to contain the majority of the observations. In particular, $u = 2.6762 = q_{47.89\%}$ for the LNGPD model, and thus, the threshold is on the 47.89th quantile. This further implies that a total of 410 observations out of the 785 observations are above the threshold and hence, are classified as extreme observations. For the GGPD model, $u = 2.0525 = q_{38.47\%}$ and thus, the threshold is on the 38.47th quantile, implying that a total of 483 observations out of the 785 observations are extreme observations. For the NGPD model, $u = 1.8156 = q_{32.10\%}$ and thus, the threshold is on the 32.10th quantile, implying that a total of 533 observations out of the 785 observations are extreme observations. For the new two-component hybrid HNGPD model estimated using the unsupervised iterative estimation algorithm, $u = 4.4253 = q_{68.79\%}$, and thus, the threshold is on the 68.79th quantile, implying that a total of 245 observations out of the 785 observations are extreme observations. This seems more reasonable given the nature of rainfall intensity data, where the bulk of observations of rainfall intensity are clustered around the mean, and a few other extreme rainfall intensity observations are also present in the data set. From the HNGPD model, we also obtained $w_2 = 0.3221$, which gives the probability of exceeding the threshold u. Thus, there exists a less than 0.5 probability of exceeding the threshold u, and this makes a lot of sense and justifies our use of non-uniform weights for each component of the model, in contrast to the use of uniform weights.

It can also be observed from Table 5 that the graphical approaches all selected a lower threshold value u, beyond which a GPD is fitted to the tail of the data distribution than the one estimated using the UIA. Again, the graphical approaches suggest the presence of many outliers or extreme observations in the data set. In particular, $u = 2.9010 = q_{53.89\%}$ for the MEP approach, and thus, the threshold is on the 53.89th quantile. This further implies that a total of 362 observations out of the 785 observations are above the threshold, and hence, they are classified as extreme observations. For the Hill plot approach, $u = 2.8609 = q_{53.50\%}$,

and thus, the threshold is on the 53.50th quantile, implying that a total of 365 observations out of the 785 observations are extreme observations. For the stability plot approach, $u = 2.7333 = q_{48.92\%}$, and thus, the threshold is on the 48.92th quantile, implying that a total of 401 observations out of the 785 observations are extreme observations. These threshold points selected by the graphical approaches are significantly lower than the ones selected by the UIA, which offered a more realistic threshold point given the nature of our data.

In conclusion, a new two-component hybrid model, suitable for modeling data with high right skewness and estimated using an unsupervised iterative estimation algorithm, has been introduced in this paper. A demonstration of the hybrid model's flexibility and robustness in capturing the unique characteristics of such data through application to a real data set has been carried out, and the results obtained have also been compared with those of other hybrid models in its class. Moreover, through empirical analyses on synthetic data sets, the unsupervised iterative estimation algorithm for the estimation of the parameters of the hybrid model has shown high accuracy and efficiency, making it a valuable tool for further practical applications. This new hybrid model and estimation technique hold significant promise for a wide range of fields where right-skewed data sets are prevalent, such as finance, environmental studies, signal processing, biomedicine, etc. Furthermore, given that the estimation algorithms also help in the automatic selection of a threshold beyond which extreme observations are observed in a distribution, it could become an effective tool in tail risk modeling. In finance, it could allow for more accurate estimation of Value-at-Risk (VaR) and Expected Shortfall (ES), ensuring that institutions can quantify potential catastrophic losses with greater reliability. In insurance, it could improve the modeling of extreme claims arising from natural disasters or systemic shocks, supporting fair premium setting and reinsurance strategies. Within the context of environmental studies, it could enhance the prediction of rare but severe climatic events, such as extreme rainfall or heatwaves, critical for climate resilience planning. Future research could extend this work by exploring multi-component extensions and applying the model and the estimation algorithm to more diverse data sets. Additionally, integrating this hybrid model into more complex statistical frameworks, such as regression models and machine learning algorithms, could further enhance its utility and scope of application. In general, our contribution in this paper provides a substantial advancement in the modeling of data exhibiting high skewness to the right, offering a powerful and versatile tool for statisticians, data scientists, and users of statistics.

Funding: This research was funded by the Free University of Bozen-Bolzano via the research project "EXTReme Events in Mountain Environments (EXTREME- RC 2022).

Data Availability Statement: The data set used in this article can be made readily available upon request from the author.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Blum, P.; Dacorogna, M. Extreme forex moves. *Risk-Lond.-Risk Mag. Ltd.* **2003**, *16*, 63–66.
2. Embrechts, P.; Kluppelberg, C.; Mikosch, T. *Modelling Extremal Events for Insurance and Finance*; Springer: Berlin/Heidelberg, Germany, 1997.
3. Carreau, J.; Bengio, Y. Hybrid Pareto model for asymmetric fat-tailed data: The univariate case. *Extremes* **2009**, *1*, 53–76. [CrossRef]
4. Dacorogna, M.; Kratz, M. Living in a stochastic world and managing complex risks. *SSRN* **2015**. [CrossRef]
5. Rangaswamy, M.; Michels, J.; Himed, B. Statistical analysis of the non-homogeneity detector for faSTAPi applications. *Digit. Signal Process.* **2004**, *14*, 253–267. [CrossRef]
6. Digham, F.; Alouini, M.S.; Simon, M.K. On the energy detection of unknown signals over fading channels. *IEEE Trans. Commun.* **2007**, *55*, 21–24. [CrossRef]

7. Broadwater, J.B.; Chellapa, R. Adaptive threshold estimation via extreme value theory. *IEEE Trans. Signal Process.* **2010**, *58*, 490–500. [CrossRef]
8. Mandava, A.; Shahram, L.; Regentova, E. Reliability assessment of microarray data using fuzzy classification methods: A comparative study. In *Advances in Computing and Communications*; Communications in Computer and Information Science; Springer: Berlin/Heidelberg, Germany, 2011; Volume 190, pp. 351–360.
9. Sermpezis, P.; Spyropoulos, T. Modelling and analysis of communication traffic heterogeneity in opportunistic networks. *IEEE Trans. Mob. Comput.* **2015**, *14*, 2316–2331. [CrossRef]
10. Rossi, F.; Fiorentino, M.; Versace, P. Two component extreme value distribution for flood frequency analysis. *Water Resour. Res.* **1984**, *20*, 847–856. [CrossRef]
11. Davison, A.C.; Smith, R.L. Models for exceedances over high thresholds. *J. R. Stat. Soc. Ser. (Methodol.)* **1990**, *52*, 393–442. [CrossRef]
12. Furrer, E.; Katz, R. Improving the simulation of extreme precipitation events by stochastic weather generators. *Water Resour. Res.* **2008**, *44*, W12439. [CrossRef]
13. Singh, V.; Li, C.; Mishra, A. Simulation of the entire range of daily precipitation using a hybrid probability distribution. *Water Resour. Res.* **2012**, *48*, W0352. [CrossRef]
14. Kollu, R.; Rayapudi, S.; Narasimham, S.; Pakkurthi, K. Mixture probability distribution functions to model wind speed distributions. *Int. J. Energy Environ. Eng.* **2012**, *3*, 1–27. [CrossRef]
15. Osatohanmwen, P.; Oyegue, F.; Ogbonmwan, S.; Muhwava, W. A General Framework for Generating Three-Components Heavy-Tailed Distributions with Application. *J. Stat. Theory Appl.* **2024**, *23*, 290–314. [CrossRef]
16. Guillen, M.; Buch-larsen, T.; Nielsen, J.P.; Bolance, C. Kernel density estimation for heavy-tailed distributions using the champernowne transformation. *Statistics* **2007**, *39*, 503–516. [CrossRef]
17. Pickands, J., III. Statistical inference using extreme order statistics. *Ann. Stat.* **1975**, *3*, 119–131. [CrossRef]
18. Cooray, K.; Ananda, M.M.A. Modeling actuarial data with a composite lognormal-Pareto model. *Scand. Actuar. J.* **2005**, *2005*, 321–334. [CrossRef]
19. Scollnik, D.P. On composite Lognormal-Pareto models. *Scand. Actuar. J.* **2007**, *2007*, 20–33. [CrossRef]
20. Cooray, K. The Weibull–Pareto Composite Family with Applications to the Analysis of Unimodal Failure Rate Data. *Commun. Stat. Theory Methods* **2009**, *38*, 1901–1915. [CrossRef]
21. Cooray, K.; Gunasekera, S.; Ananda, M. Weibull and inverse Weibull composite distribution for modeling reliability data. *Model Assist. Stat. Appl.* **2010**, *5*, 109–115. [CrossRef]
22. Scollnik, D.P.; Sun, C. Modeling with Weibull-Pareto models. *N. Am. Actuar. J.* **2012**, *16*, 260–272. [CrossRef]
23. Nadarajah, S.; Bakar, S. New composite models for the Danish fire insurance data. *Scand. Actuar. J.* **2014**, *2014*, 180–187. [CrossRef]
24. Bakar, S.A.A.; Hamzah, N.A.; Nadarajah, S. Modeling loss data using composite models. *Insur. Math. Econ.* **2015**, *61*, 146–154. [CrossRef]
25. Debbabi, N.; Kratz, M. A new unsupervised threshold determination for hybrid models. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 3440–3444.
26. Levenberg, K. A method for the solution of certain nonlonear problems in least squares quart. *Appl. Math* **1944**, *2*, 164–168. [CrossRef]
27. Marquardt, D. An algorithm for least-squares estimation of nonlinear parameters. *SIAM J. Appl. Math.* **1963**, *11*, 431–441. [CrossRef]
28. Kakutani, S. A generalization of Brouwer's fixed point theorem. *Duke Math. J.* **1941**, *3*, 457–459. [CrossRef]
29. Debbabi, N.; Kratz, M.; Mboup, M. *A Self-Calibrating Method for Heavy Tailed Modeling: Application in Neuroscience and Finance*; ESSEC Working Paper 1619; ESSEC: Cergy, France, 2016.
30. Scarrott, C.J.; MacDonald, A. A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT-Stat. J.* **2012**, *10*, 33–60. [CrossRef]
31. Hu Y.; Scarrott, C.J. evmix: An R Package for Extreme Value Mixture Modeling, Threshold Estimation and Boundary Corrected Kernel Density Estimation. *J. Stat. Softw.* **2018**, *84*, 1–27. [CrossRef]
32. Hill B.M. A simple general approach to inference about the tail of a distribution. *Ann. Stat.* **1975**, *3*, 1163–1174. [CrossRef]
33. Beirlant, J.; Goegebeur, Y.; Segers, J.; Teugels, J. *Statistics of Extremes: Theory and Applications*; Wiley Series in Probability and Statistics; John Wiley and Sons: Hoboken, NJ, USA, 2004. [CrossRef]
34. Hosking, J.; Wallis, J. Parameter and quantile estimation for the generalized pareto distribution function. *Technometric* **1987**, *29*, 339–349. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.