Earth Science Informatics (2022) 15:1607–1620
https://doi.org/10.1007/s12145-022-00811-w

RESEARCH ARTICLE

A notable bounded probability distribution for environmental and lifetime data

Hassan S. Bakouch[1,2] · Tassaddaq Hussain[3] · Christophe Chesneau[4] · Tamás Jónás[5]

Received: 26 December 2021 / Accepted: 17 April 2022 / Published online: 12 May 2022

Abstract

In this article, we introduce a notable bounded distribution based on a modification of the epsilon function that creates an upper bound on the domain of a distribution. Further, a key feature of the distribution is to have asymptotic connections with the famous Lindley distribution, which is a weighted variant of the exponential distribution and also a mixture of exponential and gamma distributions. In some ways, the proposed distribution provides a flexible solution to the modeling of bounded characteristics that can be almost well-fitted by the Lindley distribution if the domain is restricted. Moreover, we have also explored its application, particularly with reference to lifetime and environmental points of view, and found that the proposed model exhibits a better fit among the competing models. Namely, we demonstrate the practical applicability of the new distribution on two data sets containing lifetime data, as well as on two other data sets of rainfall data. Further, from the annual rainfall analysis, the proposed model exhibits a realistic return period of the rainfall.

Keywords Epsilon distribution · Lindley distribution · Practical analysis · Applications · Hydrological measure

Introduction

While observing real life phenomena, one usually comes across finite range of changes. Such finite changes generally give rise to bounded domain distributions. Among these bounded distributions, an upper bound is very helpful in analysing the annual stream flow and annual rainfall data (see (Phien and Ajirajah 1984)). The most common bounded domain distributions are the uniform, power, Bates, arcsine, Kumaraswamy, Topp-Leone, beta, triangular, raised cosine, and von Mises distributions. As an alternative to these distributions, (Dombi et al. 2018) recently introduced the epsilon distribution (EpD). Mathematically, it is based on the epsilon function defined by

$$\varepsilon_{\lambda,d}(x) = \begin{cases} \left(\frac{d+x}{d-x}\right)^{\lambda\frac{d}{2}}, & \text{if } -d < x < d, \\ 0, & \text{otherwise,} \end{cases}$$

where $\lambda \in \mathbb{R}$, $\lambda \neq 0$ and $d > 0$. This function is derived from the first-order epsilon differential equation, and it has the following exponential limit property: For any $x \in (-d, +d)$, if $d \to \infty$, then $\varepsilon_{\lambda,d}(x) \to e^{\lambda x}$. Hence, a continuous random variable X is said to have an epsilon distribution with the parameters $\lambda > 0$ and $d > 0$ if its cumulative distribution function (CDF) is given by

$$F^*_{\lambda,d}(x) = \begin{cases} 0, & \text{if } x \leq 0, \\ 1 - \varepsilon_{-\lambda,d}(x), & \text{if } 0 < x < d, \\ 1, & \text{if } x \geq d. \end{cases}$$

As a result, the epsilon distribution is a bounded domain distribution with two parameters, and it satisfies the following limit property: $\lim_{d\to+\infty} F^*_{\lambda,d}(x) = F^*_\lambda(x)$, where $F^*_\lambda(x)$ is the CDF of the exponential distribution with parameter λ. Among the applications, according to (Dombi et al. 2018), the epsilon distribution can be used to describe the mortality and useful life cycle in the sense of reliability management under the assumption of a typical bathtub-shaped failure (hazard) rate.

In this paper, we propose a notable two-parameter distribution that is also based on the epsilon function, but connected to the famous Lindley distribution, instead of the exponential distribution. The Lindley distribution has a

Communicated by: H. Babaie

✉ Tamás Jónás
jonas@gtk.elte.hu

Extended author information available on the last page of the article.

plural interest. First, it was created by Lindley (see (Lindley 1958) and (Lindley 1965)). It was first coined to express the distinction between fiducial and posterior distributions, and it has been widely used in mathematical theory and practice in recent years. Let us recall that the CDF of the Lindley distribution with a parameter $\lambda > 0$ is given by

$$F_{\lambda}^{o}(x) = \begin{cases} 0, & \text{if } x \leq 0, \\ 1 - \left(1 + \frac{\lambda x}{1+\lambda}\right) e^{-\lambda x}, & \text{if } x > 0. \end{cases}$$

The Lindley distribution has been used to analyze large amounts of data, especially in the context of stress resistance and reliability modeling. There is a substantial literature on the Lindley distribution. Let us mention the Lindley distribution's dominance over the exponential distribution of banking customers' waiting times until service, as highlighted by (Ghitany et al. 2008), the Lindley distribution's applications in lifetime data in the context of competing risks, as presented by (Mazucheli and Achcar 2011), and a comparison study of the adequacy of exponential and Lindley distributions, as presented by (Shanker and Mishra 2013) and (Shanker et al. 2015), among others.

By capturing the idea of the epsilon distribution and adapting it to reach the Lindley distribution as a limit, we motivate the use of the following function:

$$F_{\lambda,d}(x) = \begin{cases} 0, & \text{if } x \leq 0, \\ 1 - \left(1 + \dfrac{\lambda}{1+\lambda} \dfrac{dx}{d-x}\right) \varepsilon_{-\lambda,d}(x), & \text{if } 0 < x < d, \\ 1, & \text{if } x \geq d, \end{cases} \quad (1)$$

where $\lambda > 0$ and $d > 0$. The distribution defined by $F_{\lambda,d}(x)$ is called the epsilon-Lindley distribution (EpLD). Then, one can prove that it is a valid CDF, which satisfies $\lim_{d \to +\infty} F_{\lambda,d}(x) = F_{\lambda}^{o}(x)$; the Lindley distribution is a limit case of the EpLD, which is a rare property for a bounded support distribution. Furthermore, the related functions, such as the probability density function (PDF) and hazard rate function (HRF) are very flexible in their behaviour, as shown later. More precisely, by using a graphical analysis, the PDF adopts various shapes, like skewed to the right with J-shapes as well as an upside down U-shape. In all these cases, we observe a positive skewness and leptokurtic nature of the curve, which clearly indicates that it is designed to model the heavy-tailed phenomenon. Such phenomena are generally common in reliability applications, queuing theory and environmental aspects. In this regard, we focus on the environmental aspect and also lifetime direction. Our application section will help the reader to reach a decision to forecast the next generation's future in a better way. Environmental data analysis is based on the most efficient bounded models. One can mention the three-parameter lognormal distribution, generalized extreme value type II distribution, generalized extreme value type III distribution, three-parameter gamma distribution, and three-parameter log-Pearson distribution, among others. The non-closed form of the CDF in these popular hydrological models is a fundamental flaw, whereas the suggested model is based on only two parameters and has a closed form of its basic functions, including its CDF, which makes the determination of the return period considerably easier.

Further, the HRF adopts various shapes, from bathtub to increasing failure rate with a left skewed J-shape. This functional flexibility is a true plus for the EpLD from the modeling viewpoint. On the mathematical plan, the EpLD is a weighted version of the EpD. This weighted version not only models ascertainment biases but also a linear combination of probability distributions. We thus develop the statistical features offered by the EpLD through diverse aspects, including theoretical and practical facts. The practical lines interested in fitting, modeling and analysis of lifetime and environmental data are outlined by the proposed model. Here, we demonstrate the practical applicability of the EpLD on two data sets containing lifetime data, as well as on two other data sets of rainfall data. Further, from the annual rainfall analysis, we found a realistic return period of the rainfall by the proposed model.

The organization of the paper is as follows. Section Some related functions, properties with estimations presents some other functions of interest in the EpLD, like moments and parameter estimation. Section Model compatibility and its application to real-world data covers the application area of the proposed model. Section Conclusions and future research plans deals with conclusion and closing comments about the proposed distribution with future research plans.

Some related functions, properties with estimations

Related functions

We now illustrate the shape behavior of the main functions of the EpLD. First, let us focus on the CDF as defined in Eq. 1. Figure 1 presents some graphs of this CDF for several values of the parameters.

Figure 1 depicts that for smaller values of both d and λ, the convergence of CDF to 1 is very slow compared to that for larger d and λ values.

Let us now focus on the related PDF. The PDF of the EpLD is expressed as

$$f_{\lambda,d}(x) = \begin{cases} \frac{\lambda d^2}{(1+\lambda)(d^2-x^2)} \left[1+\lambda - \frac{d+x-\lambda dx}{d-x}\right] \varepsilon_{-\lambda,d}(x), & \text{if } 0 < x < d, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

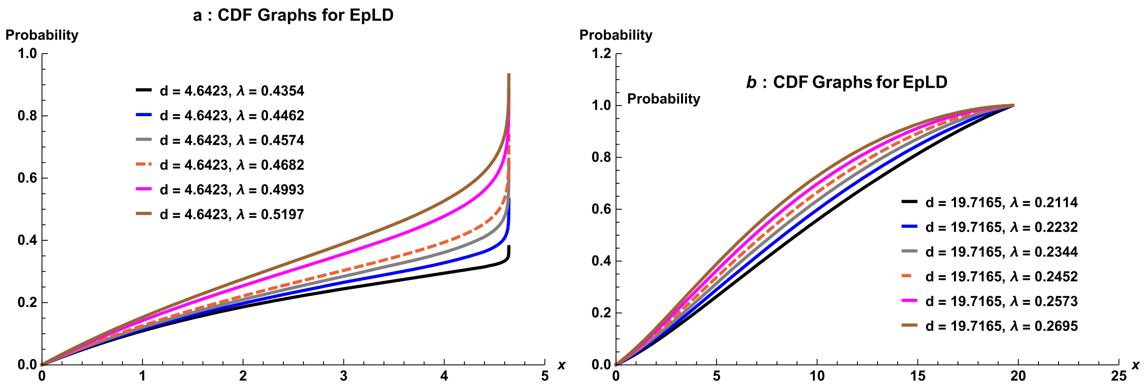


Fig. 1 Graphs of the CDF of the EpLD

Figure 2 presents some graphs of this PDF for several parameter values.

From Figs. 2 and 3, we see that the PDF of the EpLD also adopts various shapes, like unimodal shapes skewed to the right, various J-shapes, as well as an upside down U-shapes. In particular, Fig. 2 indicates that, when the rate parameter λ decreases, it increases the probability of events whatever the boundary value of d is. Thus, the PDF of the EpLD distribution is extremely flexible, and, as developed in the introductive section, motivates the use of the EpLD for various modelling purposes, including the heavy-tailed phenomenon.

In addition to its practical ability, the PDF of the EpLD has interesting mathematical decompositions. Indeed, one can view $f_{\lambda,d}(x)$ as follows:

- It is a weighted version of the PDF of the EpD, because it can be written as $f_{\lambda,d}(x) = w_{\lambda,d}(x) f^*_{\lambda,d}(x)$, where

  $$w_{\lambda,d}(x) = 1 - \frac{d + x - \lambda d x}{(1+\lambda)(d-x)}$$

  and $f^*_{\lambda,d}(x)$ refers to the PDF of the epsilon distribution.

- If $\lambda > 2/d$, by noticing that

  $$\frac{d + x - \lambda d x}{(1+\lambda)(d-x)} = \frac{1}{1+\lambda}\left[\frac{\lambda}{4}(d-x) + \left(1 - \frac{\lambda}{4}(d+x)\right)\frac{d+x}{d-x}\right],$$

  we can also write $f_{\lambda,d}(x)$ as a linear combination of PDFs and lenght biased PDFs of the EpD as

  $$\begin{aligned} f_{\lambda,d}(x) &= \left(1 - \frac{\lambda}{4(1+\lambda)}(d-x)\right) f^*_{\lambda,d}(x) \\ &\quad - \frac{\lambda d}{(1+\lambda)(\lambda d - 2)}\left(1 - \frac{\lambda}{4}(d+x)\right) f^*_{\lambda-2/d,d}(x) \\ &= \left(1 - \frac{\lambda d}{4(1+\lambda)}\right) f^*_{\lambda,d}(x) \\ &\quad - \frac{\lambda d}{(1+\lambda)(\lambda d - 2)}\left(1 - \frac{\lambda d}{4}\right) f^*_{\lambda-2/d,d}(x) \\ &\quad + \frac{\lambda}{4(1+\lambda)} x f^*_{\lambda,d}(x) + \frac{\lambda^2 d}{4(1+\lambda)(\lambda d - 2)} x f^*_{\lambda-2/d,d}(x). \end{aligned} \tag{3}$$

  This expansion is useful to determine several probabilistic quantities related to the EpLD.

As a major reliability function (see, e.g., (Nair et al. 2018)) of the EpLD, the HRF is specified as

$$h_{\lambda,d}(x) = \begin{cases} \frac{\lambda d^2}{(1+\lambda)(d^2-x^2)}\left[1+\lambda-\frac{d+x-\lambda d x}{d-x}\right]\left(1+\frac{\lambda}{1+\lambda}\frac{dx}{d-x}\right)^{-1}, & \text{if } 0 < x < d, \\ 0, & \text{otherwise.} \end{cases}$$

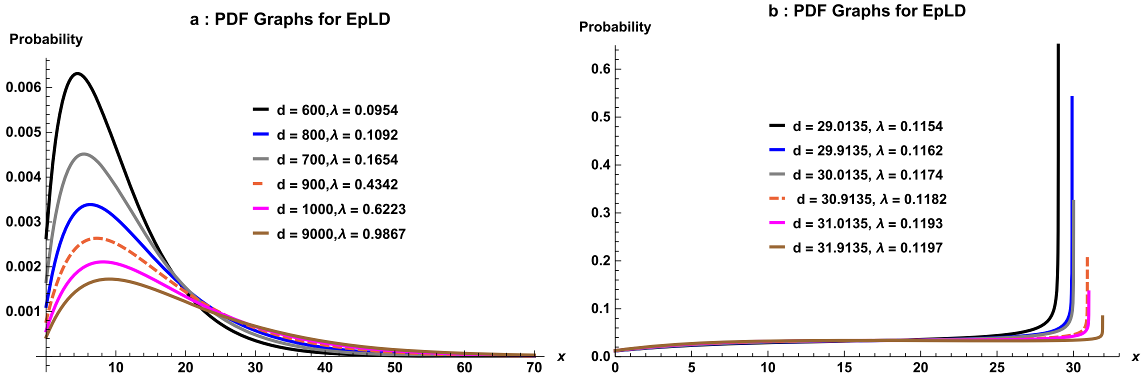


Fig. 2 Graphs of the PDF of the EpLD; unimodal shapes skewed to the right and J shapes

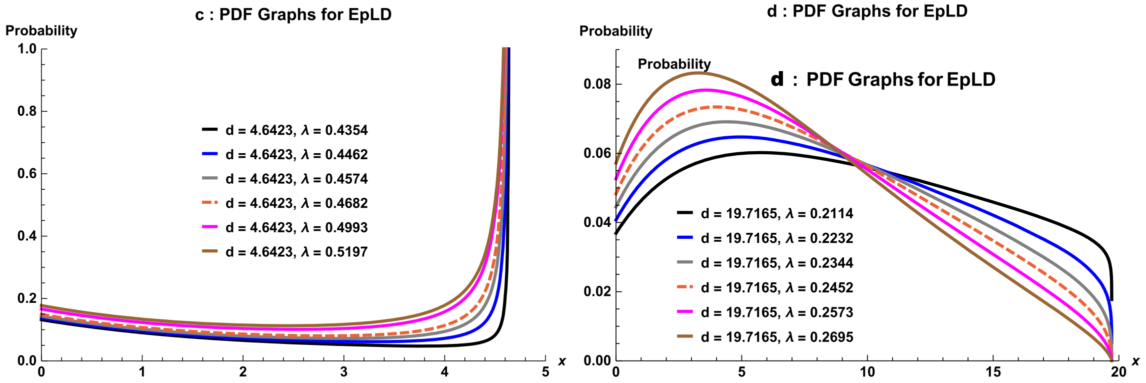


Fig. 3 Graphs of the PDF of the EpLD; U-J shapes and upside down U-shapes

Figure 4 provides some graphs of this HRF for selected values of the parameters.

The graphs in Fig. 4 clearly portray the HRF behaviour like increasing and bathtub-shaped in an impressive way.

We end this part by discussing the quantile analysis of the EpLD. As we know, in traditional probability and statistics as well as in stochastic analysis, the quantile function (QF) deals with a valuable way of describing a static or vigorous distribution, as a result, knowing how to use this function indicates certain advantages not available straight from the CDF or PDF. For example, the simplest way of simulating any non-uniform random variable is by applying its QF to uniform deviates. Similarly, from an environmental point of view, these functions usually help environmental scientists to calculate the return period and return level of any distribution.

In view of above, the QF of the EpLD, denoted by $Q_{\lambda,d}(u)$ with $u \in (0,1)$, is the solution of the following non-linear equation:

$$\left(1+\frac{\lambda}{1+\lambda}\frac{dQ_{\lambda,d}(u)}{d-Q_{\lambda,d}(u)}\right)\varepsilon_{-\lambda,d}[Q_{\lambda,d}(u)] = 1-u.$$

To evaluate $Q_{\lambda,d}(u)$ at given u and parameters, it is clear that a mathematical software is required.

Moments

In mathematics, and in statistics in particular, the word moments of a function are reckonable procedures associated to the shape of the function's graph. If the function represents density or mass function, then the first moment represent the center of the mass or expected value, and the second moment is the rotational inertia or the variance. So, the moments about the origin of the EpLD can be determined by using the expansion in Eq. 3. For a random variable $Y_{\lambda,d}$ following the epsilon distribution with parameters λ and d and a random variable X following the EpLD, the r-th moment of X can be obtained as

$$\begin{aligned}\mu'_r &= E(X^r)\\ &= \left(1-\frac{\lambda d}{4(1+\lambda)}\right)E(Y^r_{\lambda,d})\\ &\quad -\frac{\lambda d}{(1+\lambda)(\lambda d-2)}\left(1-\frac{\lambda d}{4}\right)E(Y^r_{\lambda-2/d,d})\\ &\quad +\frac{\lambda}{4(1+\lambda)}E(Y^{r+1}_{\lambda,d})+\frac{\lambda^2 d}{4(1+\lambda)(\lambda d-2)}E(Y^{r+1}_{\lambda-2/d,d}).\end{aligned}$$

The r-th and $(r+1)$-th moments of $Y_{\lambda,d}$ are well established, see (Okorie and Nadarajah 2019). In a similar way, we can express the incomplete moments of X in terms of

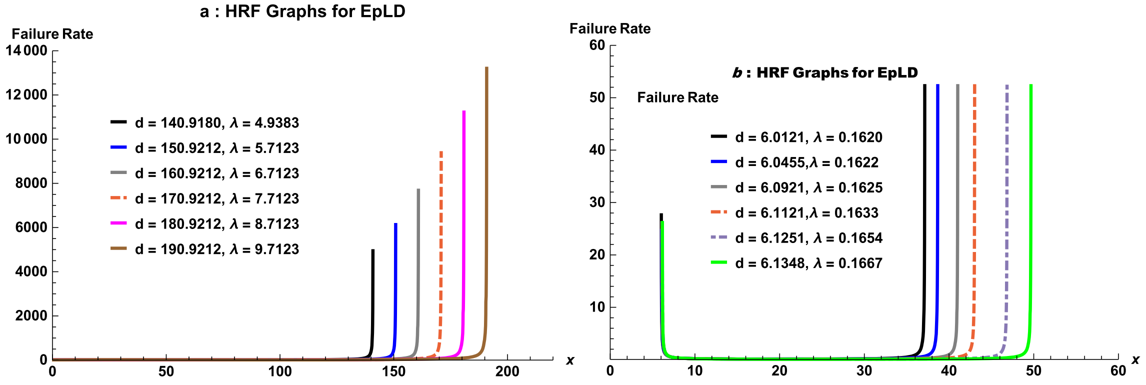


Fig. 4 Graphs of the HRF of the EpLD

incomplete moments of $Y_{\lambda,d}$. Thus, the mean and variance of X can be obtained.

Similarly, the ratio of third mean moment to the square of second mean moment are the skewness and the ratio fourth moment about mean to second moment about mean is the kurtosis. From these moments, we are now able to interpret the shape and kurtosis behaviour of the EpLD. In this regard, Fig. 5 portrays that the proposed model can exhibit versatile shapes ranging from negative to positive behaviour. In addition, we see that the distribution also has the leptokurtic, mesokurtic and platykurtic behaviour.

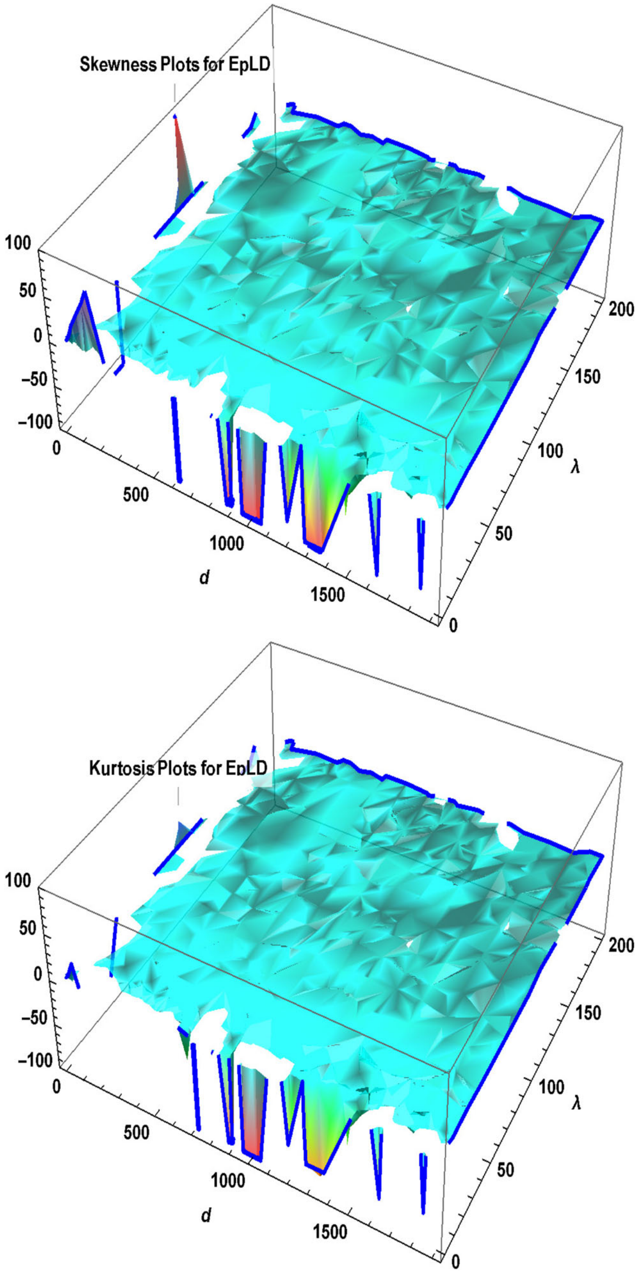


Fig. 5 Graphs of skewness and kurtosis of the EpLD

Parameters estimation

Due to the importance of statistical inference, here, we adopt the maximum likelihood method, giving the maximum likelihood estimates (MLEs) of the unknown parameters. MLEs are designed to follow the regularity conditions, which are usually helpful for constructing the confidence intervals and the test statistics. For these estimators, the large sample theory yields straightforward approximations that work well in finite samples. In order to achieve better approximation distributions, statisticians frequently strive to estimate quantities, such as the distribution of a test statistic that depends on the sample size. The resulting MLE approximation in distribution theory can be handled analytically or numerically with ease. Only complete samples are used to calculate the MLEs of the EpLD parameters. In this regard, let $x_1, \ldots, x_n$ be a realization of a random sample of size n from the EpLD given by Eq. 2. Then, the log-likelihood function of the EpLD is given by

$$\begin{aligned}\ell(\lambda, d) &= 2n\log d + n\log\lambda - n\log(1+\lambda)\\ &\quad -\left(1+\frac{d\lambda}{2}\right)\sum_{i=1}^{n}\log\left(\frac{d+x_i}{d-x_i}\right) \\ &\quad +\sum_{i=1}^{n}\log\left[d\lambda + x_i((d-1)\lambda-2)\right] - 3\sum_{i=1}^{n}\log(d-x_i).\end{aligned} \tag{4}$$

The log-likelihood can be maximized either directly by using the Mathematica [12.0] or by solving the nonlinear likelihood equations obtained by differentiating Eq. 4. In particular, we have

$$\begin{aligned}\frac{\partial\ell(\lambda,d)}{\partial\lambda} &= \frac{n}{\lambda} - \frac{n}{1+\lambda} - \frac{d}{2}\sum_{i=1}^{n}\log\left(\frac{d+x_i}{d-x_i}\right)\\ &\quad +\sum_{i=1}^{n}\frac{d+(d-1)x_i}{d\lambda+((d-1)\lambda-2)x_i} = 0.\end{aligned}$$

The MLE of the parameter λ is obtained by solving the nonlinear system $\partial\ell(\lambda,d)/\partial\lambda = 0$. As mentioned earlier, this equation cannot be solved analytically, so we prefer to use statistical packages like Mathematica [12.0]. For this purpose, we use Global MLE of the proposed model that take the Lindley distribution MLE as seed value. However, we observe that we cannot obtain the estimate of d from Eq. 4. Consequently, we adopt the methodology described as follows: since d is free from x and it is the upper limit in the domain of x, we consider $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, the ordered sample corresponding to $x_1, x_2, \ldots, x_n$, and, based on them, the estimate d as $\hat{d} = \max(x_1, x_2, \ldots, x_n) + \upsilon$, where $\upsilon > 0$ denotes an arbitrary constant. However, when we start estimating d using the sample, we undertake that all the elements in the sample are in the domain of the random variable. This is because the sample should be comprised of independent observations. Since the value of parameter d determines the domain of attraction of the random variable that has a EpLD, in the estimation, it is a necessity that d is greater than the maximal element in the sample. So we are looking for the robust value of $\hat{d}$, which will probably be established under the condition that $\hat{d}$ is greater than the largest element in the sample. For more details, see (Dombi et al. 2018), (Dombi et al. 2019), (Dombi and Jónás 2020) and (Dombi and Jónás 2021).

Model compatibility and its application to real-world data

In this section, we concentrate on the modelling process's model selection and model validation. However, model selection is a challenging task and the prime of a suitable model and it is produced with the use of well-considered judgement based on whatever information is available. It is essential that the chosen model be malleably sufficient to model the confronted data amply, while considering the settlement between simplicity of evaluation and the intricacy of the model. Moreover, outstanding attention must be devoted to modeling behavior for large and small values of the variable of interest. In this regard, the modelling process includes validating the model, which includes various goodness-of-fit tests and graphical procedures. These statistical techniques for assessing hypothesised models are known as goodness-of-fit tests. An unsatisfactory fit, either analytical or graphical, may occur for the following reasons: i) The model is incorrectly specified. ii) The model specification is correct, but unfortunately carries a huge bias. In general, validation necessitates more data, other information, and further testing, as well as a careful examination of the consequences.

Goodness-of-fit tests

As in such tests, researchers usually make a null hypothesis, H_0: The given data comes from a CDF with a specified form. For this purpose, we have considered four tests. The first test is the famous χ^2 test (Chi Square), due to Karl Pearson. It includes grouping observed data into intervals and may be used to assess the fit of data to any specified distribution (continuous or discrete). When using this test, a sample of size n is assumed, with each observation falling into one of k potential classifications. The observed and expected frequencies in the interval i are denoted by o_i and e_i, respectively. The test statistic is

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}.$$

However, this test has the advantages of being easy to apply and being applicable even when parameters are unknown (see (Murthy et al. 2004)). In addition, this test is not of much use in small or sometimes even modest size samples (see (Murthy et al. 2004)). The next three tests are based on the empirical cumulative distribution function (ECDF) and hence are often referred to as ECDF tests.

Kolmogorov-Smirnov (KS) Test

The Kolmogorov-Smirnov (KS) test is grounded on the ECDF. Given n ordered observations $Z_1, Z_2, \ldots, Z_k$, then the ECDF is defined as $E_k = m_i/k$ where m_i is the number of points less than Z_i and the Z_i are ordered from smallest to largest value. At the value of each ordered data point, this step function rises by $1/k$. The greatest distance between the hypothesised CDF and ECDF is the test statistic KS. The mathematical expression of the KS test statistic is given by

$$\mathrm{KS} = \max_{1 \le i \le k} \left\{ \frac{i}{k} - z_i, z_i - \frac{i-1}{k} \right\},$$

where $z_i = F(Z_i)$, and F is the theoretical CDF of the distribution being tested. Other goodness-of-fit tests, like the Anderson-Darling test and the Cramér-von Mises test, are alternatives of the KS test. As these modified tests are usually measured to be more powerful than the conventional KS test, many analysts prefer them.

Anderson-Darling (AD_0^*) Test

The Anderson-Darling (AD_0^*) test is an alternative of the KS test and usually attaches more weight to the tails than the KS test. Its test statistic is

$$A_0^* = \left(\frac{2.25}{k^2} + \frac{0.75}{k} + 1\right)\left\{-k - \frac{1}{k}\sum_{i=1}^{k}(2i-1)\log(z_i(1-z_{k-i+1}))\right\}.$$

Cramér-von Mises (CVM_0^*) Test

The CVM_0^* test is also a modification of the KS test, which is usually considered to be more powerful than the original KS test. The CVM_0^* test statistic is expressed as

$$W_0^* = \sum_{i=1}^{k}\left(z_i - \frac{2i-1}{2k}\right)^2 + \frac{1}{12k}.$$

A relative comparison of the selection of these tests indicates that: i) The ECDF tests are more powerful than the χ^2 test. ii) The KS test is the most well-known ECDF test, but it is often much less powerful than the other ECDF tests (AD_0^* and CVM_0^* tests) (see (Murthy et al. 2004)). Moreover, we have also applied information criteria for model selection purposes, such as Akaike information criterion (AIC), Bayesian information criterion (BIC), corrected Akaike information criterion (AICc), Hannan-Quinn information criterion (HQIC) and consistent Akaike information criterion (CAIC). The following are the definitions of AIC, AICc, HQIC, and CAIC:

$$\mathrm{AIC} = 2\hbar - 2l, \quad \mathrm{AICc} = \mathrm{AIC} + \frac{2\hbar(\hbar+1)}{n-\hbar-1}, \quad \mathrm{BIC} = \hbar\log(n) - 2l,$$

$$\mathrm{HQIC} = -2l + \hbar\log(\log(n)), \qquad \mathrm{CAIC} = -2l + \frac{2\hbar n}{n-\hbar-1},$$

where l denotes the estimate of the maximm log-likehood function, $\hbar$ is the number of parameters to be estimated and n is the number of data.

Along with these model selection procedures, we have also used the Kullback-Leibler information criterion philosophy and applied the Vuong test proposed by (Vuong 1989).

Vuong test

The Vuong test is a closeness test based on the likelihood-ratio-based test for model selection using the Kullback-Leibler information criterion philosophy. This test may be used for non-nested models, and it generally compares the null hypothesis that two competing models are equally near to the actual data against the alternative that one model performs better. Further discussion about the Vuong test can be found in (Hussain et al. 2019).

Competing models

We compare the proposed model with the following well-known models: epsilon probability distribution (EpD) (see (Dombi et al. 2018)), two-parameter Lindley distribution (TPLD) (see (Shanker et al. 2015)), A quasi Lindley distribution (QLD) (see (Shanker and Mishra 2013)), Lindley distribution (LD) (see (Lindley 1958)) and exponential distribution (ED). For the sake of transparency, these competitors are defined by the following PDFs:

- for the EpD:

  $$f_{\lambda,d}^{EpD}(x) = \frac{\lambda d^2}{d^2 - x^2}\varepsilon_{-\lambda,d}(x), \qquad 0 < x < d,$$

  and $f_{\lambda,d}^{EpD}(x) = 0$ for $x \notin [0, d]$, with $d > 0$,
- for the TPLD:

  $$f_{\lambda,d}^{TPLD}(x) = \frac{\lambda^2(1+dx)}{d+\lambda}e^{-\lambda x}, \qquad x > 0,$$

  and $f_{\lambda,d}^{TPLD}(x) = 0$ for $x \leq 0$, with $\lambda > 0$ and $d > -1$,
- for the QLD:

  $$f_{\lambda,d}^{QLD}(x) = \frac{\lambda(d+\lambda x)}{d+1}e^{-\lambda x}, \qquad x > 0,$$

  and $f_{\lambda,d}^{QLD}(x) = 0$ for $x \leq 0$, with $\lambda > 0$ and $d > -1$,
- for the LD:

  $$f_{\lambda}^{LD}(x) = \frac{\lambda^2(1+x)}{1+\lambda}e^{-\lambda x}, \qquad x > 0,$$

  and $f_{\lambda}^{LD}(x) = 0$ for $x \leq 0$, with $\lambda > 0$,
- for the ED:

  $$f_{\lambda}^{ED}(x) = \lambda e^{-\lambda x}, \qquad x > 0,$$

  and $f_{\lambda}^{ED}(x) = 0$ for $x \leq 0$, with $\lambda > 0$.

We consider four different real-world data sets.

Lifetime data sets

Data sets I and II. The first and second data sets are taken from (Walpole et al. 2012) and (Andrews and Herzberg 1985), respectively. The first data are about the length of life in years, measured to the nearest tenth of 30 similar fuel pumps, while the second data represent the life of fatigue fracture of Kevlar 373 epoxy that is subjected to constant pressure at the 90 stress level until all have failed. The measurements of the first data set are 2.0, 3.0, 0.3, 3.3, 1.3, 0.4, 0.2, 6.0, 5.5, 6.5, 0.2, 2.3, 1.5, 4.0, 5.9, 1.8, 4.7, 0.7, 4.5, 0.3, 1.5, 0.5, 2.5, 5.0, 1.0, 6.0, 5.6, 6.0, 1.2, 0.2. The second data set measurements are given as: 0.0251, 0.0886, 0.0891, 0.2501, 0.3113, 0.3451, 0.4763, 0.5650, 0.5671, 0.6566, 0.6748, 0.6751, 0.6753, 0.7696, 0.8375, 0.8391, 0.8425, 0.8645, 0.8851, 0.9113, 0.9120, 0.9836, 1.0483, 1.0596, 1.0773, 1.1733, 1.2570, 1.2766, 1.2985, 1.3211, 1.3503, 1.3551, 1.4595, 1.4880, 1.5728, 1.5733, 1.7083, 1.7263, 1.7460, 1.7630, 1.7746, 1.8275, 1.8375, 1.8503, 1.8808, 1.8878, 1.8881, 1.9316, 1.9558, 2.0048, 2.0408, 2.0903, 2.1093, 2.1330, 2.2100, 2.2460, 2.2878, 2.3203, 2.3470, 2.3513, 2.4951, 2.5260, 2.9911, 3.0256, 3.2678, 3.4045, 3.4846, 3.7433, 3.7455, 3.9143, 4.8073, 5.4005, 5.4435, 5.5295, 6.5541, 9.0960. In this regard, we have compiled the descriptive statistics, which are listed in Table 1, and the total time on test (TTT) plots introduced by (Aarset 1987), which are portrayed in Fig. 6 for Data sets I and II.

Table 1 Descriptive statistics for Data sets I and II

Dataset	Sample size	Mean	Median	Standard deviation	Skewness	Kurtosis	$\frac{Skewness}{Kurtosis}$
I	30	2.7967	2.1500	2.2273	0.3412	1.5689	0.2175
II	76	1.9592	1.5335	1.6753	1.9796	8.1608	0.2426

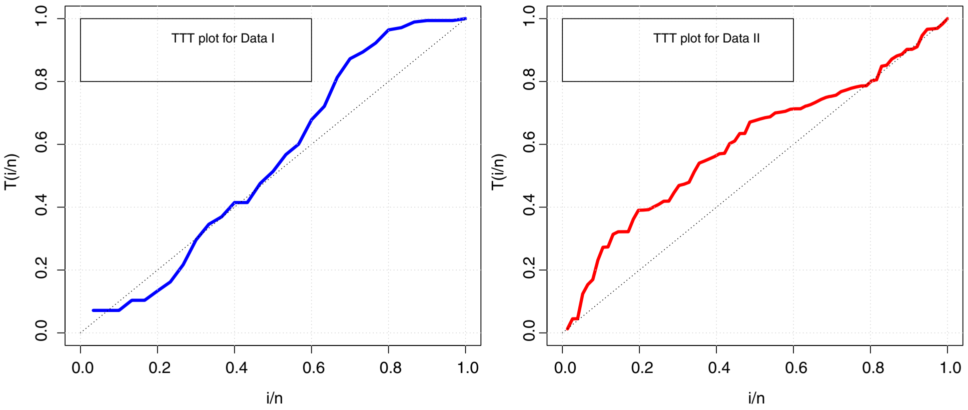


Fig. 6 Estimated TTT plots of Data sets I and II

Discussion and analysis of lifetime data sets

Tables 1 and 2 reveal that theoretical and observed descriptive statistics show a remarkable closeness to each other and it seems that both data sets are being simulated by the proposed model.

Note that as the parameter d specifies the support of the PDF of the EpLD in Equation (2), i.e., it is positive only if $x \in (0, d)$. This means that the value of parameter d must meet the requirement $d > \max_{i=1,2,\ldots,n}(x_i)$, see (Dombi et al. 2019). That is why we observed that $\hat{d}$ is large for any data set, as shown in the related tables. From the TTT-plot for Data set I, we can see that the curve has three characteristic phases: (1) a first convex phase, where the failure rate is decreasing; (2) a second quasi linear phase with a constant failure rate; (3) and a third concave phase, where the failure rate is increasing. That is, the TTT-plot for Data set I (see Fig. 6) portrays a bathtub-shaped like failure rate curve. Noting the TTT-plot for Data set II in Fig. 6, from which we can conclude that it exhibits an increasing failure rate phenomenon of the empirical failure rate function. Hence, both of the above data sets are efficiently modelled by the proposed model. These results are in line with the fact that the HRF of the EpLD can be bathtub-shaped or increasing (see Fig. 4). Since d is large, the EpLD is almost identical to the LD, and so they have almost the same goodness-of-fit statistics values. Such a suitability of the proposed model is reflected in Tables 3 and 4, where the EpLD yields a smallest value of the goodness-of-fit statistics along with highest p-vlaue for χ^2 statistics. In addition, we have also assessed the performance of the model with respect to the LD via the log-likelihood ratio test, which is usually applicable for nested models, and we drew the same conclusion.

However, Tables 5 and 6 portray that the QLD and LD yield minimum values of information criterion, which seem to be a penalty of over parametrization, particularly with reference to the LD model. Previously we pointed out that the LD distribution may be viewed as an asymptotic EpLD distribution, i.e., if $d \to \infty$, then the EpLD distribution is identical to the LD distribution. We can observe a practical implication of this finding in Tables 3–6. Namely, when a data set can be modelled well by the LD distribution, then it can also be modelled well by the EpLD distribution with a sufficiently large value of the parameter d, and vice versa. Certainly, in such a case, the estimates of the λ parameter and the corresponding goodness-of-fit statistics are very close.

Furthermore, Table 7 also pleads for the suitability of the proposed model. But Vuong statistics show that QLD and LD are strong competitors for the proposed model.

Furthermore, by the histogram analysis performed in Fig. 7, we see that the proposed model matches the data in a better way than the competing models.

Environmental Data Sets

Data sets III and IV. The third and fourth data sets are the total amount of rainfall in mm of Pakistani cities Lasbella and Bunji, which covers a period of 30 years (1981 to 2010) with 30 values of annual rainfall in each

Table 2 Theoretical statistics from the EpLD

Data set	Sample size	Mean	Median	Standard deviation	Skewness	Kurtosis	$\frac{Skewness}{Kurtosis}$
I	30	2.7802	2.2285	2.2505	1.1766	4.2477	0.2769
II	76	1.9589	1.5739	1.7362	1.4309	5.4976	0.2602

Table 3 MLEs and goodness-of-fit statistics for Data set I

Dist.	$\hat{\lambda}$	$\hat{d}$	CVM$_0^*$	AD$_0^*$	KS	$\chi^2(df)$	p-value
EpLD	0.5839	1219.9011	0.5059	0.0728	0.1408	1.0760(2)	0.5839
EpD	0.3575	6.31×10^{6}	0.7569	0.1189	0.1358	1.6246(2)	0.4438
TPLD	0.3575	-1.92×10^{-26}	0.7569	0.1189	0.1358	1.6246(2)	0.4438
QLD	0.5123	1.3097	0.5512	0.0809	0.1341	1.0990(2)	0.5772
LD	0.5834	…	0.5061	0.0729	0.1409	1.0767(2)	0.5831
ED	0.3576	…	0.7569	0.1189	0.1358	1.6246(2)	0.4438

Table 4 MLEs and goodness-of-fit statistics for Data set II

Dist.	$\hat{\lambda}$	$\hat{d}$	CVM$_0^*$	AD$_0^*$	KS	$\chi^2(df)$	p-value
EpLD	0.7948	99910.0267	1.4902	0.2667	0.1156	6.8497(6)	0.3349
EpD	0.5104	40763.1154	3.2134	0.5746	0.1663	13.4562(6)	0.0363
TPLD	0.5104	8.29×10^{-10}	3.0187	0.5746	0.1663	13.4562(6)	0.0363
QLD	0.9542	0.1499	0.6003	0.1016	0.1025	9.3621(8)	0.3126
LD	0.7947	…	1.4902	0.2668	0.1156	6.8496(6)	0.3348
ED	0.5104	…	3.0188	0.5746	0.1663	13.4562(6)	0.0363

Table 5 Estimates of the maximum log-likelihood and information criteria for Data set I

Distribution	$-l$	AIC	AICC	BIC	HQIC	CAIC
EpLD	60.7645	125.529	125.973	128.331	123.977	125.973
EpD	60.8525	125.705	126.149	128.507	124.153	126.149
TPLD	60.8529	125.706	126.15	128.508	124.154	126.15
QLD	60.4864	124.973	125.417	127.775	123.421	125.417
LD	60.7668	123.534	123.676	124.935	123.982	123.676
ED	60.8528	123.706	123.848	125.107	124.154	123.848

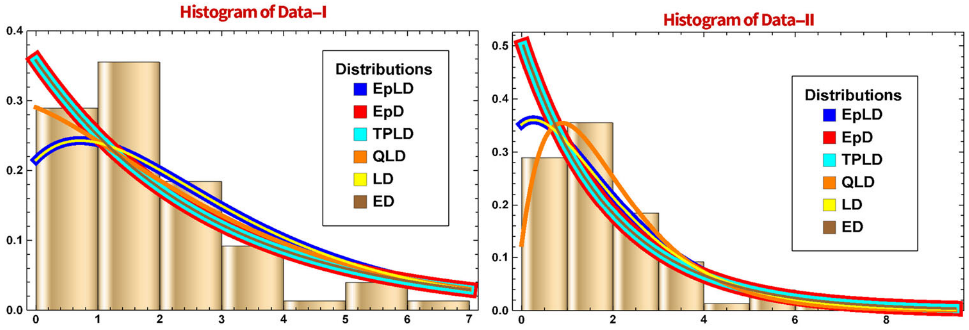
Table 6 Estimates of the maximum log-likelihood and information criteria for Data set II

Distribution	$-l$	AIC	AICC	BIC	HQIC	CAIC
EpLD	123.674	251.348	251.512	256.009	250.279	251.512
EpD	127.114	258.228	258.392	262.889	257.159	258.392
TPLD	127.114	258.228	258.392	262.889	257.159	258.392
QLD	121.65	247.3	247.464	251.961	246.231	247.464
LD	123.675	249.35	249.404	251.681	250.281	249.404
ED	127.114	258.228	258.392	262.889	257.159	258.392

Table 7 Vuong test statistics for Data sets I and II

Models	Data set I	Suitability	Data set II	Suitability
EpLD- EpD	2.7872	EpLD	32.5790	EpLD
EpLD-TPD	2.7872	EpLD	32.5804	EpLD
EpLD-QLD	-0.9287	Indecisive	-15.3623	QLD
EpLD-LD	1513.008	EpLD	-1321291	LD
EpLD-ED	2.7872	EpLD	32.5790	EpLD

Fig. 7 Data sets I and II fits via histograms



set. They were reported by (Hussain et al. 2019). The third data set measurements are as follows: 138.11818, 89.5, 246.5, 142.6, 143.5, 47.4, 105.7, 182.6, 153.7, 119.9, 56.5, 272.8, 99.9, 426.1, 205.6, 169.8, 308.3, 80.5, 104.0, 37.7, 223.0, 9.2, 474.6, 25.3, 209.6, 182.5, 196.2, 254.9, 103.6, 117.6. The fourth data set contains the following rainfall measurements: 248.8, 82.2, 102.2, 217.9, 113.2, 248.2, 244.1, 122.2, 144.9, 63.2, 62.8, 139, 228.7, 216.4, 144.8, 252.6, 144.8, 157.2, 168.5, 139.1, 74.3, 154.6, 339.4, 154.1, 156.3, 200.7, 97.5, 96.3, 155.2, 298.8. The descriptive statistics of these data sets and corresponding theoretical statistics from the EpLD are presented in Tables 8 and 9, respectively. Box-plots of the data are given in Fig. 8.

In order to analyse the environmental data, we have also checked some features of the environmental data, namely homogeneity, independence and stationarity. For this purpose, we applied the Mann-Whitney (M-W) test for testing homogeneity and stationarity, and the Mann-Kendall (M-K) test for trend detection. In this regard, we have observed that both data sets accept the hypotheses of homogeneity and stationarity at a 5 percent level of significance with Z-scores of 0.3568 and -0.2777, respectively. Similarly, the hypothesis of independence and identically distributed distribution is accepted at a 5 percent level of significance with Z-scores of 0.4817 and -0.6943, respectively. For details of these tests, readers are referred to (Haktanir et al. 2013).

Analysis and discussion of environmental data

From Table 8 and Fig. 8 as well as Table 9, it is obvious that the empirical and theoretical aspects of the data sets in the presence of outliers in Data set III are in close agreement. These indicate that the model can effectively be used if the data are positively skewed and leptokurtic in nature, which are the obvious characteristics of environmental data. Such findings are further consolidated by viewing Tables 10 and 11, which portray that the EpLD exhibits minimum values of goodness-of-fit statistics.

Tables 10 and 11 indicate that ECDF test statistics for goodness-of-fit tests are low, which ensures that the EpLD is a good competitor to the QLD and LD.

However, likelihood aspects and information criterion values also favour the proposed model, which can be visualized in Tables 12 and 13, respectively.

Furthermore, the shape of our proposed model, as shown in Fig. 9, matches the data in a better way compared to the competing models.

Furthermore, the Vuong statistics as depicted in Table 14 also show the capability of the proposed model.

Hydrological parameters

The annual series is very common in frequency analysis for two reasons. The earliest is its accessibility, as most data are managed in such a way that the annual series is commonly available. The other one is that there is a simple hypothetical basis for deducing the frequency of annual series data beyond the range of observation (see (World Meteorological Organization 2009)). Moreover, we observed that both series are valid and approved by the M-W and M-K tests, as shown in the earlier section.

Therefore, it can be settled that, according to the consequences of relevant tests, both the annual rainfall

Table 8 Descriptive statistics for Data sets III and IV

Data set	Sample size	Mean	Median	Standard deviation	Skewness	Kurtosis	$\frac{Skewness}{Kurtosis}$
III	30	164.241	143.05	108.163	1.1201	4.2629	0.2627
IV	30	165.6	154.35	70.3731	0.5800	2.7073	0.2142

Table 9 Theoretical statistics from the EpLD

Data set	Sample size	Mean	Median	Standard deviation	Skewness	Kurtosis	$\frac{Skewness}{Kurtosis}$
III	30	163.952	138.302	114.483	1.0747	4.0723	0.2639
IV	30	176.0789	160.9816	102.6899	0.5248	2.5549	0.2054

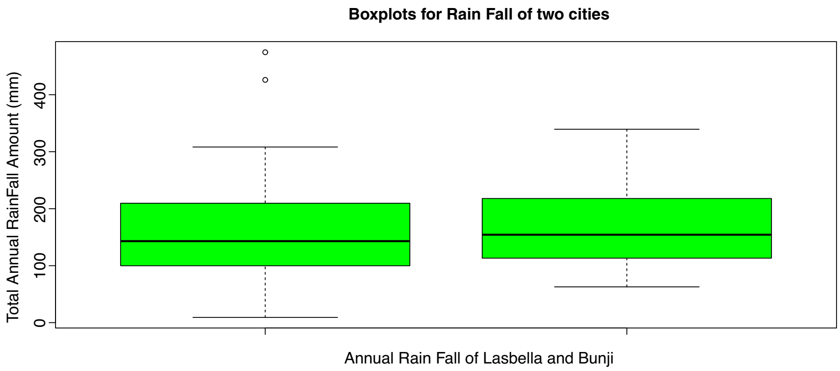


Fig. 8 Box-plots for Data sets III and IV

Table 10 MLEs and goodness-of-fit statistics for Data set III

Dist.	$\hat{\lambda}$	$\hat{d}$	CVM_0^*	AD_0^*	KS	$\chi^2(df)$	p-value
EpLD	0.0122	8329.81	0.0472	0.2778	0.1117	0.2294(2)	0.8916
EpD	0.0041	9.29×10^6	1.8655	0.3667	0.2417	3.6401(2)	0.1620
TPLD	0.00411	0.0000	0.3667	1.8655	0.2417	3.6401(2)	0.1620
QLD	0.0121	0.0159	0.0479	0.2796	0.1122	0.2350(2)	0.8891
LD	0.0121	…	0.0473	0.2779	0.1121	0.2296(2)	0.8915
ED	0.0061	…	0.3549	1.8494	0.2224	3.3397(2)	0.1883

Table 11 MLEs and goodness-of-fit statistics for Data set IV

Dist.	$\hat{\lambda}$	$\hat{d}$	CVM_0^*	AD_0^*	KS	$\chi^2(df)$	p-value
EpLD	0.0123	498.906	0.2002	1.3785	0.1650	2.5612(1)	0.1096
EpD	0.0060	9.25×10^7	0.8954	4.6593	0.3156	10.7892(1)	0.0010
TPLD	0.0060	6.56×10^{-13}	0.8954	4.6594	0.3156	10.3156(1)	0.0013
QLD	0.0121	-4.12×10^{-27}	0.3099	1.8943	0.1898	2.7965(1)	0.0945
LD	0.0120	…	0.3164	1.9301	0.1904	2.8808(1)	0.0896
ED	0.0060	…	0.8954	4.6593	0.3156	10.7892(1)	0.0010

Table 12 Estimates of the maximum log-likelihood and information criteria for Data set III

Distribution	$-l$	AIC	AICC	BIC	HQIC	CAIC
EpLD	179.123	362.247	362.691	365.049	360.695	362.691
EpD	183.04	370.08	370.524	372.882	368.528	370.524
TPLD	183.04	370.08	370.524	372.882	368.528	370.524
QLD	179.125	362.251	362.695	365.053	360.699	362.695
LD	179.13	360.26	360.403	361.661	360.708	360.403
ED	183.04	368.08	368.223	369.481	368.528	368.223

Table 13 Estimates of the maximum log-likelihood and information criteria for Data set IV

Distribution	$-l$	AIC	AICC	BIC	HQIC	CAIC
EpLD	173.351	350.703	351.147	353.505	349.151	351.147
EpD	183.287	370.575	371.019	373.377	369.023	371.019
TPLD	183.287	370.575	371.019	373.377	369.023	371.019
QLD	174.435	352.869	353.314	355.672	351.318	353.314
LD	174.576	351.152	351.294	352.553	351.6	351.294
ED	183.287	368.575	368.717	369.976	369.023	368.717

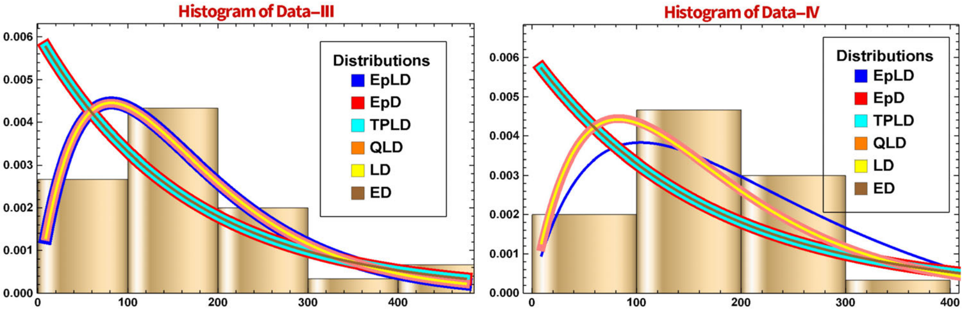


Fig. 9 Data sets III and IV fits via histograms

Table 14 Vuong test statistics for Data sets III and IV

Models	Data set III	Suitability	Data set IV	Suitability
EpLD- EpD	19.4212	EpLD	56.0414	EpLD
EpLD-TPD	19.4213	EpLD	56.0412	EpLD
EpLD-QLD	1954.888	EpLD	8.5923	EpLD
EpLD-LD	-39411	LD	9.8278	EpLD
EpLD-ED	16.7418	EpLD	56.0414	EpLD

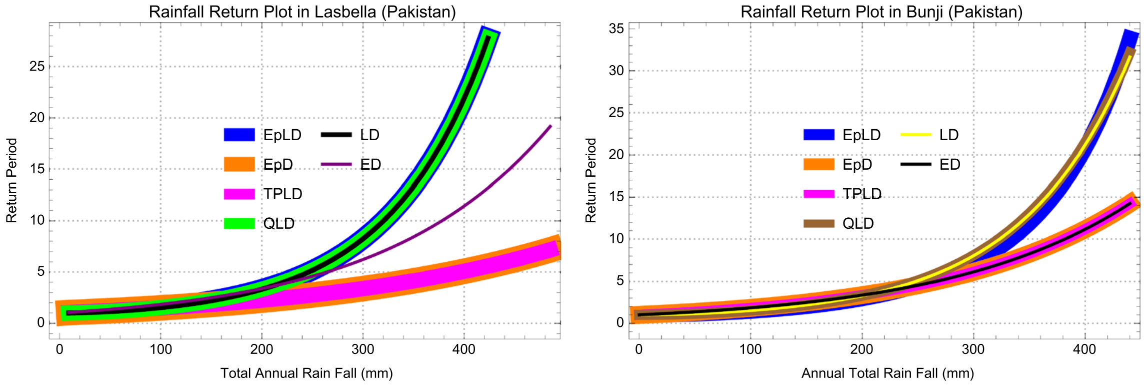


Fig. 10 Return periods of the competing models for Data sets III and IV

series documented at Lasbella (Pakistan) and Bounji (Pakistan) are homogeneous, independent, non-periodic and trend-free. Hence, classical frequency analyses are applied to all of the annual rainfall series. From the above mentioned analysis, we can conjecture that the EpLD is a suitable model for the above mentioned data sets, so we have decided to portray its return period for those interested in environmental data, which is being studied in the coming subsection.

Return period

The average number of years in which an event is predicted to be equalled or exceeded only once is the return period $\mathfrak{T}$ of a particular level. The return period is the reciprocal of the probability of exceeding the threshold in a particular year (see (World Meteorological Organization 2009)). The link between the annual return time and the exceedance probability may be stated as follows if the yearly exceedance probability is designated $1/\mathfrak{T}$. Since the probability of exceedance is $P\left(X > x_{\mathfrak{T}}\right) = 1/\mathfrak{T}$, this implies a return level with a return period of $\mathfrak{T} = 1/\mathfrak{p}$ is a high threshold $x_{\mathfrak{T}}$ whose probability of exceedance is $\mathfrak{p}$.

In this regard, we have found that the EpLD yields a realistic return period that can be visualized from Fig. 10.

Conclusions and future research plans

In this article, we proposed a notable bounded distribution under the name epsilon Lindley distribution (EpLD). Since the Lindley distribution is a limit case of the EpLD, which is a rare property for a bounded support distribution, the EpLD may be treated as a bounded alternative to the Lindley distribution. Therefore, this new distribution provides a flexible solution to the problem of modeling bounded characteristics. We pointed out that the PDF and HRF of the EpLD are very flexible, i.e., they can exhibit various shapes. The fact that the PDF of the EpLD can have a positive skewness and a leptokurtic nature indicates that this new distribution can be used to model the heavy-tailed phenomenon, which is generally common in reliability applications, queuing theory and environmental aspects. We found that the HRF of the EpLD can adopt various shapes from bathtub to increasing failure rate with a left skewed J-shape. All that make this distribution suitable for modeling purposes in a wide range of practical problems. The environmental data analyses are mainly based on the most efficient bounded models, such as the three-parameter lognormal distribution, generalized extreme value type II distribution, generalized extreme value type III distribution, three-parameter gamma distribution, and three-parameter log-Pearson distribution. A significant issue in many common hydrological models is the non-closed form of the CDF. At the same time, the EpLD has only two parameters and has a closed form of its basic functions, including its CDF. This property of the EpLD makes it useful in practical hydrological modeling applications. We should mention that the EpLD is a weighted variant of the EpD from a theoretical standpoint. This weighted variant includes a linear mixture of probability distributions as well as ascertainment biases. In this study, we estimated the parameters of EpLD using the maximum likelihood method. Next, we studied the applications of the proposed distribution both in lifetime and environmental data modeling. Based on the empirical results, we could conclude that the proposed methodology works quite well. In particular, for the annual rainfall data, the EpLD yields a realistic return period when compared with other competing models.

As part of our research activities, we plan to study how the epsilon-Lindley distribution can be utilized in other areas of statistics, including regression analyses and classification modeling.

Acknowledgements The authors are thankful to the Editor-in-Chief and the anonymous referees for their meticulous and thorough reading, which significantly enhanced the readability of this paper.

Funding Open access funding provided by Eötvös Loránd University.

References

Andrews DF, Herzberg AM (1985) Data: a collection of problems from many fields for the student and research worker. Springer, New York

Aarset MV (1987) How to identify a bathtub hazard rate. IEEE Trans Reliab 36(1):106–108

Dombi J, Jónás T, Tóth ZE (2018) The epsilon probability distribution and its application in reliability theory. Acta Polytechnica Hungarica 15(1):216–197

Dombi J, Jónás T, Tóth ZE, Árva G (2019) The omega probability distribution and its applications in reliability theory. Qual Reliab Eng Int 35(2):600–626

Dombi J, Jónás T (2020) On an alternative to four notable distribution functions with applications in engineering and the business sciences. Acta Polytech Hung 17:231–252

Dombi J, Jónás T (2021) Advances in the theory of probabilistic and fuzzy data scientific methods with applications. In: Studies in Computational Intelligence, vol 814. Springer, Berlin/Heidelberg, pp 1–186

Ghitany ME, Atieh B, Nadarajah S (2008) Lindley distribution and its Applications. Mathematical Computation and Simulation 78(4):493–506

Haktanir T, Bajabaa S, Masoud M (2013) Stochastic analyses of maximum daily rainfall series recorded at two stations across the mediterranean sea. Arab J Geosci 6:3943–3958. https://doi.org/10.1007/s12517-012-0652-0

Hussain T, Bakouch HS, Chesneau C (2019) A new probability model with application to heavy-tailed hydrological data. Environ Ecol Stat 26:127-151. https://doi.org/10:1007/s10651-019-00422-7

Lindley DV (1958) Fiducial distributions and Bayes theorem. J R Stat Soc A 20:102–107

Lindley DV (1965) Introduction to Probability and Statistics from a Bayesian Viewpoint Part II: inference. Cambridge University Press, New York

Mazucheli J, Achcar JA (2011) The Lindley distribution applied to competing risks lifetime data. Computer Methods Programs in Biomedicine 104(2):188–192

Murthy DNP, Xie M, Jiang R (2004) Weibull Models. Wiley, New Jersey

Nair U, Sankaran PG, Balakrishnan N (2018) Reliability modelling and analysis in discrete time. Academic Press, https://doi.org/10.1016/C2014-0-01528-6

Okorie IE, Nadarajah S (2019) On the omega probability distribution. Qual Reliab Eng Int 35(6):2045–2050

Phien HN, Ajirajah TJ (1984) Applications of the log Pearson type-3 distribution in hydrology. J Hydrol 73(3-4):359–372. https://doi.org/10.1016/0022-1694(84)90008-8

Shanker R, Mishra A (2013) A quasi Lindley distribution. Biometrics and African Journal of Mathematics and Computer Science Research 6(4):64–71

Shanker R, Hagos F, Sujatha S (2015) On modeling of Lifetimes data using exponential and Lindley distributions. Biometrics & Biostatistics International Journal 2(5):1–9

Vuong QH (1989) Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57(2):307–333

Walpole RE, Myers RH, Myers SL (2012) Probability and Statistics for Engineers and Scientists. Pearson Education, Boston

World Meteorological Organization (2009) Guide to Hydrological Practices Volume II Management of Water Resources and Application of Hydrological Practices WMO-No. 168 6^{th}, Chairperson, Publications Board World Meteorological Organization (WMO) 7 bis, avenue de la Paix P.O. Box 2300 CH-1211 Geneva 2, Switzerland

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Hassan S. Bakouch[1,2] · **Tassaddaq Hussain**[3] · **Christophe Chesneau**[4] · **Tamás Jónás**[5]

Hassan S. Bakouch
hassan.bakouch@science.tanta.edu.eg

Tassaddaq Hussain
tafkho2000@gmail.com

Christophe Chesneau
christophe.chesneau@unicaen.fr

[1] Department of Mathematics, Faculty of Science, Tanta University, Tanta, Egypt

[2] Department of Mathematics, College of Science, Qassim University, Buraydah, Saudi Arabia

[3] Mirpur University of Science and Technology (MUST), Mirpur (AJK), Pakistan

[4] Laboratoire de Mathématiques Nicolas Oresme, Université de Caen, Caen, France

[5] Faculty of Economics, Eötvös Loránd University, Budapest, Hungary