Note on *y*-truncation: a simple approach to generating bounded distributions for environmental applications

Advances in Water Resources, 2007, v.30, p.113-117.

Key words
Bounded distribution; truncated distribution; exponential distribution;
raindrop distribution

W. E. Bardsley
Department of Earth & Ocean Sciences
University of Waikato
Hamilton
New Zealand

web@waikato.ac.nz

Abstract

It may sometimes be desirable to introduce bounds into probability distributions to formalise the presence of upper or lower physical limits to data to which the distribution has been applied. For example, an upper bound in raindrop sizes might be represented by introducing an upper bound to an exponential drop-size distribution. However, the standard method of truncating unbounded probability distributions yields distributions with non-zero probability density at the resulting bounds. In reality it is likely that physical bounding processes in nature increase in intensity as the bound is approached, causing a progressive decline in observation relative frequency to zero at the bound. Truncation below a *y*-axis point is proposed as a simple alternative means of creating more natural truncated probability distributions for application to data of this type. The resulting "*y*-truncated" distributions have similarities with the traditional truncated distributions but probability densities have the desirable feature of always declining to zero at the bounds. In addition, the *y*-truncation approach can also serve in its own right as a means of creating a rich new class of bounded probability distributions when transformations of *y*-truncated distributions are included.

1. Introduction

There will sometimes be a need to impose upper or lower bounds on the *x*-axis of unbounded standard probability distributions to better represent recorded variables which are constrained in their magnitude range because of some physical bounding process. For example, exponential distributions can give reasonable first approximation to histograms of raindrop diameters [6][8], but a truncated exponential distribution might be utilised to allow for the physical reality of an upper limit to raindrop size while still maintaining the approximate exponential form. Distribution bounding in this way is distinct from bounding resulting from censored observations which cannot be recorded beyond a certain magnitude.

The familiar text-book truncated distributions are referred to in this paper as "*x*-truncated" distributions and are natural choices for representing data with censoring beyond threshold values [4]. However, their value for representing situations with a physical upper or lower bound is more questionable because the mathematical truncation process creates finite probability density at the bounds. In reality it would be more likely that environmental bounding processes such as raindrop break-up or maximum storm precipitation limitation will progressively increase in intensity as the bound is approached, giving rise to observation frequencies which decline progressively to zero at the bound.

One approach would be to discard the original unbounded distribution altogether and replace it with an existing flexible bounded distribution such as the beta distribution which include unimodal forms declining to zero at the bounds. It is likely, however, that the original unbounded distribution would have had some history of successful data fitting in the field and the users would prefer to modify this distribution so as to

incorporate an upper or lower bound. For this situation, it would be helpful to have an alternative truncation procedure available such that the resulting bounded distributions possess probability density declining to zero as the bounds are approached. This brief paper describes a particularly simple "*y*-truncation" process which achieves this end.

2. Definitions

A "*y*-truncated distribution" is defined here to be that portion of the original distribution which is above a given horizontal line drawn parallel to the *x*-axis, with rescaling to an integral of 1.0. There is a connection between *x*- and *y*-truncated distributions in that any *y*-truncated distribution also defines a corresponding *x*-truncated distribution which can be thought of as a two-component finite mixture distribution comprised of the *y*-truncated distribution and a rectangular distribution defined over the same *x*-axis interval. The *y*-mode of truncation is a purely geometrical contrivance and is somewhat analogous to the familiar hydrograph quickflow separation concept, although for the purposes of this paper *y*-truncated distributions are defined as always arising from horizontal separation lines.

It is assumed $w(x)$ is a probability density function (pdf) subject to *y*-truncation by a horizontal line corresponding to some value of probability density $0 < y_* < y_{\max}$, where $y_{\max}$ is the maximum value of $w(x)$. Two general $w(x)$ forms will be considered here: form (i) is defined such that $w'(x) < 0$ for distributions defined over $0 \le x \le \infty$, and form (ii) represents unimodal $w(x)$ types. For example, (i) includes the exponential distribution and (ii) incorporates a wide variety of standard distributions including the normal, lognormal, Gumbel, and inverse Gaussian distributions.

With respect to (i), introducing an upper bound at $x = \xi$ gives the general definition of the cumulative distribution function (cdf) of this class of *y*-truncated distributions as

$$F(x) = \frac{W(x) - w(\xi)x}{W(\xi) - w(\xi)\xi} \qquad 0 \le x \le \xi \qquad (1)$$

where $W(x)$ is the cdf corresponding to the probability density function $w(x)$. If $F(x)$ in (1) is mapped to the interval *A,B* then this gives:

$$F(x) = \frac{W(v_x) - w(\xi)v_x}{W(\xi) - w(\xi)\xi} \qquad v_x = \xi\left(\frac{x - A}{B - A}\right) \qquad A \le x \le B \qquad (2)$$

and the associated *y*-truncated pdf is given by

$$f(x) = \frac{\xi \; w(v_x) - w(\xi)}{(B - A) \; W(\xi) - w(\xi)\xi} \qquad (3)$$

For the unimodal case (ii), a horizontal *y*-truncation line through $w(x)$ now defines two bounding values $\xi_1 < \xi_2$ on the *x* axis such that $w(\xi_1) = w(\xi_2)$ and ξ_1 and ξ_2 are respectively to the left and right of the $w(x)$ mode. Mapping the resulting *y*-truncated distribution to some *A,B* interval then gives the *y*-truncated cdf and pdf respectively as:

$$G(x) = \frac{W(k_x) - W(\xi_1) - w(\xi_1)(k_x - \xi_1)}{W(\xi_2) - W(\xi_1) - w(\xi_1)(\xi_2 - \xi_1)} \qquad A \le x \le B \tag{4}$$

$$g(x) = \frac{[w(k_x) - w(\xi_1)](\xi_2 - \xi_1)/(B - A)}{W(\xi_2) - W(\xi_1) - w(\xi_1)(\xi_2 - \xi_1)} \tag{5}$$

where $k_x = \xi_1 + (\xi_2 - \xi_1)(x - A)/(B - A)$.

The *r*th moment about zero μ_r' of any *y*-truncated distribution can be conveniently represented as a function of the equivalent moments of a rectangular distribution and the corresponding *x*-truncated distribution possessing the same bounding points ξ_1 , ξ_2:

$$\mu_r' = \frac{\tilde{\mu}_r' \, [W(\xi_2) - W(\xi_1)] - w(\xi_2)(\xi_2 - \xi_1)\breve{\mu}_r'}{W(\xi_2) - W(\xi_1) - w(\xi_2)(\xi_2 - \xi_1)} \tag{6}$$

where $\tilde{\mu}_r'$ is the *r*th moment about zero of the corresponding *x*-truncated distribution and $\breve{\mu}_r'$ is the *r*th moment about zero of a rectangular distribution defined over ξ_1 , ξ_2. The lower bound ξ_1 is set to zero for the specific case $w'(x) < 0$ over the region $0 \le x \le \infty$. Expressions for the moments of a number of *x*-truncated distributions are summarised in [3] and could be incorporated in (6) as $\tilde{\mu}_r'$ to provide expressions for moments about zero for a range of specific *y*-truncated distributions. Central moment expressions of *y*-truncated distributions can also be obtained by way of (6) using standard relations linking absolute and central moments.

3. Example

The *y*-truncation process and an estimation example is given here with respect to the *y*-truncated exponential distribution with an upper bound at some point $x = \xi$. From (1) and (3) the cdf and pdf of the *y*-truncated exponential distribution are obtained respectively as:

$$F(x) = \frac{1 - \exp(-x/\alpha) - \exp(-\xi/\alpha)\,(x/\alpha)}{1 - \exp(-\xi/\alpha)(1 + \xi/\alpha)} \qquad \alpha > 0, \quad 0 \le x \le \xi \tag{7}$$

$$f(x) = \frac{[\exp(-x/\alpha) - \exp(-\xi/\alpha)]/\alpha}{1 - \exp(-\xi/\alpha)\,(1 + \xi/\alpha)} \qquad (8)$$

where α is the mean of the untruncated exponential distribution.

Some example plots of (8) for are shown in Fig. 1 for ξ=1 and various values of α, illustrating decline of $f(x)$ to zero at $f(\xi)$. Both *x*- and *y*-truncated exponential distributions can have similar properties even though they differ in near-boundary form. For example, the hazard function $h(x)$ of both distribution types is of similar form (Fig.2) , with $h(x)$ increasing slowly at first and then more rapidly toward infinity as the upper bound is approached.

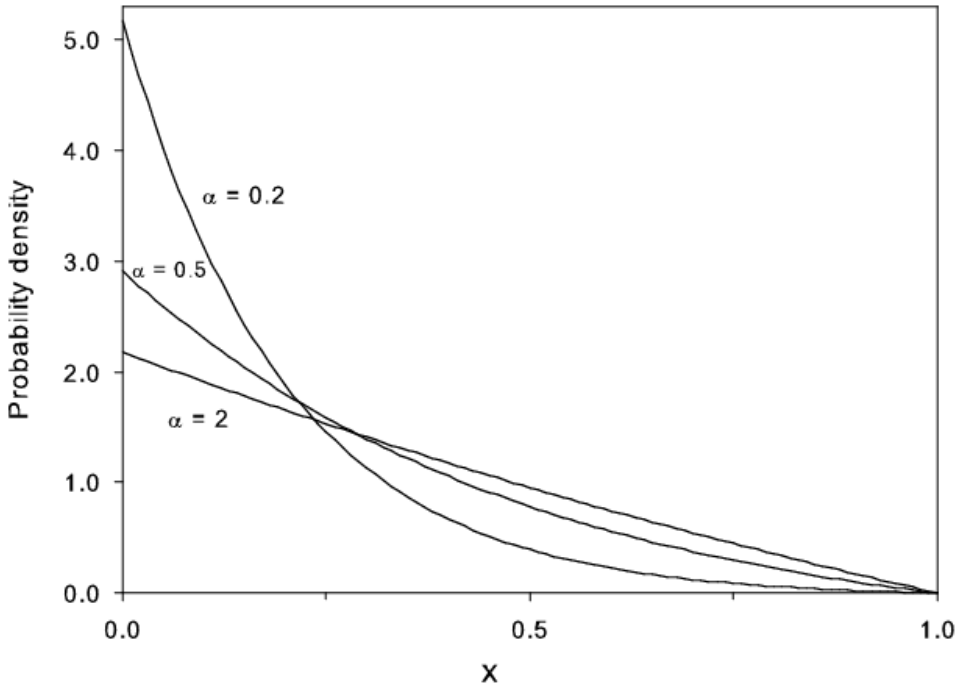

Fig. 1. Selected *y*-truncated exponential distributions for ξ=1 and various values of α.

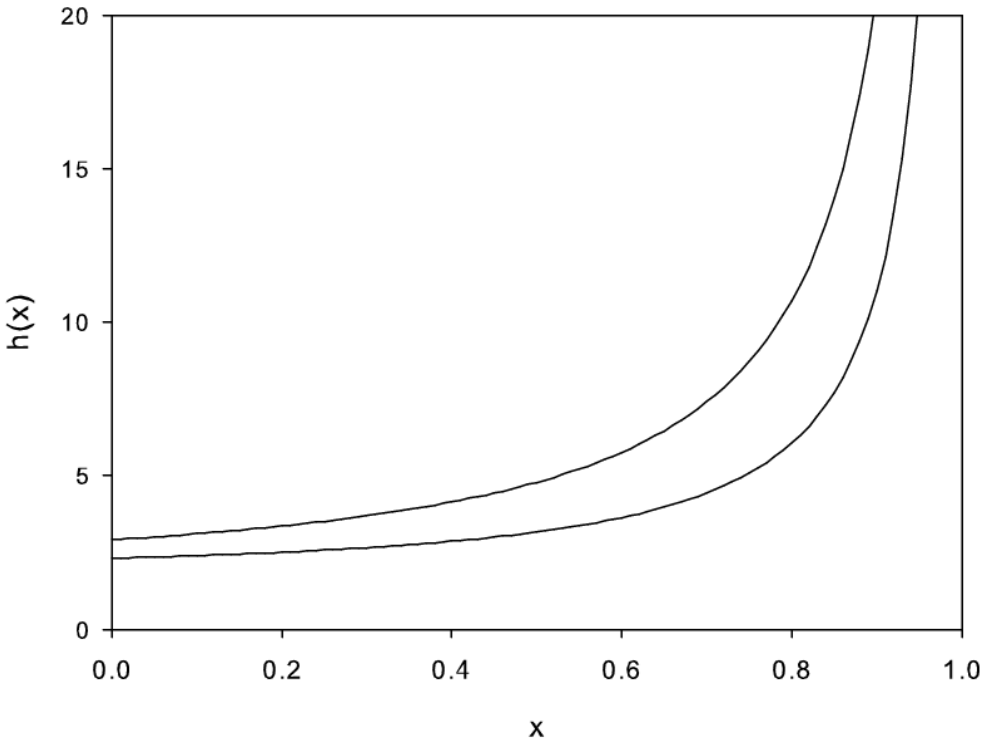

Fig. 2. Comparison of hazard functions $h(x)$ for *y*- and *x*-truncated exponential distributions (upper and lower plots respectively). Both truncated distributions are derived from truncation of an exponential distribution with mean value α =0.5 and with upper truncation at *x* = ξ = 1.

Substituting for $\tilde{\mu}_r'$ and $\breve{\mu}_r'$ in (6) yields the *y*-truncated exponential distribution expressions for μ_1' and μ_2' respectively as:

$$\mu_1' = [\alpha - \exp(-\xi/\alpha)(\xi + \alpha)]/[1 - \exp(-\xi/\alpha)] \tag{9}$$

$$\mu_2' = [2\alpha^2 - \exp(-\xi/\alpha)(\xi^2 + 2\xi\alpha + 2\alpha^2)]/[1 - \exp(-\xi/\alpha)] \tag{10}$$

which also defines the distribution variance through $\mu_2 = \mu_2' - \mu_1'^2$.

The expressions (9) and (10) are not helpful for yielding simple moment estimators for *y*-truncated exponential distributions and bounded distributions generally are awkward to deal with when considering optimal estimators applicable to small samples. It might happen that some suitable modification could be applied to methodologies developed for parameter estimation for *x*-truncated exponential distributions [2], although investigating formal estimation procedures is beyond the scope of this paper. However, some environmental recordings such as raindrop frequencies will generate large samples which allow application of more informal estimation techniques.

Assuming a large sample, the estimation approach utilised here is to first obtain $\hat{\xi}$ as distribution-free estimate of ξ and then estimate α by maximising the liklihood function with ξ fixed at $\hat{\xi}$. This estimation procedure was applied to 1000 "data" values which might be called raindrop diameters but were generated as random numbers from (7) with $\alpha = 1$ and $\xi = 3,$ with the simulations being achieved though generating random $F(x)$ values from the standard rectangular distribution coupled with numerical solution to obtain each *x*.

The Robson-Whitlock bound estimator [5] was utilised to estimate ξ as:

$$\hat{\xi} = 2X_N - X_{N-1} \tag{11}$$

where X_N and X_{N-1} are respectively the largest and second largest recorded observation in the sample. An alternative bound estimator such as [7] could have been employed to give a reduced estimation error at the expense of some increase in complexity.

For the simulated sample X_N and X_{N-1} were respectively 2.90 and 2.86, giving $\hat{\xi}$ =2.94 from (11). With ξ fixed at 2.94, $\hat{\alpha}$ was obtained as 1.20 through maximising the likelihood function – in this case using the Excel solver. Fig. 3 shows the actual *y*-truncated exponential pdf from which data were simulated together with the estimated pdf as obtained from inserting the $\hat{\xi}, \hat{\alpha}$ estimates into (8). Fig 4 shows an expansion of the lower part of Fig. 3, illustrating that a simple exponential distribution fitted to the data gives a poor fit near ξ because of the bounded nature of the data. Fitting was via equating the exponential α parameter to the sample mean. The simulated data set is listed in an Excel spreadsheet as an electronic attachment to this paper.

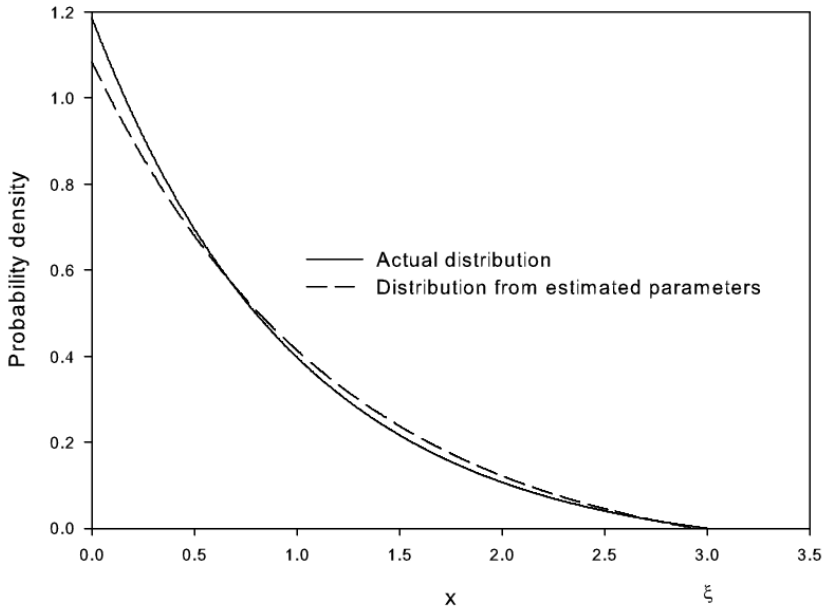

Fig.3. The *y*-truncated exponential distribution (α =1, ξ=3) used to simulate 1000 random values, together with the distribution defined by the $\hat{\alpha}$, $\hat{\xi}$ estimates from the sample.

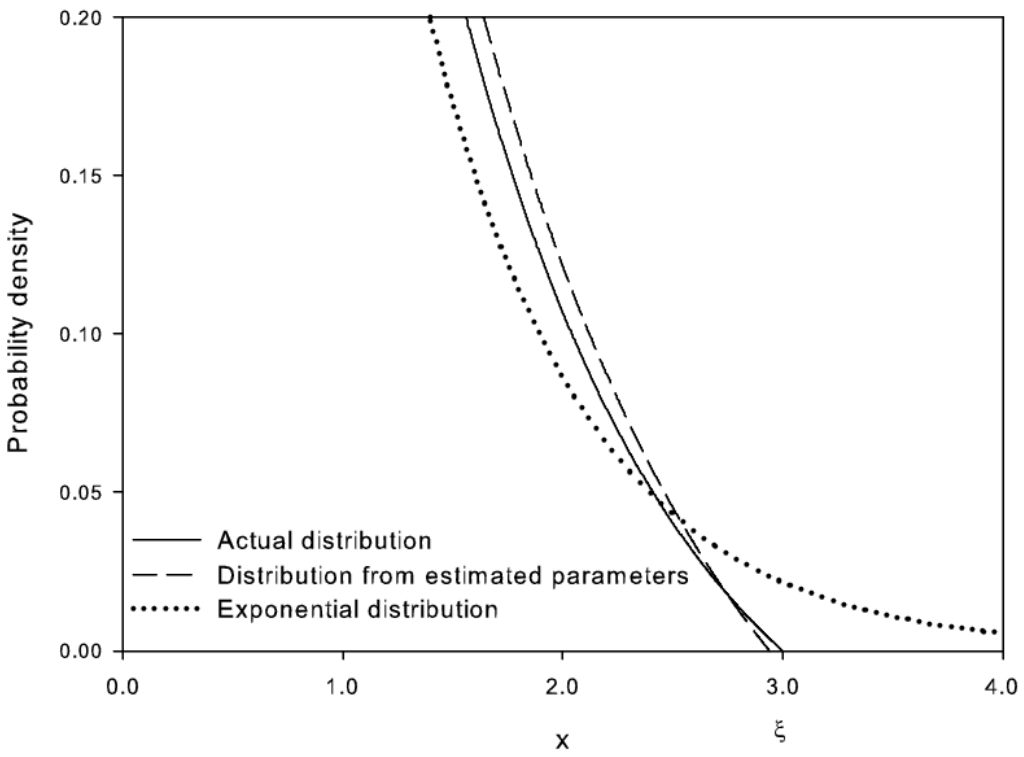

Fig. 4. Expansion of the lower portion of Fig. 3, also showing part of an exponential distribution with the same mean value as the simulated sample.

The reasonably close correspondence between the actual and estimated distributions in Fig. 3 is aided by the dual advantages of large sample size and the fact that the data were generated exactly from a *y*-truncated exponential distribution. However, two-stage estimation methods should be applicable to any large real-world data sample provided the recorded values are not too far removed from some *y*-truncated exponential distribution form.

Some properties of the *y*-truncated exponential distribution may find application in areas outside of environmental science. In particular, the *x*-truncated exponential distribution has been argued as an alternative for describing component life testing situations because the increasing hazard function is seen as more realistic than the constant $h(x)$ of the exponential distribution [2]. However, if the failure process is such that upper time bound is never reached in practice then the *y*-truncated exponential distribution would be the more appropriate alternative to the exponential distribution.

4. Discussion and conclusion

The main attribute of *y*-truncated distributions is their more natural appearance for environmental applications in comparison with the traditional *x*-truncated distributions. There is, however, a need for further work to develop estimation theory up to a comparable level with *x*-truncated distributions and bounded distributions in general.

An additional more general aspect of the *y*-truncation process is that it generates a rich new class of bounded probability distributions when transformation of variables is taken into account. For example, a power transformation applied to random variables from a *y*-truncated exponential distribution yields a bounded distribution with respective cdf and pdf given respectively by:

$$F(x) = \frac{1-\exp[-(x/\alpha)^c] - \exp[-(B/\alpha)^c](x/\alpha)^c}{1-\exp[-(B/\alpha)^c][1+(B/\alpha)^c]} \qquad \alpha, c > 0, \quad 0 < x \le B \qquad (12)$$

$$f(x) = \frac{\{\exp[-(x/\alpha)^c] - \exp[-(B/\alpha)^c]\}(x/\alpha)^{c-1} c/\alpha}{1-\exp[-(B/\alpha)^c][1+(B/\alpha)^c]} \qquad (13)$$

where c and $\gamma = B/\alpha$ are shape parameters and the *y*-truncated exponential distribution corresponds to the special case of $c = 1$.

The pdf (13) is unimodal for $c > 1$ and some of the unimodal forms have similarities with some beta distribution unimodal forms (Fig. 5), suggesting that (13) might sometimes be suitable as an alternative to the beta distribution. One possible application here would be to multiply (13) by a scale factor and use this simply as an empirical function to describe quickflow hydrographs as an alternative to similar application of the beta distribution. [1][9]. The use of (13) in this way may or may not offer an improvement in representation of hydrograph form over the beta distribution but does have the convenience of avoiding the need to import a beta function macro for hydrograph evaluation in a standard Excel spreadsheet.

Many other bounded probability density functions could be derived through transformations of different *y*-truncated distributions and this opens an interesting area for further work in distribution development. The utility value of these new distributions will ultimately depend on the extent to which workable estimation methodologies can be developed.

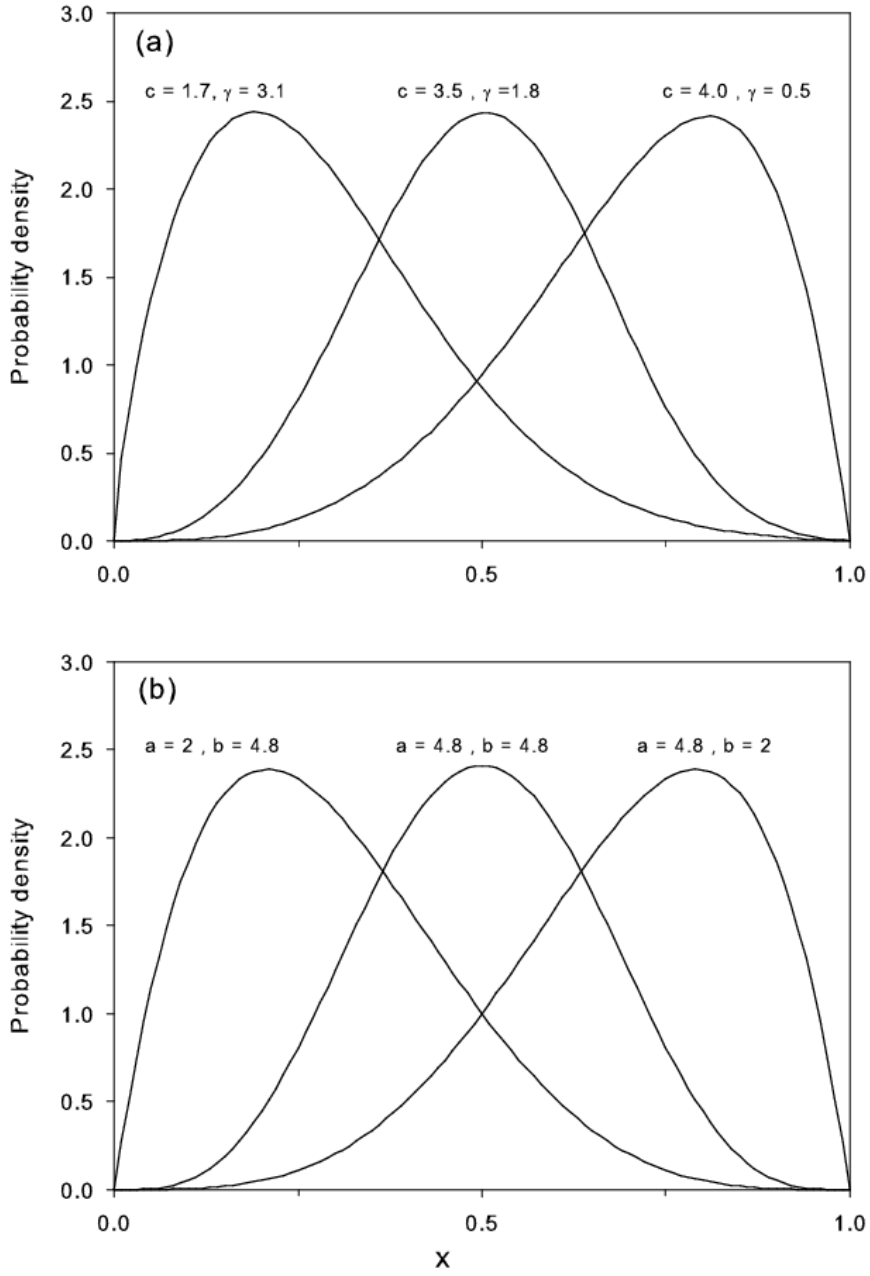

Fig. 5. Unimodal distributions on the 0,1 interval: (a) some unimodal forms of (13) for different shape parameter combinations; (b) some unimodal beta distribution forms for selected shape parameters. The beta distribution is defined here as $f(x) = B(a,b)^{-1} x^{a-1}(1-x)^{b-1}$ where a and b are shape parameters, and $B(.)$ is the beta function.

References

[1] Bhunya PK, Mishra SK, Ojha CSP, Berndtsson R. Parameter estimation of beta distribution for unit hydrograph derivation. *J Hydrol Eng* 2004;9: 325-332.

[2] Hannon PM, Ram CD. Estimation of parameters for the truncated exponential distribution. *Commun Statist - Theory Meth* 1999;28: 2591-2612.

[3] Jawitz JW. Moments of truncated continuous univariate distributions. *Adv Water Resour* 2004;27: 269-281.

[4] Loáiciga HA, Michaelsen J, Hudak PF. Truncated distributions in hydrologic analysis. *Water Resour Bulletin* 1992;28: 853-863.

[5] Robson DS, Whitlock JH. Estimation of a truncation point. *Biometrika* 1964;51: 33-39.

[6] Smith PL. Raindrop size distributions: exponential or gamma - does the difference matter? *J App Meteor* 2003;42: 1031-1034.

[7] Van Der Wat PA. Note on estimation of bounds of random variables. *Biometrika* 1980;67 713-714.

[8] Uijlenhoet R, Stricker JNM. A consistent rainfall parameterization based on the exponential raindrop size distribution. *J Hydrol* 1999;218: 101-127.

[9] Yue S, Ouarda TBMJ, Bobée B, Legendre P, Bruneau P. Approach for describing statistical properties of flood hydrograph. *J Hydrol Eng* 2002;7: 147-153.