SA

SA 00 (2025) 1–4

The affine growth of river heights

Z.W.T. Mason

Sheffield, UK

Abstract

River heights form a skewed distribution which is variously fit by the log-normal, Gamma, Generalized Extreme Value, Weibull and Pareto distributions. For Bernoulli trials it has been recently shown affine returns may be approximated by an appropriately scaled logit-normal distribution. A good fit for some of the river heights is performed by this latter distribution whose parameters are derived using a mixture of Maximum Likelihood Estimation and a grid search.

1. Introduction

River height data is influenced by factors like rainfall, snowmelt and animals such as beavers and humans. Different distributions are used depending on whether the user is modelling typical levels, extremes or floods. The log-normal distribution is used for daily river height measurements during non-extreme conditions. The Gamma for seasonal river heights with consistent variability. The Generalized Extreme Value (GEV) for extreme river heights, such as annual maximum water levels or flood peaks. The Weibull for river heights with moderate skewness. The Pareto Distribution for the upper tail of river height data, particularly for extreme flood events where a few rare events dominate.

By way of example, if there were some upper scale applicable then the logit-normal distribution may be appropriate. The logit-normal distribution is much overlooked but important, for example it moderates the growth of a well-mixed epidemic from exponential to logistic as shown by the author [1].

The author [2] also showed, for stock prices, that affine returns on coin tosses can be approximated by a stretched-out logit-normal distribution for the case when the growth converges to a finite support. The same shapes were also evident in the case where the support grows exponentially with t when $\beta > 1$ in equation 1.

$$\left(0, \delta\frac{1-\beta^t}{1-\beta}\right) \tag{1}$$

The process itself being described by the random variable S_t growing from $S_0 = 0$ with probability p

$$S_{t+1} = \beta S_t + \delta,\ 0 < \beta \neq 1,\ \delta > 0 \tag{2}$$

and shrinking with probability $q = 1 - p$

$$S_{t+1} = \alpha S_t,\ 0 < \alpha < 1 \tag{3}$$

When $\beta < 1$ the probability distribution converges. Applying the equations to the distribution leads to shrinks and shifts which, at the tails, means only one copy is applied. The resulting equations then show that both tails are power laws. i.e. The logit-normal distribution may be appropriate for the extreme heights. In fact, it should be expected

that only the one distribution should cover all the natural cases rather than the mish-mash of the existing approach. Whether the model is appropriate though hinges on whether it fits the data.

2. Parameter estimation

The Maximum Likelihood Estimators for the log-normal and logit-normal distributions are the same as for the normal distribution fitted to the values transformed by the appropriate function, so the fitting is done in a highly distorted space. The probability density function for the scaled logit-normal is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \frac{L}{x(L-x)} \exp\left(-\frac{(\log(x) - \log(L-x) - \mu)^2}{2\sigma^2}\right) \tag{4}$$

Taking logarithms

$$-\log(\sigma) - \frac{1}{2}\log(2\pi) + \log(L) - \log(x) - \log(L-x) - \frac{(\log(x) - \log(L-x) - \mu)^2}{2\sigma^2} \tag{5}$$

Differentiating with respect to μ, σ, L results in, respectively

$$\frac{\log(x) - \log(L-x) - \mu}{\sigma^2} \tag{6}$$

$$-\frac{1}{\sigma} + \frac{(\log(x) - \log(L-x) - \mu)^2}{\sigma^3} \tag{7}$$

$$\frac{1}{L} - \frac{1}{L-x} + \frac{\log(x) - \log(L-x) - \mu}{\sigma^2} \frac{1}{L-x} \tag{8}$$

Resulting in the usual, if L were equal to one, estimates of

$$\mu\Sigma_i 1 = \Sigma_i (\log(x_i) - \log(L - x_i)) \tag{9}$$

$$\sigma^2\Sigma_i 1 = \Sigma_i (\log(x_i) - \log(L - x_i) - \mu)^2 \tag{10}$$

The final derivative was not summed over the data and used to estimate L but was used to estimate the error. The mean square error between the formula and the histogram was also calculated. Finally, the result was checked to see whether it could be improved upon by searching through a grid of parameters.

3. Data analysis

3.1. Data for the Kanawha River, West Virginia

Fourteen years of water gauge height, from 2008/1/14 to 2022/1/11, for the Kanawha River in West Virginia were downloaded from Kaggle [3].

The Maximum Likelihood Estimation was performed with L from 6.3 to 15 metres with the mean square error decreasing to monotonically. Subsequently a matrix of tests was performed with a minimum at $L = 15.7\,\mathrm{m}$, $\mu = -2.28$, $\sigma^2 = 0.2555$ is shown in figure 1.

3.2. Data for the Pembina River, Alberta

One year of water gauge height, from 2024/5/31 to 2025/5/31, for the Pembina River near Entwistle were downloaded from the government of Alberta [4].

The Maximum Likelihood Estimation was performed with L from 2.9 to 10 metres with the mean square error decreasing monotonically. Subsequently a matrix of tests was performed with a minimum at $L = 9.93\,\mathrm{m}$, $\mu = -1.42$, $\sigma^2 = 0.01098$ which is shown in figure 2.

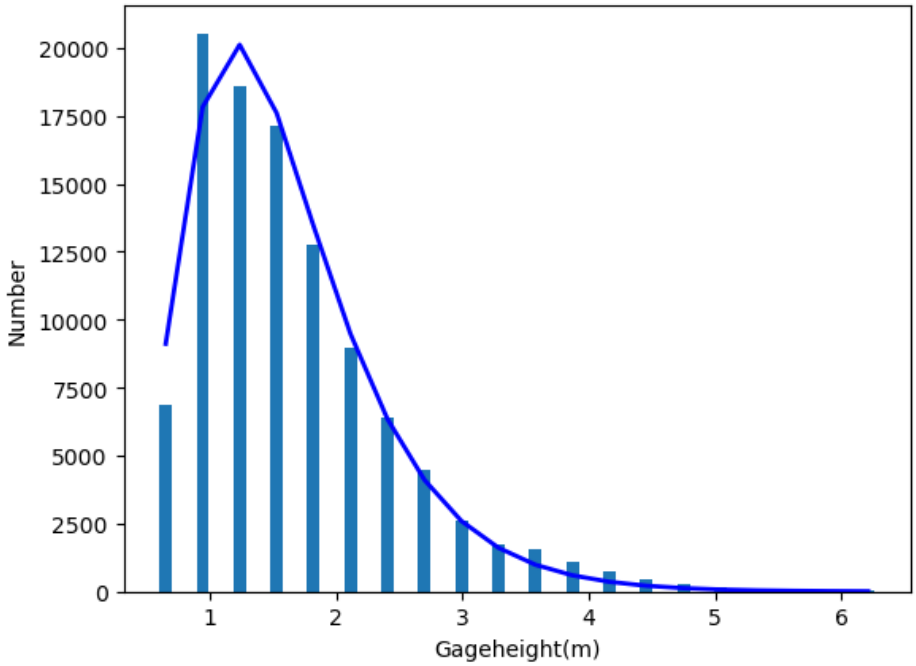

Figure 1. Fit of Kanawha River gauge heights

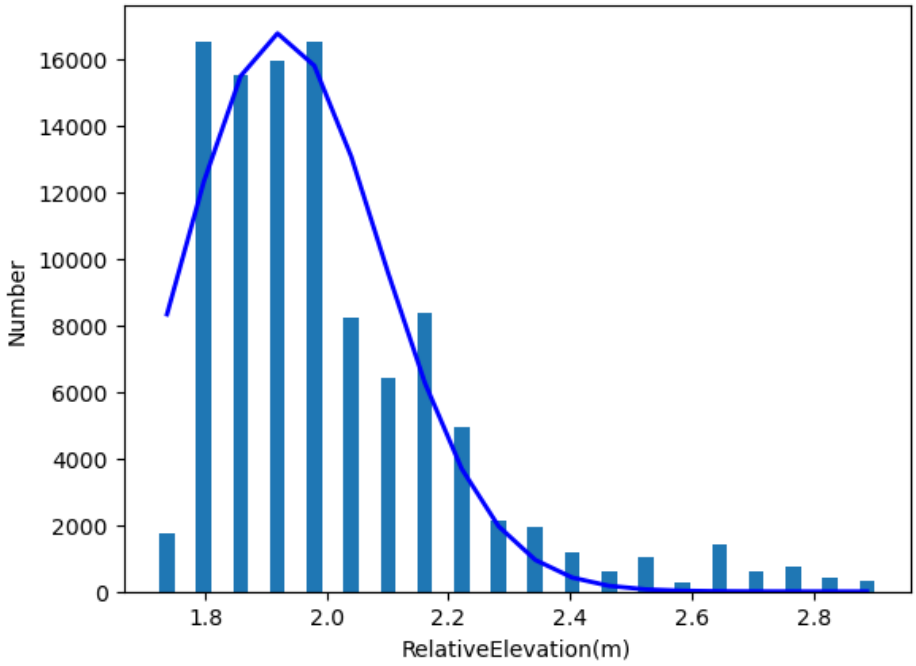

Figure 2. Fit of Pembina River gauge heights

3.3. Data for the River Don at Fishlake, South Yorkshire

Twelve years of water gauge height, from 2012/11/26 to 2025/5/30, for the River Don at Fishlake, South Yorkshire were downloaded from RiverLevels.uk [5]. This includes the flooding of 2019.

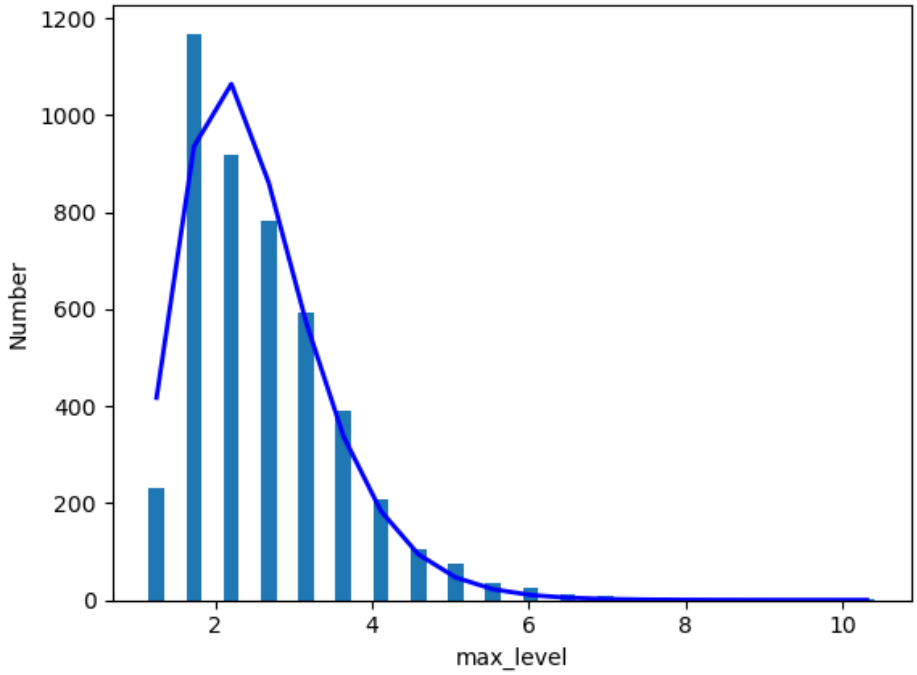

Figure 3. Fit of River Don maximum gauge heights

The Maximum Likelihood Estimation was performed with L from 10.35 to 20 metres with the mean square error decreasing monotonically. Subsequently a matrix of tests was performed with a minimum at $L = 24.64\,\text{m}$, $\mu = -2.24$, $\sigma^2 = 0.16068$ which is shown in figure 3.

4. Discussion

That the Maximum Likelihood Estimation wasn't sufficient to determine the parameters may be due to the support being semi-infinite. i.e. The length scale tends to infinity. A rough grid search, although ugly and a misuse of cheap computing power, seemed to be sufficient.

The fits for the Kanawha and the Don seem reasonable, for the Pembina less so. Obviously more data needs to be fitted and comparison needs to be made with the other distributions. However, using a scaled logit-normal distribution does appear to be a promising option.

References

[1] Z. Mason, Naturally extending the standard sir model to stochastic growth, TBDdoi:10.31219/osf.io/y6ckv.
[2] Z. Mason, Affine returns on bernoulli trials in finance, TBD.
[3] J. A. Jewel, 3 years water gage height data kanawha river wv, https://www.kaggle.com/datasets/jewelshaikh/3years-water-gage-height-data-kan (2023).
[4] Pembina river near entwistle - wsc, https://rivers.alberta.ca/ (2025).
[5] River don at fishlake, https://riverlevels.uk/river-don-fishlake (2025).