Published by
NIGERIAN SOCIETY OF PHYSICAL SCIENCES
Available online @ https://journal.nsps.org.ng/index.php/jnsps

J. Nig. Soc. Phys. Sci. 6 (2024) 2077

Journal of the Nigerian Society of Physical Sciences

Assessing model selection techniques for distributions use in hydrological extremes in the presence of trimming and subsampling

Sunday Samuel Bako[a,b], Norhaslinda Ali[c,*], Jayanthi Arasan[c]

[a]*Institute for Mathematical Research, Universiti Putra Malaysia, Selangor, Malaysia*
[b]*Department of Mathematical Sciences, Kaduna State University, Kaduna, Nigeria*
[c]*Department of Mathematics and Statistics, Universiti Putra Malaysia, Selangor, Malaysia*

Abstract

A suitable probability distribution is required to quantify and estimate hydraulic structure design for risk evaluation and management. The inability of model selection criteria to differentiate, in some cases, among candidate distributions used in the analysis of hydrological extremes is often criticised. This study verifies, with the aid of model selection techniques, the potential utility of trimming and subsampling in distinguishing between candidate distributions, which might not be feasible using the traditional goodness of fit method alone, when samples available are small. The performance of the proposed method is evaluated through its application to real and simulated yearly peak rainfall datasets. The proposed approach is then compared with several standard model selection techniques. Results show that the model selection techniques with the aid of subsampling are effective in identifying the true parent distribution for the untrimmed samples given a two-parameter distribution; contrarily, they are inefficient where a distribution with a three-parameter is the parent distribution. However, as trimming is introduced, all model selection methods recognise the true parent distribution for a three-parameter distribution. Overall, utilising trimming and subsampling with the aid of model selection methods yields promising outcomes in the analysis of hydrological extreme frequencies. Drawing from the results of numerical simulation and examination of observed data, the use of trimming and subsampling can be a viable tool in differentiating among candidate distributions used in the investigation of hydrological extreme frequencies.

DOI:10.46481/jnsps.2024.2077

Keywords: Flood frequency analysis, Model selection, Trimming, Subsampling

Article History :
Received: 16 April 2024
Received in revised form: 16 August 2024
Accepted for publication: 24 August 2024
Published: 02 September 2024

Communicated by: P. Thakur

1. Introduction

Flooding, a global natural phenomenon that affects humans negatively, requires different strategies to combat. Some of these strategies include the construction of culverts, downstream barricades, flood projections for awareness and possible evacuation, and effective land management in terms of streamflow characteristics, Kidson and Richards [1]. Considerably, flood frequency analysis can provide risk evaluation for these strategies. Wallis [2] defined Flood Frequency Analysis (FFA) as a statistical approach used to identify the fundamental prob-

*Corresponding author Tel. No: +60-126-910-694.
Email address: `norhaslinda@upm.edu.my` (Norhaslinda Ali)

ability distribution that generates observed floods and models the exceedance probability of a return period of statistical significance. Flooding is generally set off due to different combinations of meteorological features, and these give rise to various underlying mechanisms generating floods. Yan *et al.* [3] classified flood as flood triggered by excessive rainfall, flood resulting from rain on snow, and flood arising from snowmelt. Floods as a result of snowmelt are usually of long duration, while floods triggered by rainfall are usually of short duration, Ref. [4]. Though streamflow data also provides reliable estimates for the analysis of floods, it is common practice in hydrological studies to use rainfall observations in modelling flood design associated with a specified recurrence interval as they are readily available in space and time, Flamini *et al.* [5]. The primary objective of flood frequency analysis is to demonstrate the correlation between the magnitude of hydrological extremes and their rate of occurrence using a probability distribution, Refs. [6, 7]. Hydraulic structure design and the handling of river basin administration rely on the assessment of factors that triggered flooding, such as excessive rainfall or peak streamflow, ideally acquired from long historical observations. Notwithstanding, the scant availability of hydrometeorological monitoring devices, prevalent in less economically developed countries, and the necessity for estimation of return values linked to increasing recurrence intervals, resulted in the use of statistical-based techniques such as the at-site analysis of floods. This study examines how well candidate probability distributions can describe observed floods caused by heavy rainfall.

In practical application, the fundamental probability distribution that generates an observed flood at a site or region is unknown, Hamed and Rao [7]. Identifying and effectively discriminating among these probability distributions is an existing challenge in hydrological extremes. Methods related to parameter estimations are well-established in the literature and give good outcomes for different probability distributions. However, this is not the case with techniques for selecting a suitable underlying distribution, Laio *et al.* [8]. The World Meteorological Organization (WMO) operational hydrology reports by Cunnane [9] contain several kinds of probability distributions examined in separate circumstances around the world. Graphical and statistical testing procedures are methods used in selecting probability distribution in the analysis of flood frequency. The ineffectiveness of model selection criteria to differentiate between distributions for similar applications in some cases and the subjective nature of the graphical methods are often criticised, Onoz and Bayazit [10]. Therefore, choosing an appropriate probability distribution is still an existing challenge. Other approaches are needed to discriminate further among distributions, especially in situations where this discrimination is difficult between probability distributions. The primary design behind this work is to identify the probability distribution nearer to the distribution that is believed to generate the observed data, often referred to as the parent distribution in the context of flood frequency analysis.

Many studies have been conducted previously on using goodness-of-fit tests for comparison of various probability distributions used in flood frequency analysis, a few are discussed herein. Langat *et al.* [11] describe current approaches such as the goodness-of-fit tests used in identifying the most suitable probability functions most appropriate for calculating the highest, lowest, and average stream flows of daily discharges. Lognormal and generalised extreme value distribution functions were a suitable fit for average annual stream flows. The study suggests using the selected probability distribution model to try other methods for choosing distributions in frequency analysis, such as theoretical evaluation. This can help confirm the appropriateness of the selected distributions. Das [12] employed a subsampling scheme in examining flood frequency. The study investigates the sensitivity of annual maximum flood data to subsampling in selecting the best probability distribution. The study examined a situation with a large available dataset by using six probability distributions and the Anderson-Darling (AD) test in the analysis. The suggested method shows plausible prospects for large samples. Chen *et al.* [13], showed a technique for selecting best-fit flood frequency distributions using combined criteria. Eight distributions and five selection criteria; Kolmogorov Smirnov test, Anderson-Darling criterion (ADC), Akaike Information Criterion (AIC), Akaike Information Criterion-corrected (AICc), and Bayesian Information Criterion (BIC)-were adopted. The study demonstrated that using the composite criterion improves performance. The objective of Hassan *et al.* [14] is to find a suitable model amidst five probabililty distributions for the yearly highest peak flow data from various sites. The AD technique was adopted for model selection. The Pearson generalised logistic distributions show a good fit. See Refs. [10, 15–19] for more on the suitability of distributions for analysis of flood frequency in hydrological extremes.

This study aims to verify using the goodness of fit technique the potential effectiveness of trimming and subsampling in discriminating among probability distributions used in hydrological extremes for at-site frequency analysis when the discrimination between statistical distributions is difficult and the available samples are small, which are common situations experience in the analysis of hydrological extremes. The purpose of trimming is to reduce the undesirable influence that lower observations may have on the tail behaviour at the high end of a distribution. The rationale for utilising subsampling for selecting an optimal distribution lies in the fact that the subsamples represent samples of smaller sizes than the original sample. Thus, drawing a reasonable number of subsamples from the original samples is expected to yield a pool of samples that will enhance the identification of an optimal distribution for a gauged site. This study evaluates the performance of the proposed methodology by applying it to simulated and real annual peak rainfall datasets. The technique utilised in this study and the findings to be derived will not only give an important complement to model selection methods but will also assist in decision-making for minimizing uncertainty in flood design estimation, thereby helping in hazard cushioning for hydraulic structures planning, design, and management.

Table 1: Model selection criteria.

Method	Test statistic
AIC	$-2\ln[L(D\|\hat{\theta}) + 2m]$
AICc	$-2\ln[L(D\|\hat{\theta}) + 2m(\frac{n}{n-m-1})]$
BIC	$-2\ln[L(D\|\hat{\theta}) + \ln(n)m]$
AD	$-n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)[\ln F(y_i) + \ln 1 - F(y_{(n-i+1)})]$

2. Methods

The model selection problem in flood frequency analysis may be stated as follows: Suppose there exists a set of n observations, $R = \{y_1, y_2, ..., y_n\}$, sorted in increasing sequence. In practical situations, researchers typically cannot identify the underlying parent distribution $f(y)$ of the random variable R. Let $K_j, j = 1, 2, ..., N_k$, be N_k operating probability distributions models that can take the form $K_j = g_j(y, \hat{\theta})$ utilising parameters $\hat{\theta}$ derived from the provided data set R. The parameter $\hat{\theta}$ describes the likelihood distribution of y. The model selection objective is to determine the optimal operating model K_j that best approximates the parent distribution $f(y)$.

Trimming and a subsampling scheme can be viable tools for flood frequency analysis. Trimming can eliminate the undesirable effects of smaller observations on the extreme tail characteristics of a distribution describing extreme events. A resampling method that draws smaller sample sizes from an original sample is known as subsampling. The subsamples come from the same unknown probability distribution as the original sample. As stated in Politis and Romano [20], subsampling is an intensive numerical simulation technique often used in statistical inference. It generates a set of observations, each selected from the true sample of a given length. The samples drawn in subsampling are done without replacement and the subsample size must be less than the record length of the dataset. The rationale behind employing trimming and subsampling for distribution selection lies in the ability of trimming to mitigate the adverse impacts of smaller observations on the upper tail behaviour of a distribution that characterizes extreme events, Ref. [21, 22]. Meanwhile, subsamples represent samples from the actual unknown distribution, analogous to the original sample, Ref. [20].

2.1. Model selection methods

This study examines four of the frequently used model selection techniques in hydrological extremes; the Akaike information criterion, the AIC-corrected, the BIC, and the ADC techniques. The mathematical expression for these methods are as shown in Table 1 where n denotes the size of the sample, m is the number of estimated parameters of the jth operational model, and $L(D|\hat{\theta}) = \prod_{i=1}^{n} g_j(y_i\hat{\theta})$ represents the probability function assessed at the point $\theta = \hat{\theta}$, Linhart and Zucchini [23].

The Kullback-Leibler information which assesses the difference between the actual model $f(x)$ and the model closer to it $K_j = g_j(x, \hat{\theta})$ is utilised by the Akaike information criteria, Akaike [24]. The highest value of the log-likelihood function

Table 2: Candidate distributions.

Distribution	Probability distribution function	Parameters
Normal	$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$	(μ, σ)
Gumbel	$\exp\left\{-\exp\left[-\left(\frac{y-\mu}{\sigma}\right)\right]\right\}$	(μ, σ)
EV2	$\exp\left\{-\left[\left(\frac{y-\mu}{\sigma}\right)\right]^{-\xi}\right\}$	(μ, σ, ξ)
GEV	$\exp\left\{-\left[1+\xi\left(\frac{y-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}$	(μ, σ, ξ)
P3	$\frac{1}{\mu\Gamma(\xi)}(\frac{y-\sigma}{\mu})^{\xi-1}e^{-(\frac{y-\sigma}{\mu})}$	(μ, σ, ξ)
LP3	$\frac{1}{\mu\Gamma(\xi)}(\frac{z-\sigma}{\mu})^{\xi-1}e^{-(\frac{y-\sigma}{\mu})}; z = \log y$	(μ, σ, ξ)
LN	$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}(z-\mu)^2}; z = \log y$	(μ, σ)

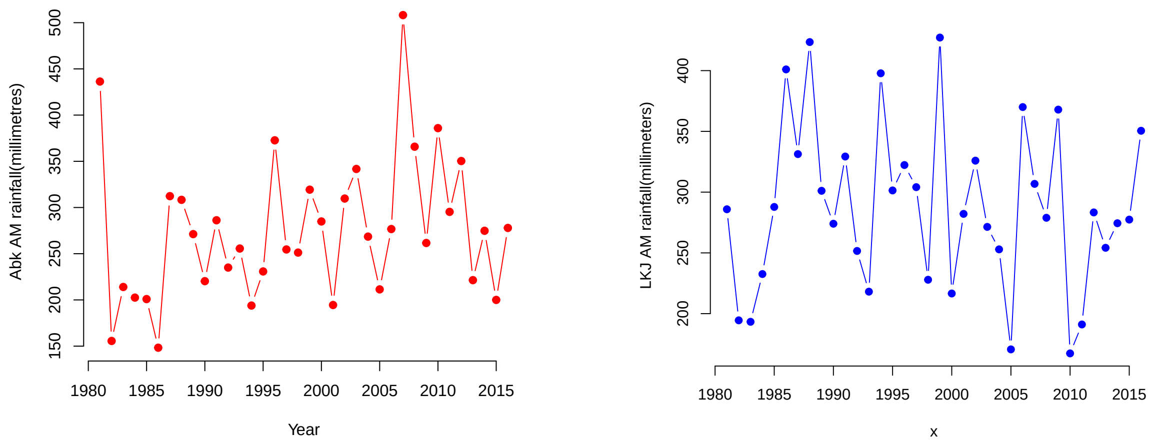
is utilise for model selection, assigning a greater penalty for a higher number of estimated parameters m. In practical application, after computing the AIC_j for all distributions under consideration, the model possessing the least observation is chosen as the best approximating distribution. The difference between the AIC and the AICc is that the AICc imposes a greater penalty than AIC does for the number of estimated parameters m, Ref. [25, 26]. Schwarz [27], suggested the concept of the Bayesian information method. The BIC resembles the AIC but its formulated within a Bayesian foundation. Compared to the AIC, the BIC penalizes more heavily for the number of estimated parameters, as shown in Refs. [25, 28]. Table 1 presents the test statistic for the operative model indexed as j. The Anderson-Darling test utilises the weighted sum of squared variation between actual and conceptual distributions, particularly focusing on differences in the tails where F is the cumulative probability distributions.

2.2. Assessment of model selection criteria

In this study, we will conduct a comprehensive numerical investigation to evaluate the effectiveness of the proposed methodology when handling limited samples in the presence of trimming and subsampling. This investigation will compare the model selection criteria discussed earlier. Monte Carlo experiment will be used to perform the analysis by employing as operative models M_j, a collective of seven distributions often employed in the analysis of hydrological extremes. Table 2 gives the approximating probability distributions utilised in this study. These distributions are among the frequently used distributions in the analysis of flood frequency, Refs. [1, 7, 10, 29], and are highly recommended across the globe, Cunnane [9].

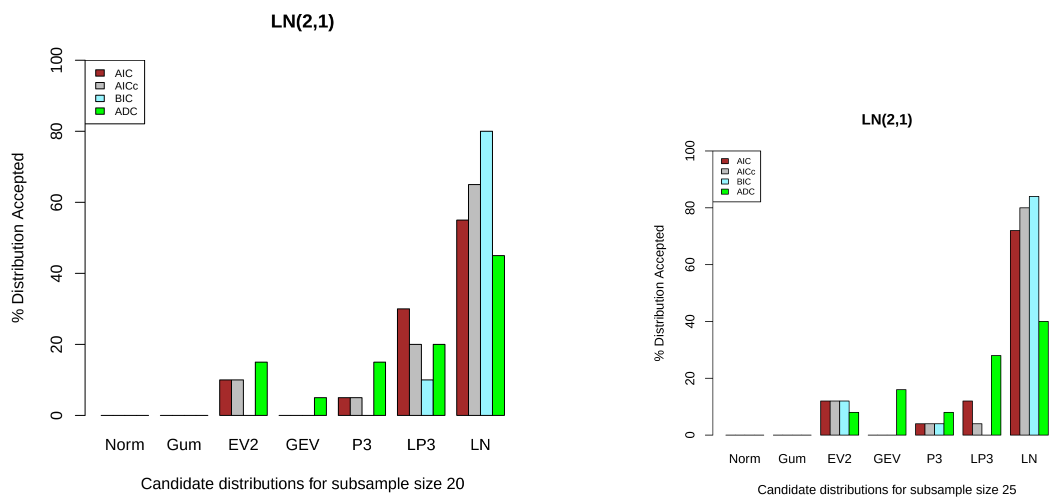
To achieve the objective of this study, the Monte Carlo experiment is organised in the following manner:

1. Let $f(z) = g_{j^*}(z, \theta^*)$ be a lognormal parent distribution with a given set of parameters. We generate 1000 subsamples each of size n from the fitted lognormal distribution.
2. Subsamples of size b are selected without replacement from the subsamples generated in (i) with $b < n$. The size of the subsample b depends on the length, n, of the sample.



(a) Station ABK. (b) Station LKJ.

Figure 1: Annual maximum plots.



(a) (b)

Figure 2: Acceptance percentage of the LN distribution for n=30 with b=20,25 for the untrimmed samples.

3. Trimming Proportion (TP) of 10% and 15% are applied to each of the subsamples of size b selected.
4. Obtain the AIC_j, $AICc_j$, BIC_j values, and ADC_j p-values, for the untrimmed and trimmed subsamples in step iii for each of the candidate distribution, $j = 1, 2, 3, .., 7$.
   The model having the lowest AIC value is chosen
5. The model M_i^* having the lowest AIC value is chosen, denoted by $AIC_{i^*} = AIC_{min}$. If i^* equals j^*, AIC is chosen as it correctly identifies the true parent distribution. The same process is repeated for AICc and BIC. For the Anderson-Darling Criterion, if the non-exceedance probability, $P(A_2)$, of the Anderson-Darling test statistic A^2 is greater than k, where k is one minus the level of significance, then we fail to accept the candidate distribution.
6. The test procedure in steps (i)-(v) are repeated for different subsample sizes, b. The frequency of selection for each of the model selection technique is recorded.

According to the handbook of flood estimation from the Institute of Hydrology Ref. [30], to perform at-site flood frequency analysis, it is recommended to have a minimum data record length of 20 years. Hence, in this study, the subsample size b will have a record length of at least 20 years. The AIC, AICc, BIC, and ADC tests are applied to every subsample to ascertain its probability distribution. The proposed trimming and subsampling methodology will be applied to an at-site annual maximum rainfall data series to determine the

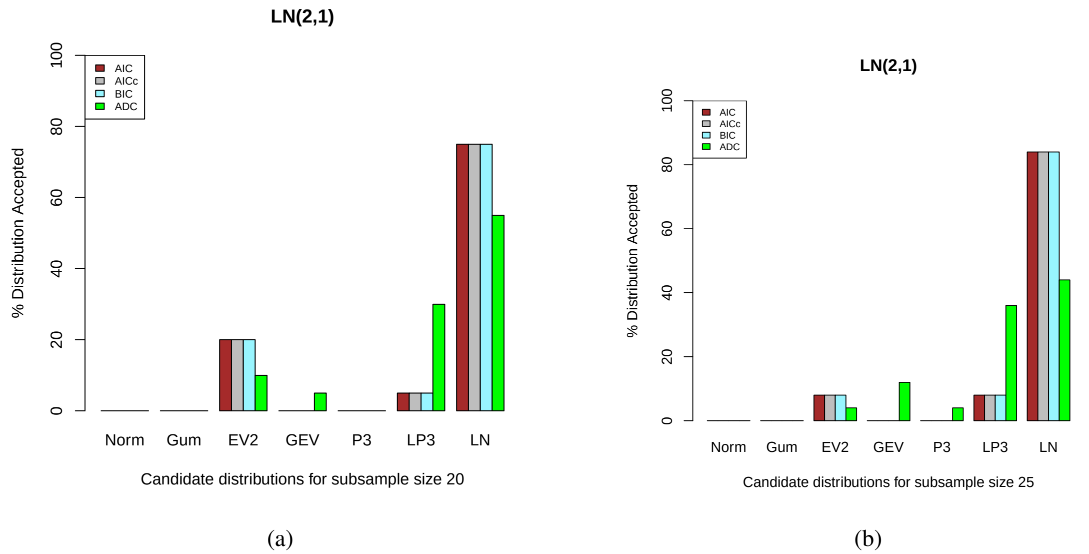


Figure 3: Acceptance percentage of the LN distribution for n=40 with b=20,25 for the untrimmed samples.

most suitable distribution in situations where the goodness-of-fit test cannot discriminate between distributions. Annual rainfall data for two cities in Nigeria obtained from the Nigeria Meteorological Agency and published by Nigeria's central bank: https://www.cbn.gov.ng will be used. In the simulation experiment, subsamples are generated from a given parent distribution, and the model selection techniques are used to establish the candidate distribution that suitably describes the underlying mechanism generating the data.

3. Results and discussion

To verify whether the model selection techniques work correctly in discriminating among probability distributions used in hydrological extremes, four conventional model selection methods; AIC, AICc, BIC, and ADC, are applied to observed Annual Maximum (AM) rainfall data on two geographical locations; Abeokuta (ABK) and Lokoja (LKJ), in Nigeria. ABK, located in southwest Nigeria, is situated on the eastern side of the Ogun River amid rocky outcroppings in a wooded savanna, it covers a region covering approximately 879 square kilometers. Lokoja, located in the north-central of Nigeria, has a tropical wet and dry savanna climate. It is located at the meeting point of the river Niger and river Benue and covers an area of 3,180 km^2.

Table 3 gives important information of ABK and LKJ data series in terms of their location, data length, and moments, while Figure 1 represents the data series plots for these locations. Table 4 displays the outcomes of using the model selection techniques alone when applied to the observed yearly peak data series for ABK and LKJ using seven candidate distributions that are frequently used in the study of hydrological extremes. The standard tests for model selection are applied to the annual peak data of ABK and LKJ to identify the optimal distribution for the observed data series. To select the optimal candidate distribution, the distribution with the least AIC, AICc, or BIC value is selected, while for the ADC, if the non-exceedance probability, $P(A_2)$, of the Anderson-Darling test statistic A^2 is greater than k, where k is one minus the level of significance, the candidate distribution is rejected, see Laio [31]. Table 4 is a summary of the test results of the model selection criteria for AIC, AICc, BIC and ADC tests. From Table 4, at location ABK, the model selection criteria (AIC, AICc, and BIC) indicate the Gumbel and lognormal distributions as the optimal distributions, with exact test values of 412.8, 413.2, and 416.0, respectively. Similarly, at location LKJ, the normal and lognormal distributions are favoured by the AIC, AICc, and BIC, with test values of 409.3, 409.7, and 412.5, respectively. For the ADC test, all candidate distributions were accepted at location ABK, given that $P(A_2) < k = 1 - \alpha$. Conversely, at location LKJ, all candidate distributions were accepted except for the EV2 distribution. Overall, when considering the observed data from both ABK and LKJ, it appears that the standard model selection criteria failed to distinguish between candidate distributions effectively. To assess the efficiency of the proposed method in selecting the true parent distribution, a Monte Carlo study was performed, and the methodology is applied to real data.

3.1. Simulation study

The primary aim of the simulation study is to evaluate how well the proposed method performs under various conditions, including different sample sizes, subsample sizes, trimming proportion and inherent statistical properties of the sample.The outcomes of the Monte Carlo experiment carried out are presented in this section. The results presentations are a bit cumbersome because the result is dependent on the parent distribution, sample size, trimming proportions, and subsampling size. For the purpose of a clearer presentation, results that are considered important for the purpose of this paper will be presented.

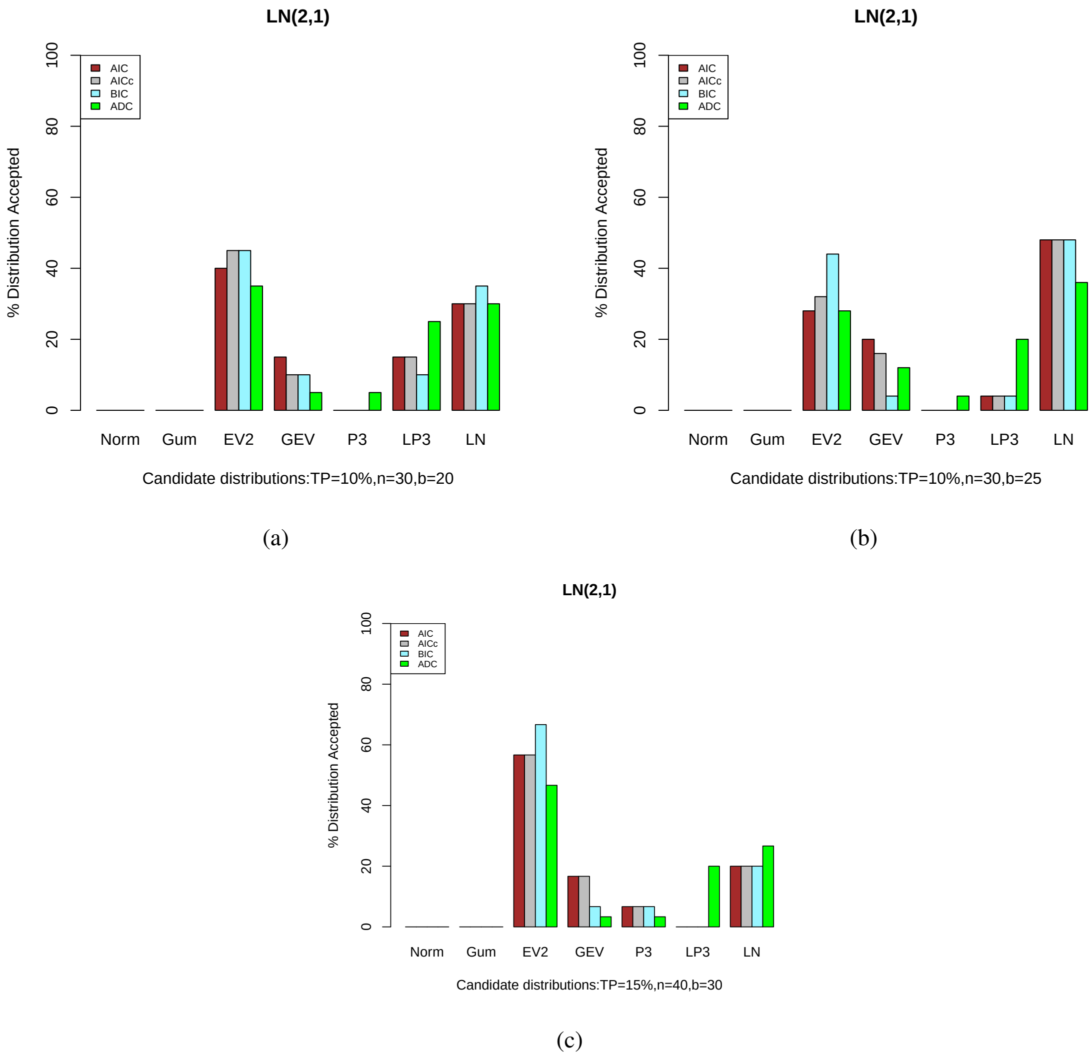


Figure 4: Acceptance percentage of candidate distribution for trimmed samples with lognormal parent distribution.

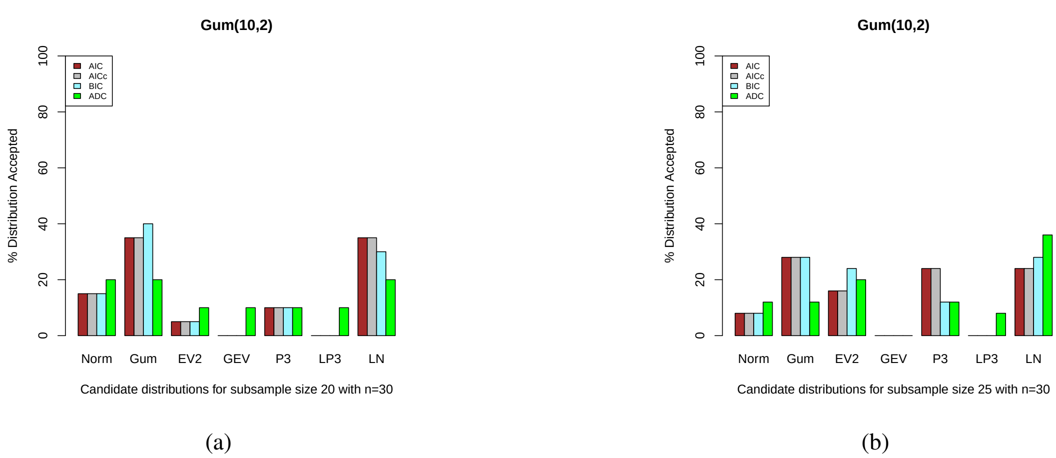
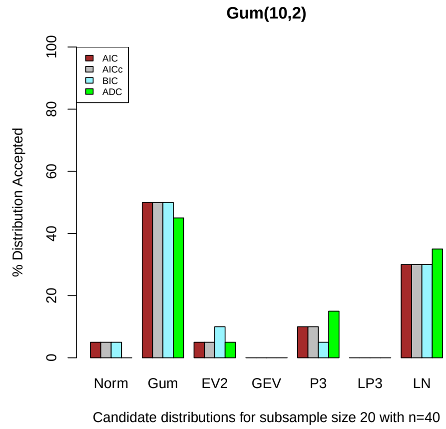
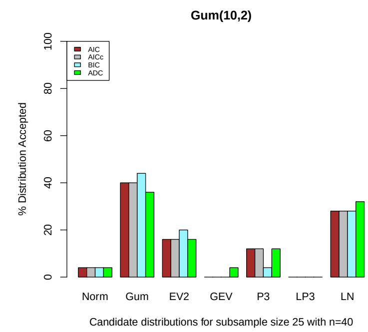


Figure 5: Acceptance percentage of the Gumbel distribution for n=30 with b=20,25 for the untrimmed samples.

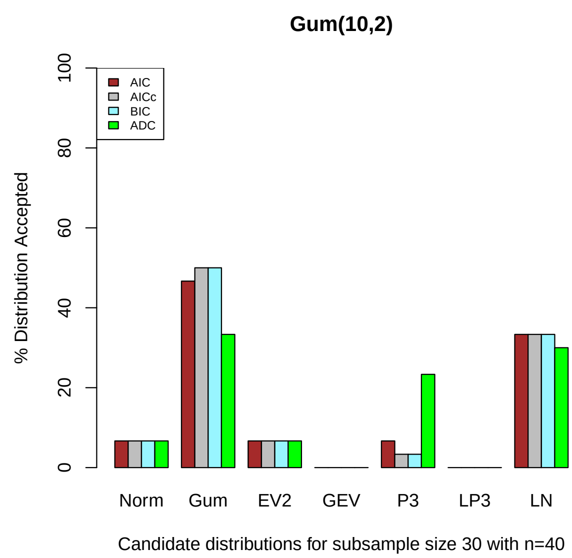
Figure 2 and Figure 3 show the acceptance percentages of the seven candidate distributions by the four model selection techniques for a lognormal parent distribution for untrimmed samples. The model selection technique correctly chooses the



(a)



(b)



(c)

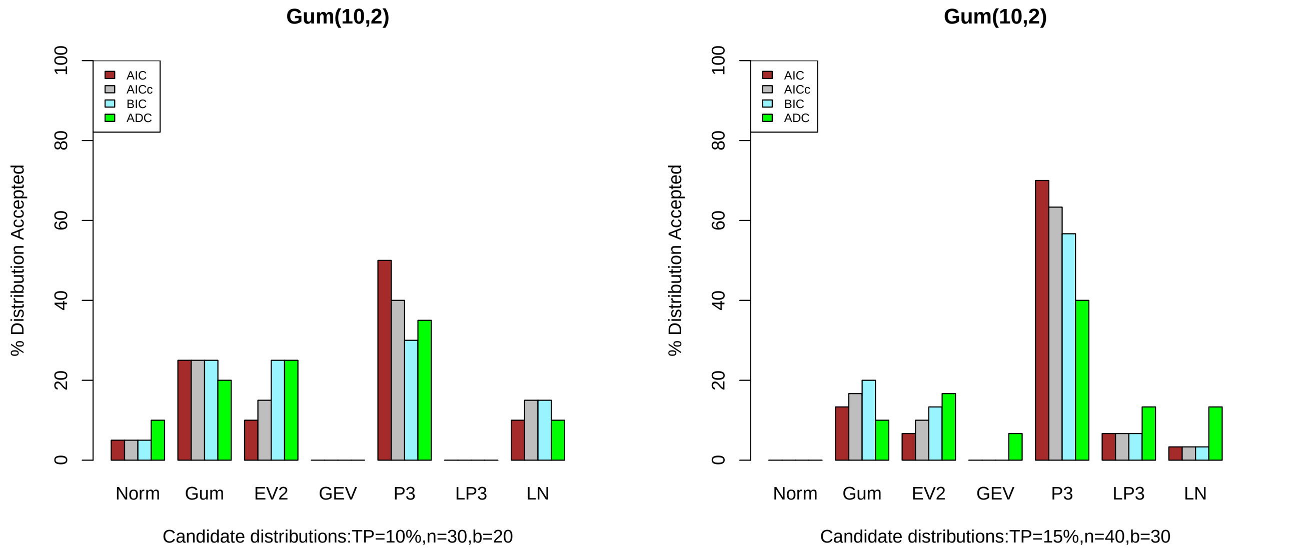
Figure 6: Acceptance percentage of the Gumbel distribution for n=40 with b=20,25,30 for the untrimmed samples.

Table 3: Statistical summary of abk and lkj annual maximum rainfall data.

Location	Data length(years)	Area (km^2)	Mean	Variance	Skewness	Kurtosis
ABK	36	879	274.9333	5706.6600	0.8977	4.0082
LKJ	36	3180	287.3889	4540.2454	0.2389	2.5100

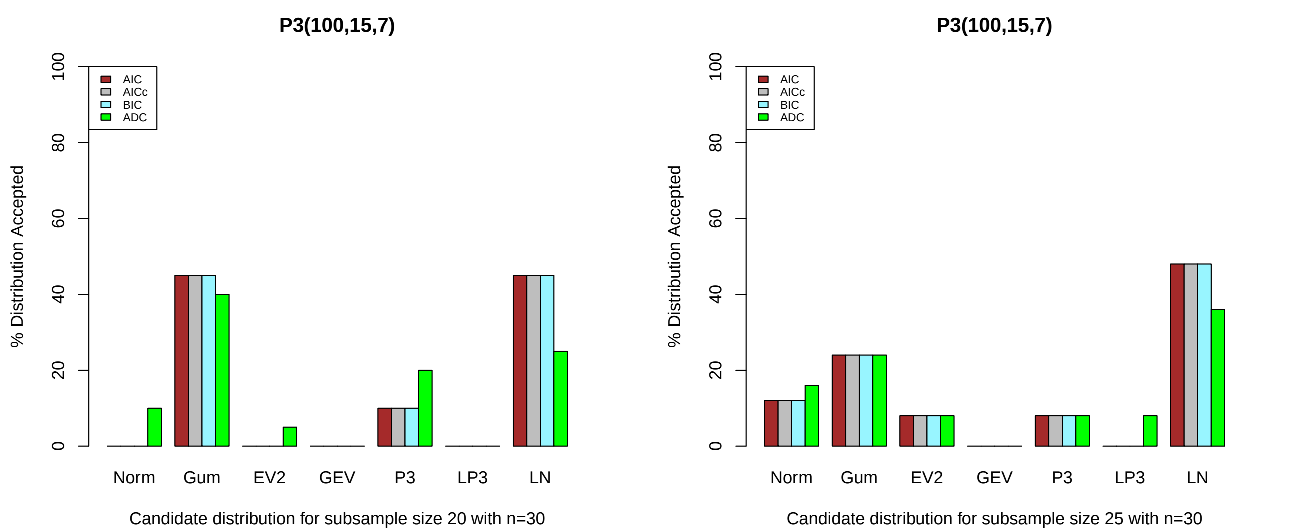
lognormal distribution as the parent distribution between 40% to 84% of cases. The Bayesian information criteria are the most effective in recognising the parent distribution than the AIC, AICc, and ADC tests in this sequence. Though the AIC and AICc show similar tendencies in selecting the parent distribution, the AICc is more effective than the AIC. The ADC selects the lognormal distribution as the parent distribution in around 40-55% of the cases.

As can be seen from Figure 4, when the lognormal distribution is used as the parent distribution for trimmed samples, results show that either the lognormal or EV2 distribution is selected as the parent distribution at 10% trimming proportion. The lognormal distribution is selected between 30-48% of the times while the EV2 distribution is chosen as the parent distribution between 28-45%. However, as the trimming proportion is increased to 15% for sample size 40, the EV2 distribution is selected by all model selection methods as the parent distribution between 56.67-66.67%. This outcome remains largely consistent even when considering lognormal distributions with different variances as the parent distribution. The tendency to select the EV2 distribution, a three-parameter distribution, by the four model selection criteria as the parent distribution can



(a) (b)

Figure 7: Acceptance percentage of candidate distribution for trimmed samples with Gumbel parent distribution.



(a) (b)

Figure 8: Acceptance percentage of the P3 distribution for n=30 with b=20,25 for the untrimmed samples.

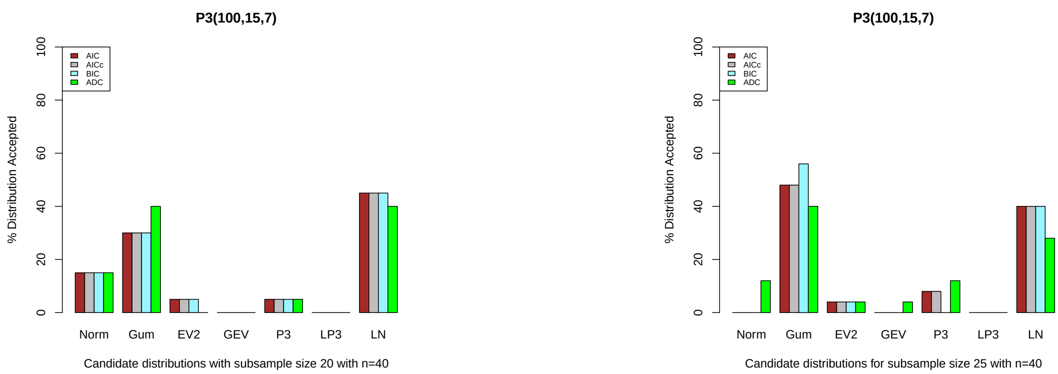
be attributed to the effect of smaller observed values on three-parameter distributions. Onoz and Bayazit [10] state that while some studies are interested in the behavioral analysis of observed flood data, others are keen on the predictive ability by evaluating how good the estimates of the distribution selected are compared to the assumed parent distribution for estimating quantiles. In such a situation, it will be interesting to see how the estimates of quantiles behave, in terms of prediction error, between the distribution selected when trimming is introduced and the true parent distribution.

By analysing Figure 5, one can observe that for the untrimmed samples, the model selection techniques couldn't distinguish between the Gumbel parent distribution and the log-normal distribution in some cases (see Figure 5(a) where only the BIC selects the true parent distribution outrightly with acceptance rate of 40%). However, as the sample size is increased to 40, the Gumbel distribution is selected by all model selection techniques as the parent distribution in around 33.33% to 50% cases for all subsample sizes, see Figure 6. The BIC is however more efficient in recognising the parent distribution than the AICc, AIC, and ADC tests, in this order.

Employing the Gumbel distribution as the parent distribu-

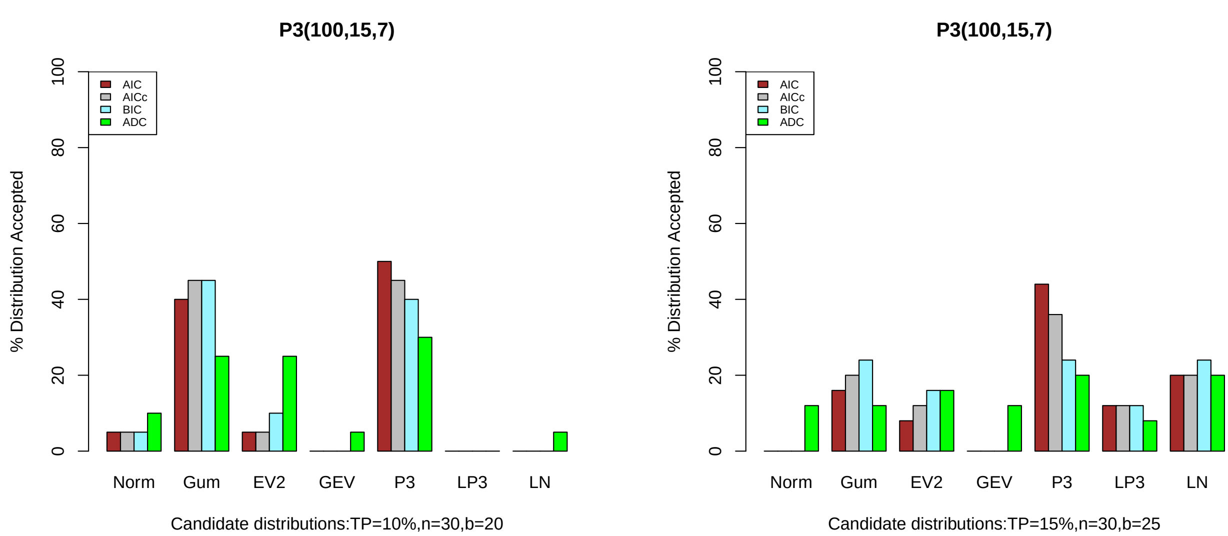
Table 4: Model selection test results for locations ABK and LKJ dataset at the 5% significance level.

		Distribution						
Location	Test	Normal	Gumbel	EV2	GEV	P3	LP3	LN
ABK	AIC	417.5	412.8	417.1	414.7	414.7	414.7	412.8
	AICc	417.9	413.2	417.5	415.5	415.5	415.5	413.2
	BIC	420.7	416.0	420.3	419.5	419.5	419.5	416.0
	ADC	0.7745	0.0685	0.8313	0.0955	0.1198	0.1062	0.0578
LKJ	AIC	409.3	410.7	416.4	410.4	410.7	410.5	409.3
	AICc	409.7	411.0	416.8	411.1	411.5	411.2	409.7
	BIC	412.5	413.8	419.6	415.1	415.5	415.3	412.5
	ADC	0.3653	0.6975	0.9869	0.3902	0.4255	0.4146	0.4733



(a) (b)

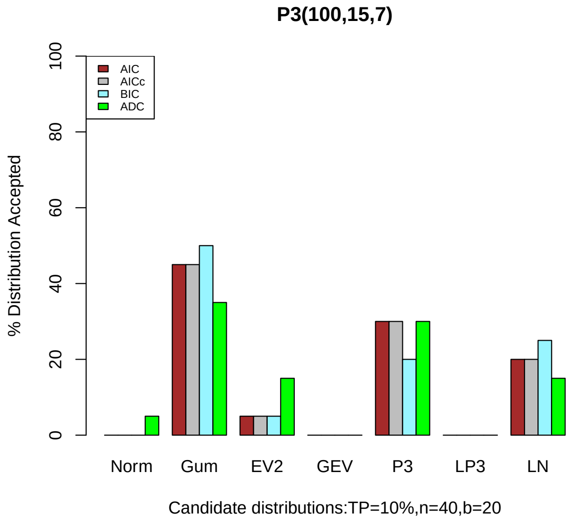
Figure 9: Acceptance percentage of the P3 distribution for n=40 with b=20,25 for the untrimmed samples.



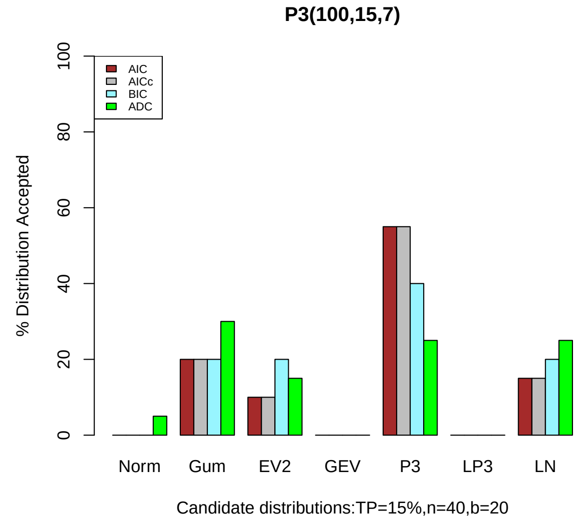
(a) (b)

Figure 10: Acceptance percentage of candidate distribution for trimmed samples with P3 parent distribution.

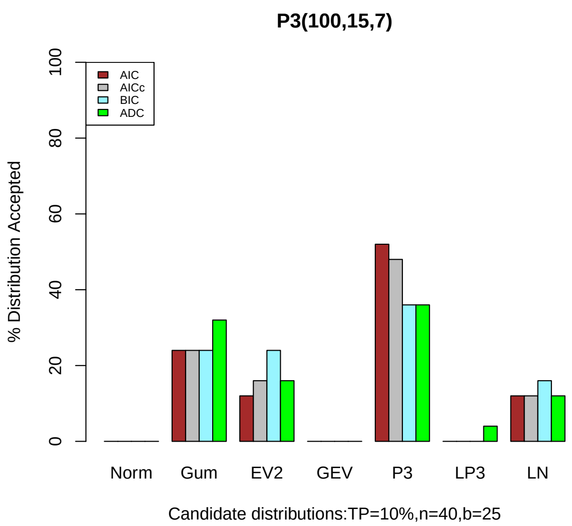
tion and introducing trimming, all model selection criteria indicate the P3 distribution as the best choice (see Figure 7). There is therefore a tendency for the model selection technique to opt for a three-parameter distribution instead of the true parent two-parameter distribution. Since one of the goals of flood frequency analysis is the extrapolation of flood quantiles, this result may not necessarily be a constraint, as in some cases, a distribution that suitably describes flood data may not be ad-



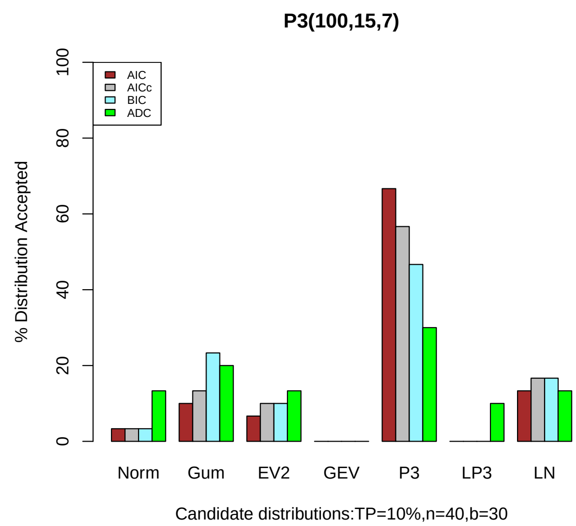
(a)



(b)



(c)



(d)

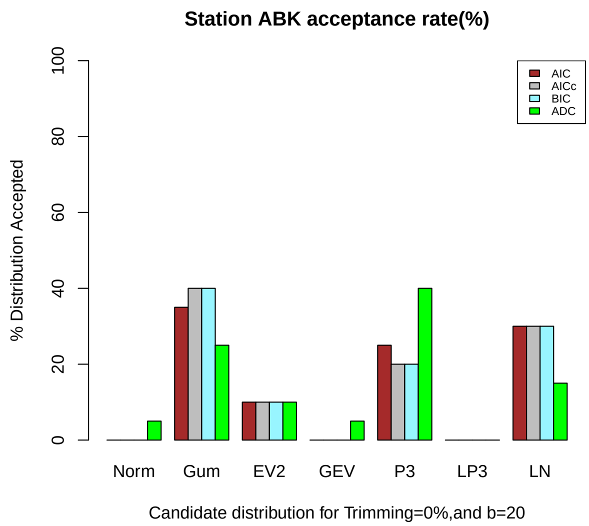
Figure 11: Acceptance percentage of candidate distribution for trimmed samples with P3 parent distribution.

equate for estimating flood quantiles. This, however, changes the perspective from the problem of recognising the true parent distribution to seeking the best optimal operational model for quantile estimation, Laio *et al.* [15].

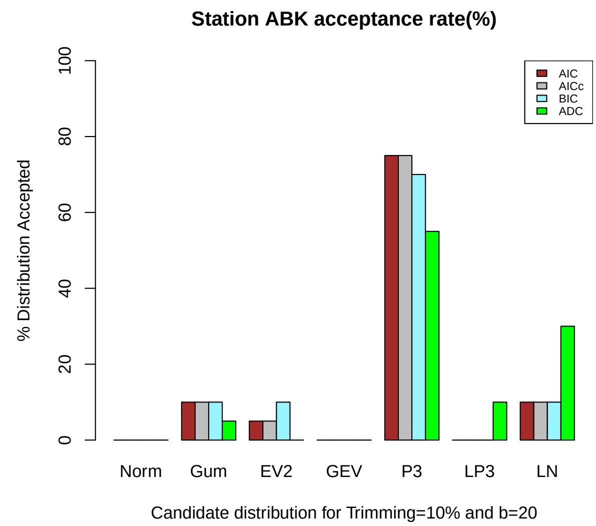
Using the Pearson type III distribution as the parent distribution, the model selection criteria are favorably disposed to selecting a two-parameter distribution instead of the three-parameter parent distribution for all samples and subsample sizes for the untrimmed sample, see Figures 8 and 9 . The tendency to select a two-parameter distribution by the model selection criteria given the parent distribution is a three-parameter distribution is perhaps implicitly a result of model parsimony, which suggests choosing the simplest possible distribution that sufficiently captures the underlying mechanism generating the observed data, Box *et al.* [32].

By analysing Figures 10 and 11 we can observe that with the introduction of trimming when Pearson type III distribution is used as the underlying distribution, all model selection methods correctly select the Pearson type III distribution between 16% to 66.67% for all sample and subsample sizes; AIC turns out to be the best test even when varying the sample size, subsample size, and trimming proportion.

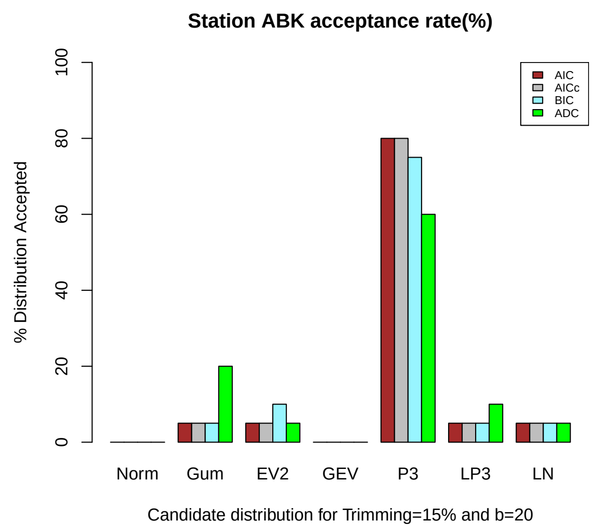
The tendency for the model selection techniques to choose the true parent distribution when trimming is introduced is a result of removing the undesirable effect smaller observations may have on the extreme of a distribution, especially for a three-parameter distribution.



(a)



(b)



(c)

Figure 12: Percentage of cases a candidate distribution is accepted for subsample size 20.

3.2. Application to real data

We return to the observed yearly peak rainfall data of ABK and LKJ with descriptive statistics as shown in Table 3. The primary objective is to assess whether trimming and subsampling, in combination with model selection techniques, can effectively differentiate among candidate distributions, which otherwise cannot be accomplished solely through the use of model selection techniques alone, as demonstrated in Table 4, particularly when working with small samples. Therefore, the proposed methodology is now applied to the real datasets.

Through a Monte Carlo simulation, 1000 subsamples of datasets with similar characteristics and record length, see Table 3, as the observed annual maximum data for ABK and LKJ are generated. The subsampling scheme is first applied to the datasets without trimming; and subsequently applied to the trimmed data.

The procedure for the implementation of the proposed method to the observed datasets is structured as follows:

i Generate 1000 subsamples with similar characteristics and record length n as the original ABK data series.

ii Subsamples of size b with $b < n$ are selected without replacement from the 1000 subsamples, where the size of the subsample depends on the record length of the ABK data series.

iii A Trimming proportion (TP) of 10% and 15% is applied to each of the subsamples of size b selected in (ii) above.

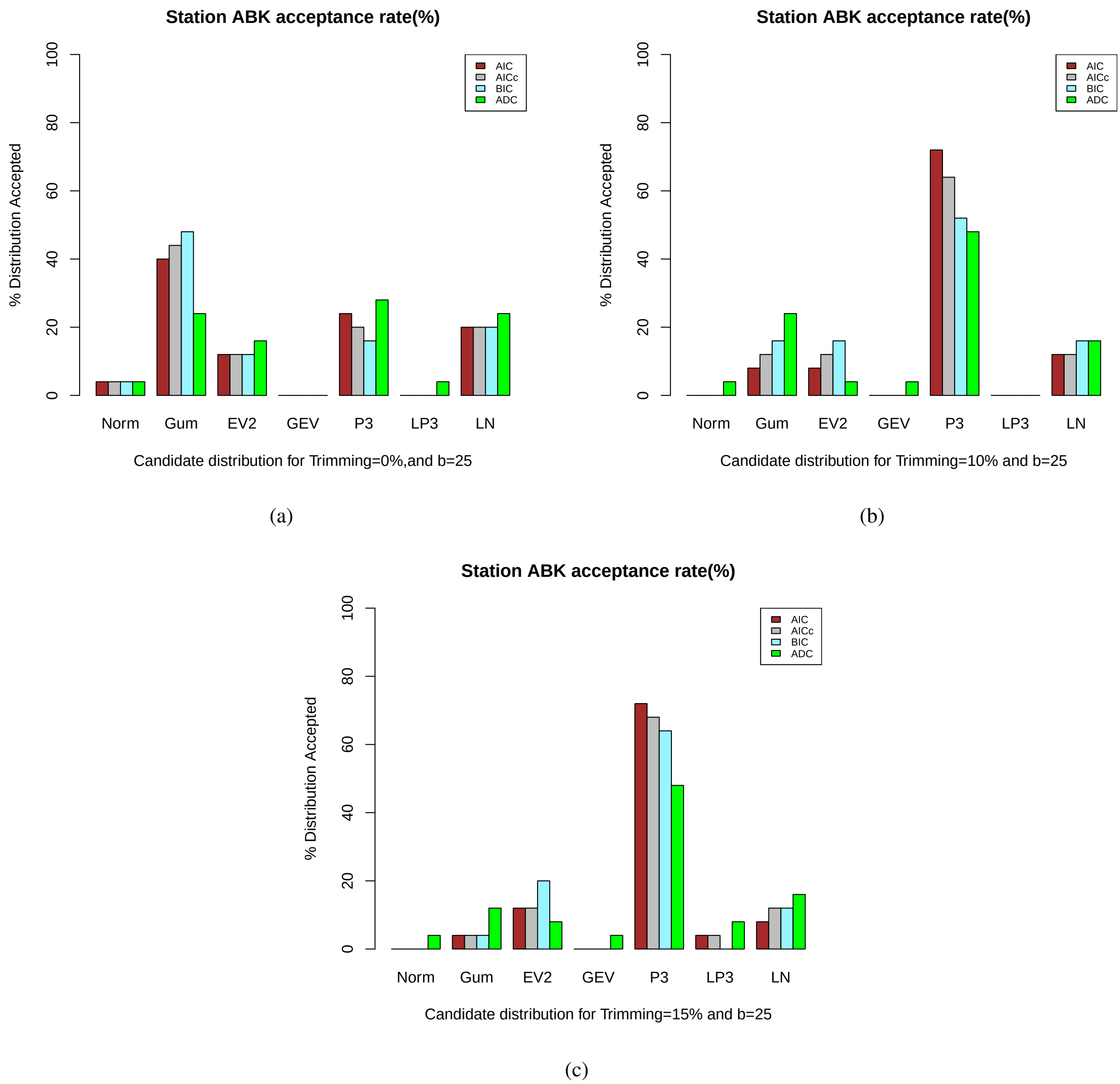


Figure 13: Percentage of cases a candidate distribution is accepted for subsample size 25.

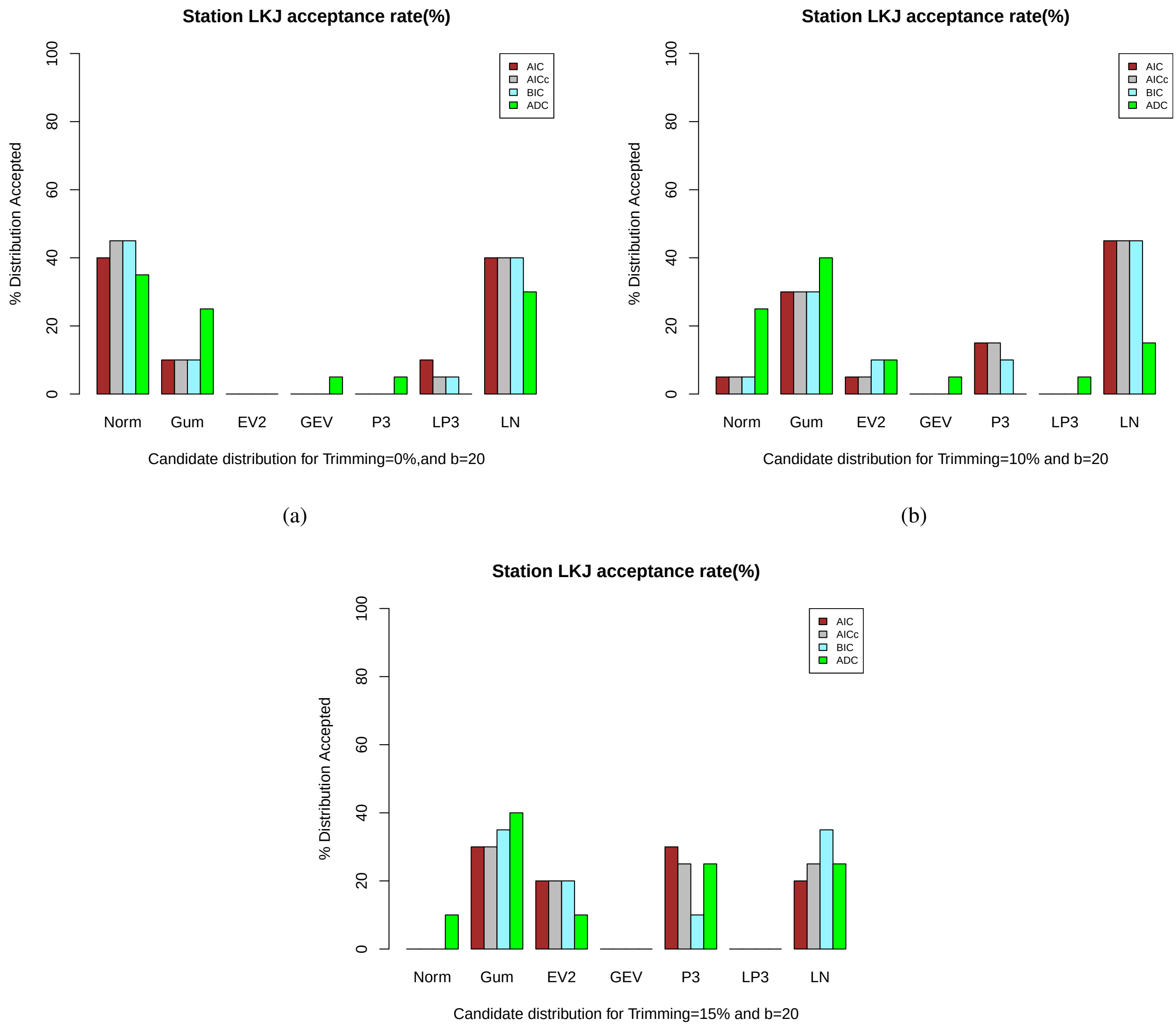
iv The four model selection techniques are then applied to the trimmed and untrimmed subsamples selected using the seven candidate distributions.

v The test statistic for each of the model selection techniques is computed and noted.

vi The procedure in steps (i)-(v) is repeated for each of the b subsamples, and the distribution that is accepted more frequently by each of the goodness of fit tests is noted. Steps $(i)-(vi)$ are also performed for the LKJ dataset.

The proposed method application to the observed station ABK data is summarised in Figures 12 and 13. These Figures show the accepted distributions for varying trimming proportions and subsample sizes. A trimming proportion of 0% implies that the dataset used is for an untrimmed sample. By analysing Figure 12(a), one can see that for the untrimmed sample, the AIC, AICc, and BIC consistently select the Gumbel distribution as optimal, with acceptance proportions of 35%, 40%, and 40%, respectively, while the Anderson-Darling test selects the P3 distribution as optimal, with an acceptance percentage of 40%.

However as trimming is introduced, the P3 distribution is consistently selected as the optimal distribution by all four goodness-of-fit tests. From Figures 12(b) and (c) , we can observe that both AIC and AICc exhibit similar maximum acceptance percentages of 75% and 80% at trimming proportions of 10% and 15%, respectively. Although the Anderson-Darling test yields the lowest acceptance percentage for the P3 distri-



(a) (b)

(c)

Figure 14: Percentage of cases a candidate distribution is accepted for subsample size 20.

bution, it consistently selects the P3 distribution as optimal for both untrimmed and trimmed data of station ABK.

Figure 13 shows the candidate distributions accepted with varying trimming proportions for a subsample size of 25. In the untrimmed station ABK data series, the behavior of the four goodness-of-fit tests remains similar compared to when the subsample size is 20. The acceptance percentage of the Gumbel distribution, as shown in Figure 13(a), increases to 40%, 44%, and 48% for the AIC, AICc, and BIC tests, respectively, compared to 35%, 40%, and 40% observed when the subsample size is 20 while, the ADC test selects the P3 distribution with an acceptance percentage of 28% for a subsample size of 25. In Figures 13(b) and (c), with trimming introduced to the station ABK data series, all four model selection methods choose the P3 distribution as the optimal distribution, with the AIC, AICc, BIC, and ADC tests showing a marked tendency towards selecting the P3 distribution accordingly. In practical situations, we don't know the true parent distribution, however, from examining Figures 12 and 13 we can observe some compelling considerations. Comparing the results from the different model selection techniques, it is evident that the ADC test consistently predisposes to a three-parameter distribution for both trimmed and untrimmed samples, whereas AIC, AICc, and BIC tend to favour two-parameter distributions for untrimmed samples and three-parameter distributions for trimmed samples. Other factors to be considered are how well different goodness-of-fit tests select a distribution. The ADC test consistently selects the P3 distributions for both the trimmed and untrimmed data sets. This demonstrates that the Anderson-Darling test possesses distinct selection abilities compared to the other three model selection methods.

We have shown the application of trimming and subsam-

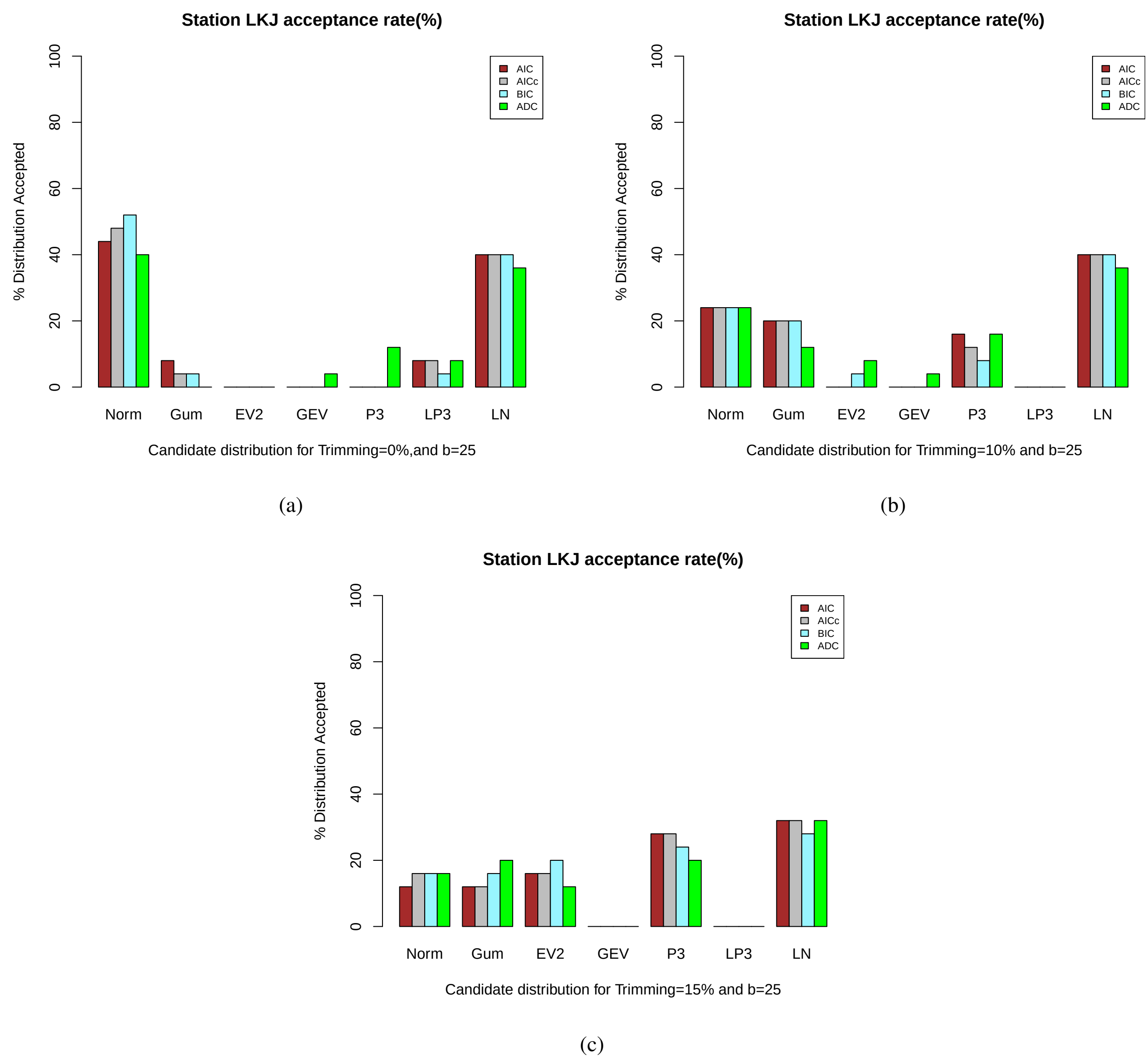


Figure 15: Percentage of cases a candidate distribution is accepted for subsample size 25.

pling, with the aid of model selection techniques, to yearly peak rainfall data from ABK. The results for datasets from LKJ are reported in Figures 14 and 15. From Figure 14(a), the normal distribution is selected as the optimal distribution by the four model selection criteria at around 40-45% when the samples are untrimmed and the subsample size is 20. The AIC test, however selects both the normal and lognormal distribution with an acceptance percentage of 40%. The results, as evident from Figures 14(b) and (c) show that the ADC test behaves somewhat differently from the other model selection criteria. At 10% trimming proportion, AIC, AICc, and BIC select the lognormal distribution as the optimal distribution 45% of the time, respectively, while the ADC test chooses the Gumbel distribution as the optimal distribution with 40% acceptance. When trimming proportion is increased for subsample size 20, some of the model selection criteria couldn't distinguish between some distributions. At 15% trimming proportion, the AIC test selects the Gumbel and P3 distribution as the optimal, while the BIC test selects the Gumbel and lognormal distribution. However, the AICc and ADC tests outright choose the Gumbel distribution as the best fit distribution.

From Figure 15, as the subsample size is increased to 25, all four model selection criteria outrightly select the normal distribution as the optimal distribution for the untrimmed samples, with acceptance of around 40% to 52%. As trimming is introduced, the lognormal distribution is selected as the optimal distribution by all model selection techniques. It is noteworthy that the ADC test leans towards the Gumbel distribution as it has the highest acceptance of 40% at subsample size 20 for the 10% and 15% trimming proportions respectively. Comparing the results for the LKJ dataset, the model selection criteria choose normal distribution as the optimal when the sample is untrimmed with a maximum acceptance rate of 44%, 48%, 52%, and 40% respectively, for the AIC, AICc, BIC, and ADC tests. For the trimmed samples, all model selection criteria select the lognormal distribution as the optimal distribution except for the ADC test, whose highest acceptance rate of 40% is the Gumbel distribution. The ADC test predisposition to select Gumbel as against the lognormal distribution chosen by the AIC, AICc, and BIC tests for the trimmed samples is not necessarily a constraint for flood quantile estimates for smaller samples. As estimating design floods is one of the objectives of frequency analysis, it will be interesting to see which distribution between lognormal and Gumbel will give quantile estimations with reduced bias and standard error.

4. Conclusion

In this study, utilising model selection techniques, we explore the potential usefulness of trimming and subsampling to distinguish between candidate distributions employed in flood frequency analysis, which otherwise may not be possible by using the standard model selection technique alone. The idea is that while the subsamples belong to the unknown true parent distribution as the real sample, trimming removes the undesirable effects smaller observations may have on the upper tail characteristics of the distribution. The proposed methodology has shown the ability to discriminate between candidate distributions, which otherwise cannot be achieved by using only the standard goodness of fit tests. In a practical situation, our opinion is that when two distinct distributions are accepted by the selection criteria for the untrimmed and trimmed datasets and the aim is to extrapolate rare events; an effective operational approach may be to use these two distributions for quantile estimates and adopt the distribution with the least estimation error.

In summary, the model selection techniques with the aid of subsampling are effective in identifying the true parent distribution for the untrimmed samples when it is a two-parameter distribution; contrarily, they are not as effective when the parent distribution is a three-parameter distribution. Nevertheless, as trimming is introduced, all model selection methods recognise the true parent distribution for a three-parameter distribution. Overall, utilising trimming and subsampling with the aid of model selection techniques yields promising outcomes in frequency analysis of hydrological extremes.

Acknowledgment

We extend our gratitude to Universiti Putra Malaysia for their financial backing of this research through the PUTRA GRANT GP/2023/9753100.

References

[1] R. Kidson & K. S. Richards, "Flood frequency analysis: assumptions and alternatives", Progress in Physical Geography **29** (2005) 392. https://journals.sagepub.com/doi/abs/10.1191/0309133305pp454ra.

[2] J. R. Wallis, *Regional frequency analysis: An approach based on l-moments*, Cambridge University Press, New York, 1997. https://doi.org/10.1017/CBO9780511529443.

[3] L. Yan, L. Xiong, G. Ruan, C. Y. Xu, P. Yan & P. Liu, "Reducing uncertainty of design floods of two-component mixture distributions by utilizing flood timescale to classify flood types in seasonally snow covered region", Journal of Hydrology **574** (2019) 588. https://www.sciencedirect.com/science/article/abs/pii/S0022169419303968.

[4] L. Yan, L. Zhang, L. Xiong, P. Yan, C. Jiang, W. Xu, B. Xiong, K. Yu, Q. Ma & C. Y. Xu, "Flood frequency analysis using mixture distributions in light of prior flood type classification in Norway", Remote Sensing **15** (2023) 401. https://doi.org/10.3390/rs15020401.

[5] A. Flammini, J. Dari, C. Corradini, C. Saltalippi & R. Morbidelli, "Areal reduction factor estimate for extreme rainfall events", In *Rainfall*, Elsevier, 2022, pp. 285–306. https://doi.org/10.1016/B978-0-12-822544-8.00014-7.

[6] S. A. Moges & M. T. Taye, "Regional flood frequency curves for remote rural areas of the Nile River Basin: The case of Baro-Akobo drainage basin, Ethiopia", In *Extreme Hydrology and Climate Variability*, Elsevier, 2019, pp. 385–393. https://www.sciencedirect.com/science/article/abs/pii/B9780128159989000300.

[7] K. Hamed, & A. R. Rao, *Flood frequency analysis*, CRC press, 2019. https://doi.org/10.1201/9780429128813.

[8] F. Laio, G. Di Baldassarre & A. Montanari, "Model selection techniques for the frequency analysis of hydrological extremes", Water Resources Research **45** (2009) W07416. https://doi.org/10.1029/2007WR006666.

[9] C. Cunnane, *Statistical distribution for flood frequency analysis*, WMO Operational Hydrology, Report No. 33, WMO-No. 718, Geneva, Switzerland, 1989. https://library.wmo.int/viewer/33760?medianame=wmo_718_#page=1&viewer=picture&o=bookmarks&n=0&q=.

[10] B. Onoz & M. Bayazit, "Best-fit distributions of largest available flood samples", Journal of Hydrology **167** (1995) 195. https://doi.org/10.1016/0022-1694(94)02633-M.

[11] P. K. Langat, L. Kumar & R. Koech, "Identification of the most suitable probability distribution models for maximum, minimum, and mean streamflow", Water **11** (2019) 734. https://doi.org/10.3390/w11040734

[12] S. Das, "An assessment of using subsampling method in selection of a flood frequency distribution", Stochastic Environmental Research and Risk Assessment **31** (2017) 2033. https://link.springer.com/article/10.1007/s00477-016-1318-3.

[13] X. Chen, Q. Shao, C. Y. Xu, J. Zhang, L. Zhang & C. Ye, "Comparative study on the selection criteria for fitting flood frequency distribution models with emphasis on upper-tail behavior", Water **9** (2017) 320. https://doi.org/10.3390/w9050320.

[14] M. Ul Hassan, O. Hayat & Z. Noreen, "Selecting the best probability distribution for at-site flood frequency analysis; a study of Torne River", SN Applied Sciences **1** (2019) 1. https://link.springer.com/article/10.1007/s42452-019-1584-z.

[15] G. Di Baldassarre, F. Laio & A. Montanari, "Design flood estimation using model selection criteria", Physics and Chemistry of the Earth Parts ABC **34** (2009) 606. https://doi.org/10.1016/j.pce.2008.10.066.

[16] K. Haddad & A. Rahman, "Selection of the best fit flood frequency distribution and parameter estimation procedure: a case study for Tasmania in Australia", Stochastic Environmental Research and Risk Assessment **25** (2011) 415. https://link.springer.com/article/10.1007/s00477-010-0412-1.

[17] A. S. Rahman, A. Rahman, M. A. Zaman, K. Haddad, A. Ahsan & M. Imteaz, "A study on selection of probability distributions for at-site flood frequency analysis in Australia", Natural hazards **69** (2013) 1803. https://link.springer.com/article/10.1007/s11069-013-0775-y.

[18] X. Zeng, D. Wang & J. Wu, "Evaluating the three methods of goodness of fit test for frequency analysis", Journal of Risk Analysis and Crisis Response **5** (2015) 178. https://www.jracr.com/index.php/jracr/article/view/151.

[19] A. O. Adeyemi, I. A. Adeleke & E. E. Akarawak, "Modeling extreme stochastic variations using the maximum order statistics of convoluted distributions", Journal of the Nigerian Society of Physical Sciences **5** (2023) 994. https://doi.org/10.46481/jnsps.2023.994.

[20] D. N. Politis, J. P. Romano & M. Wolf, *Subsampling*, Springer Science & Business Media, 2012. https://link.springer.com/book/10.1007/978-1-4612-1554-7.

[21] S. Coles, J. Bawa, L. Trenner & P. Dorazio, *An introduction to statistical modeling of extreme values*, Springer, London, 2001, p. 208. https://link.springer.com/book/10.1007/978-1-4471-3675-0.

[22] K. P. Bhattarai, "Partial L-moments for the analysis of censored flood samplesUtilisation des L-moments partiels pour l'analyse d'échantillons tronqués de crues", Hydrological sciences journal **49** (2004) 855. https://doi.org/10.1623/hysj.49.5.855.55138.

[23] W. Zucchini, "An introduction to model selection", Journal of mathematical psychology **44** (2000) 41. https://www.sciencedirect.com/science/article/abs/pii/S0022249699912762.

[24] H. Akaike, "Information theory and an extension of the maximum likelihood principle", In *Selected papers of hirotugu akaike*, Springer, New York, NY, 1998, pp. 199–213. https://link.springer.com/chapter/10.1007/978-1-4612-1694-0_15.

[25] G. Calenda, C. P. Mancini & E. Volpi, "Selection of the probabilistic model of extreme floods: The case of the River Tiber in Rome", Journal of Hydrology **371** (2009) 1. https://doi.org/10.1016/j.jhydrol.2009.03.010.

[26] K. P. Burnhan & D. R. Anderson, *Model selection and multimodel inference*, Springer, New York, 2002. https://link.springer.com/book/10.1007/b97636.

[27] G. Schwarz, "Estimating the dimension of a model", The annals of statistics **6** (1978) 461. https://www.jstor.org/stable/2958889.

[28] K. P. Burnham & D. R. Anderson, "Multimodel inference: understanding AIC and BIC in model selection", Sociological methods & research **33** (2004) 261. https://doi.org/10.1177/0049124104268644.

[29] S. B. Sunday, N. S. Agog, P. Magdalene, A. Mubarak & G. K. Anyam, "Modeling extreme rainfall in Kaduna using the generalised extreme value distribution", Science World Journal **15** (2020) 73. https://www.ajol.info/index.php/swj/article/view/203031.

[30] A. Robson & D. Reed, *Flood estimation handbook: statistical procedures for flood frequency estimation*, Institute of Hydrology, 1999. .

[31] F. Laio, Cramer–von Mises and Anderson-Darling goodness of fit tests for extreme value distributions with unknown parameters. Water Resources Research **40** (2004) W09308. https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2004WR003204.

[32] G. E. Box, G. M. Jenkins, G. C. Reinsel & G. M. Ljung, *Time series analysis: forecasting and control, 5th Edition*, John Wiley & Sons, 2015. https://www.wiley.com/en-us/Time+Series+Analysis%3A+Forecasting+and+Control%2C+5th+Edition-p-9781118675021.