Zero : Jurnal Sains, Matematika, dan Terapan
E-ISSN : 2580-5754; P-ISSN : 2580-569X
Volume 9, Number 2, 2025
DOI: 10.30829/zero.v9i2.25688
Page: 324-334

Evaluation of Best-Fit Probability Distribution Models for Monthly Rainfall in the Lake Toba Region

[1] **Syukri Arif Rafhida**
School of Data Science, Mathematics, and Informatics, IPB University, Bogor, Indonesia

[2] **Sri Nurdiati**
School of Data Science, Mathematics, and Informatics, IPB University, Bogor, Indonesia

[3] **Retno Budiarti**
School of Data Science, Mathematics, and Informatics, IPB University, Bogor, Indonesia

[4] **Mohamad Khoirun Najib**
School of Data Science, Mathematics, and Informatics, IPB University, Bogor, Indonesia

Article Info

Article history:

Accepted, 30 September 2025

Keywords:

Hydrological Modeling;
Kolmogorov–Smirnov Test;
Lake Toba;
Probability Distribution;
Rainfall.

ABSTRACT

Understanding rainfall's statistical distribution is vital for water resource management, disaster mitigation, and climate adaptation in tropical regions. This study evaluates the best-fit probability distributions for monthly rainfall in the Lake Toba region, Indonesia, using data from 34 rain gauge stations in 1972–2017 period. Ten distributions were tested, with parameters estimated by Maximum Likelihood Estimation (MLE) and model performance assessed using the Kolmogorov–Smirnov (KS) test. Results show that Generalized Extreme Value (GEV), Gamma, and Weibull distributions consistently provide the best fit for most stations and regencies, while Exponential and Inverse Gaussian perform poorly. Spatial analysis reveals variation in model suitability among regencies, influenced by local topography and microclimate. These findings highlight the importance of flexible models for hydrological planning and climate risk assessment. The study also provides valuable references for rainfall modeling and bias correction in other tropical regions.

Corresponding Author:

Sri Nurdiati,
School of Data Science, Mathematics, and Informatics.
IPB University, Bogor, Indonesia
Email: nurdiati@apps.ipb.ac.id

1. INTRODUCTION

The Lake Toba region possesses unique geographical and climatological characteristics. It is the largest volcanic and tectonic lake in Indonesia, formed by a supervolcanic eruption tens of thousands of years ago [1].

Climatologically, the area lies in the equatorial zone and exhibits a typical equatorial climate, with two wet seasons and two dry seasons each year [2]–[4]. As a result, rainfall in this region is relatively high and well-distributed throughout the year, supporting vital sectors such as agriculture, food security, tourism, as well as water and electricity supply for surrounding communities [5]. The region's topography is also highly complex, as it lies around a large caldera and is surrounded by mountains of varying elevations, further influencing local wind circulation patterns and cloud formation. In this physiographic and climatic setting, understanding the behavior of rainfall becomes pivotal for both environmental processes and socio-economic outcomes.

Rainfall is one of the most important climatological parameters, playing a vital role in various environmental and socio-economic aspects, such as water resource management [6], agriculture [7], urban planning [8], and flood disaster mitigation [9]. The statistical distribution of rainfall data is crucial in long-term hydrological planning, as knowledge of this distribution allows estimation of rainfall-event probabilities in a given area—particularly for Lake Toba, where the caldera's complex topography and distinct wet–dry seasonality intensify spatial variability and extremes [10]. Several studies have aimed to identify the most appropriate (best-fit) probability distribution models, which is essential for understanding the spatio-temporal characteristics of rainfall in a region [11]–[13].

Numerous previous studies have evaluated the most suitable probability distributions for rainfall data in various regions worldwide. Yusof et al. [11] investigated hourly rainfall data in the Federal Territory of Malaysia using Exponential, Gamma, Weibull, and Mixed-Exponential distributions, and found that mixed distributions were more appropriate for data dominated by light rainfall with occasional extreme events. This research highlighted the importance of selecting distribution models based on comprehensive goodness-of-fit tests.

Alam et al. [12] analyzed maximum monthly rainfall data in Bangladesh using Generalized Extreme Value (GEV), Pearson Type III, and Log-Pearson Type III distributions, and applied three statistical tests, including Kolmogorov–Smirnov (K–S), Anderson–Darling (A–D), and Root Mean Square Error (RMSE). Their results indicated that the GEV distribution was most frequently the best-fit model at more than one-third of the observed stations. This study demonstrates the importance of considering extreme distributions when modeling maximum rainfall.

Another study by Ximenes et al. [13] in Northeastern Brazil evaluated six two-parameter distributions (Gamma, Weibull, Log-Normal, Generalized Pareto, Gumbel, and Normal) using data from 293 stations, and determined that the Gamma and Weibull distributions provided the best performance for monthly rainfall data in semi-arid regions. Their approach involved a modification of the Shapiro-Wilk statistic (TN. SW) as the basis for model selection. This demonstrates that two-parameter distributions can be sufficiently flexible when combined with appropriate parameter estimation methods.

Although these studies that have used diverse computational environments, such as R packages, and MATLAB toolboxes, have made significant contributions to the understanding of rainfall distributions in various regions, a substantial research gap remains, particularly in the Lake Toba region, Indonesia. Previous studies in this area [14], [15] have identified the best-fit distributions, but their primary focus was on bias correction processes, and comprehensive spatial analyses of rainfall distributions in each sub-region are still lacking.

Previous studies in this area have utilized various distributions such as Generalized Extreme Value, Normal, Weibull, Gamma, Logistic, Log-Normal, Log-Logistic, and Inverse Gaussian to identify monthly rainfall and temperature distributions [14]. Because data resolution (daily vs. monthly vs. seasonal) materially affects the suitability of a probability model, it is important to note that Lake Toba's monthly rainfall totals contain very few zeros and exhibit only moderate right-skew. Accordingly, distributions without a point mass at zero and with moderate skewness (e.g., Gamma, Log-Normal, Logistic, Log-Logistic, Weibull) are generally more appropriate for this setting. Another study [15] evaluated ten types of distributions, similar to those used in [16], [17], including extreme and exponential distributions. No prior study has comprehensively analyzed monthly rainfall distribution across all regencies in Lake Toba using multiple probability models and spatial analysis.

This study aims to conduct a basin-wide evaluation of ten widely used probability distributions for monthly rainfall using observations from 34 rain-gauge stations in the Lake Toba region. Models are fitted by maximum likelihood and appraised with the Kolmogorov–Smirnov test to determine the best-fit distribution for each station–month combination. Spatial variation in best-fit models is examined across regencies and interpreted in relation to topographic and climatic controls. The resulting evidence base provides practical guidance for rainfall modeling and bias correction and supports climate-adaptation planning in Lake Toba and in other tropical regions with comparable geographical settings.

2. RESEARCH METHOD

2.1 Study Area and Datasets

Lake Toba is the largest volcanic lake in Southeast Asia [5], located in North Sumatra Province, Indonesia. The lake spans seven regencies, namely Karo (KAR), Simalungun (SIM), Dairi (DAI), Toba (TOB), Samosir (SAM), North Tapanuli (TAP), and Humbang Hasundutan (HUM). Geographically, the lake is situated between 98° 31' 2"–98° 9' 14" East Longitude and 2° 19' 15"–2° 54' 2" North Latitude. Lake Toba has a surface area of approximately 1,124 km², with a length of about 50 km, a width of about 27 km, and an average depth reaching 228 meters.

The topography surrounding the lake is dominated by hills and mountains belonging to the Bukit Barisan Mountain range, creating a complex morphological environment that influences local cloud formation and the distribution of rainfall. Figure 1 shows the mean and distribution of monthly rainfall in the Lake Toba region based on all available data from each station. The average monthly rainfall in this area generally exceeds 150 mm per month, with two main peaks occurring in April and October–December, indicating a bimodal rainfall pattern characteristic of equatorial regions. The range between the mean plus or minus one standard deviation (dashed lines) indicates considerable interannual variability, particularly during the second wet season. This pattern is consistent with the findings of [2], which showed that equatorial regions like Lake Toba receive relatively high monthly rainfall throughout the year, with less variation during the dry season compared to monsoon climates. This pattern is influenced by the dynamics of the Intertropical Convergence Zone (ITCZ), which moves seasonally in response to the position of the sun [18]. The combination of equatorial geographic position, ITCZ influence, and complex topography results in high and evenly distributed rainfall throughout the year in this region [2]–[4]. This makes Lake Toba a region with significant water potential and an important role in supporting agriculture, tropical forest ecosystems, as well as water and energy supply for the surrounding communities.

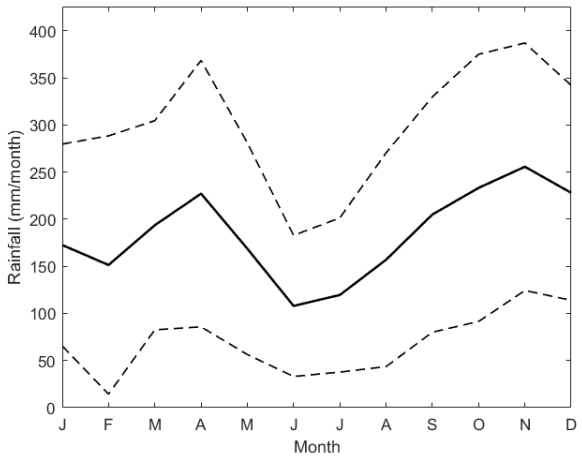

Figure 1. Annual cycle of monthly total rainfall Lake Toba. Solid lines represent the mean of monthly rainfall (mm/month) in all data period used in study, while dashed lines represent the standard deviation (mm/month).

The data used in this study consist of monthly rainfall records from 34 stations around Lake Toba, obtained from the North Sumatra Meteorological, Climatological, and Geophysical Agency (BMKG) in units of mm/month. All available monthly rainfall data recorded by BMKG were included in the analysis, without standardizing the observation periods across stations. Missing values were handled in a listwise, month-specific manner: any month flagged as null for a given station was omitted from that station's analysis for that month only, without imputation or interpolation. This approach prevents missing entries from influencing other months and avoids injecting distributional assumptions prior to model fitting. The data periods for each station vary, ranging from 1972 to 2017, depending on the data availability at each location. The locations of the stations are shown in Figure 2.

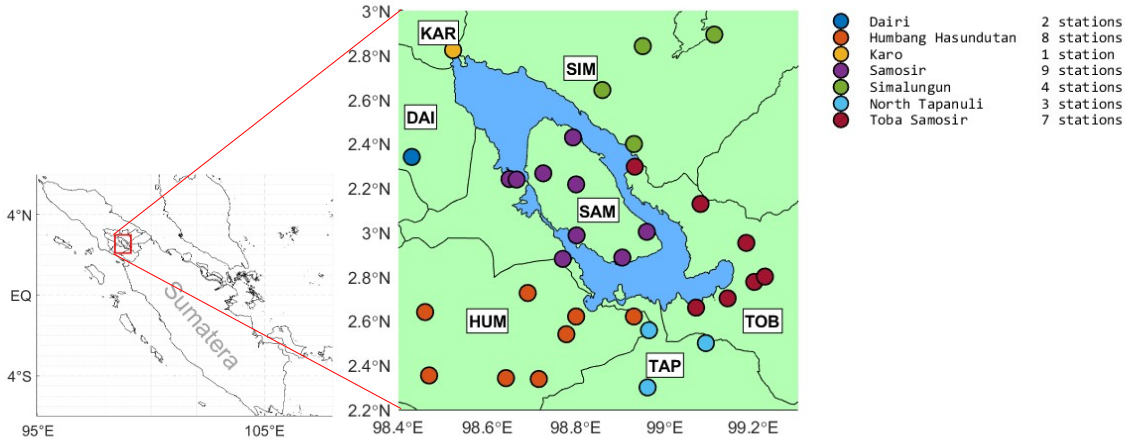

Figure 2. Location of 34 rainfall stations in the Lake Toba region, North Sumatra, Indonesia.

Figure 2 shows the spatial distribution of stations across seven regencies, represented by different colors and abbreviations: Dairi (DAI), Humbang Hasundutan (HUM), Karo (KAR), Samosir (SAM), Simalungun (SIM), North Tapanuli (TAP), and Toba Samosir (TOB). The numbers in the legend indicate the total stations available in each regency. The left panel provides the geographic context of the Lake Toba basin within Sumatra Island. To provide an overview of the rainfall data analyzed, Table 1 presents the basic statistical summary of 34 rain gauge stations in the Lake Toba region.

Table 1. Summary statistics of monthly rainfall at 34 stations in the Lake Toba region, including observation period, mean ± standard deviation, minimum, and maximum values.

Station	Region	Period	Mean and Std	Min	Max
Lae Hole	DAI	1995-2017	191.6 ± 119.2	8	767
Sitinjo	DAI	1981-2017	195.0 ± 190.4	9	3030
Baktiraja	HUM	2010-2017	145.9 ± 121.3	5	617
Dolok Sanggul	HUM	1973-2017	179.2 ± 113.8	3	828
Onanganjang	HUM	2010-2017	280.1 ± 181.6	19	993
Pakkat	HUM	2007-2017	273.8 ± 168.5	4	781
Paranginan	HUM	2010-2017	147.2 ± 81.4	21	337
Parlilitan	HUM	1991-2017	317.8 ± 164.0	14	897
Pollung	HUM	1998-2017	202.4 ± 103.8	15	594
Sijamapolang	HUM	2010-2017	177.7 ± 110.0	18	487
Merek	KAR	1973-2017	174.0 ± 110.8	7	916
Gabe Hutaraja	TAP	1986-2017	179.8 ± 103.5	2	562
Muara	TAP	1996-2017	218.3 ± 151.2	35	1350
Siborong-borong	TAP	1974-2017	196.6 ± 143.4	2	1521
Harian	SAM	1984-2017	168.2 ± 161.5	0	1830
Nainggolan	SAM	1998-2017	171.2 ± 102.5	3	540
Onan Runggu	SAM	1981-2017	166.4 ± 101.7	7	510
Palipi	SAM	2006-2017	145.0 ± 84.8	1	332
Pangururan	SAM	1973-2017	154.5 ± 99.5	0	706
Ronggur Nihuta	SAM	2006-2017	152.0 ± 90.3	4	475
Sianjur Mula-mula	SAM	2007-2017	233.7 ± 228.6	14	1840
Simanindo	SAM	1998-2017	194.3 ± 105.8	4	613
Sitio-Tio	SAM	2006-2017	176.5 ± 101.1	0	531
Marjandi	SIM	2010-2017	224.3 ± 113.7	25	492
SMPK Marihat	SIM	1971-2017	246.6 ± 121.9	6	819
Sidamanik	SIM	1974-2017	226.5 ± 127.6	4	894
Stageof Parapat	SIM	1973-2017	172.8 ± 92.7	2	577
Ajibata	TOB	2006-2017	109.6 ± 100.4	0	429
Balige	TOB	1973-2017	140.5 ± 88.3	2	460
Laguboti	TOB	1973-2017	142.6 ± 87.6	2	513
Lumban Julu	TOB	1984-2017	217.1 ± 120.2	4	647
Porsea	TOB	1995-2016	188.6 ± 115.1	2	728
Sigumpar	TOB	2006-2017	166.3 ± 94.4	16	502
Silaen	TOB	1972-2017	145.3 ± 88.0	2	461

Overall, as in table 1, the monthly rainfall means across the stations range from about 140 to 320 mm, with relatively large standard deviations, reflecting substantial interannual variability in the Lake Toba basin. Minimum values at several stations even reach 0 mm, indicating occasional dry months despite the equatorial climate. In contrast, maximum values vary between 500 and 1,800 mm/month for most stations, although one extreme record of 3,030 mm/month was observed at Sitinjo Station (DAI). Such an unusually high value is likely an outlier or a recording/unit error, since monthly rainfall of this magnitude is rarely observed in North Sumatra. This highlights the importance of data quality control before applying probability distribution analysis.

2.2. Rainfall Distribution Modeling

In this study, the selection of the best-fit probability distribution for monthly rainfall data was conducted through parameter estimation and model goodness-of-fit evaluation. This process was systematically performed using MATLAB sofware for each month at every station to determine the most appropriate distribution model in accordance with the characteristics of the rainfall data in the study area.

Various probability distribution models commonly used in hydrology and climatology studies were applied to the monthly rainfall data. Parameter estimation for each distribution was carried out using the Maximum Likelihood Estimation (MLE) method, which is recognized as an efficient and consistent approach for obtaining

distribution parameters from observational data [19]. This process was performed separately for each month and station, resulting in specific distribution parameters for every month–station combination, according to the characteristics of the data. The selection of distributions was based on previous literature comparing the performance of various distributions for monthly precipitation data [15], [16]. The probability density functions of each distribution are presented in Table 2.

Table 2. Distribution used in study

Distribution Name	Probability Density Function	Domain & Parameter
Extreme Value (EV)	$f(x \mid \mu,\sigma) = \frac{1}{\sigma} e^{-\left(\frac{x-\mu}{\sigma}\right)} e^{-e^{-\left(\frac{x-\mu}{\sigma}\right)}}$	$x \in \mathbb{R}; \sigma > 0, \mu \in \mathbb{R}$
Generalized Extreme Value (GEV)	$f(x\|\mu,\sigma,\xi) = \frac{1}{\sigma} e^{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}} \left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi-1}$	$x \in \mathbb{R}; \sigma > 0, \mu \in \mathbb{R}, \xi \in \mathbb{R}$
Logistic (LOG)	$f(x\|\mu,s) = \frac{1}{s} \frac{e^{-(x-\mu)/s}}{(1+e^{-(x-\mu)/s})^2}$	$x \in \mathbb{R}; s > 0, \mu \in \mathbb{R}$
Normal (NOR)	$f(x\|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$x \in \mathbb{R}; \sigma > 0, \mu \in \mathbb{R}$
Exponential (EXP)	$f(x \mid \mu) = \frac{1}{\mu} e^{\left(-\frac{x}{\mu}\right)}$	$x \geq 0; \mu > 0$
Gamma (GAM)	$f(x\|\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}}$	$x > 0; \alpha > 0, \beta > 0$
Inverse Gaussian (ING)	$f(x \mid \mu, \lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} e^{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}}$	$x > 0; \mu > 0, \lambda > 0$
Log-Logistic (LL)	$f(x\|\alpha,\beta) = \frac{\beta}{x}\left(1+\left(\frac{x}{\alpha}\right)^{\beta+1}\right)^{-1}$	$x > 0; \alpha > 0, \beta > 0$
Log-Normal (LN)	$f(x\|\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(ln(x)-\mu)^2}{2\sigma^2}}$	$x > 0; \sigma > 0, \mu \in \mathbb{R}$
Weibull (WB)	$f(x\|\lambda,k) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}$	$x \geq 0; \lambda > 0, k > 0$

2.3. Kolmogorov-Smirnov test (KS)

After the parameters of each distribution were estimated using the Maximum Likelihood Estimation (MLE) method, the goodness-of-fit to the observed data was evaluated using the Kolmogorov–Smirnov (KS) test. The Kolmogorov–Smirnov (KS) test measures the maximum error of the cumulative distribution function (CDF), which is defined as

$$KS = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \tag{1}$$

where F_n and F represent the empirical and theoretical CDFs, respectively [20]. The KS test measures the maximum deviation between the empirical and theoretical cumulative distribution functions (CDFs), providing a non-parametric method for comparing continuous distributions. The KS test assesses the similarity between the distributions of the observed BMKG data, the CMIP6 model data, and the bias-corrected model data. The null hypothesis of the KS test assumes that both samples are drawn from the same distribution.

This study relied exclusively on the KS test because of its simplicity, broad applicability, and minimal assumptions. Unlike chi-square tests that require data binning, and Anderson–Darling tests that emphasize tail differences, the KS test directly compares entire CDFs without binning. Although Anderson–Darling can be more powerful in detecting discrepancies in the distribution tails, it also tends to require larger samples for reliable performance [21]. In particular, Engmann and Cousineau [22] demonstrated that while Anderson–Darling occasionally offers better sensitivity, the KS test remains competitive, especially in moderate or varied sample sizes typical of rainfall datasets. Moreover, hydrological studies often adopt the KS test due to its practicality across multiple sites with disparate record lengths and bias correction often used this method [23], [24].

3. RESULT AND ANALYSIS

In this section, the results of the goodness-of-fit analysis for ten probability distributions applied to monthly rainfall data in the Lake Toba region are presented. The evaluation was conducted using the one-sample Kolmogorov–Smirnov (KS) test to determine which distribution best models the rainfall for each month at all observation stations. The objective of this analysis is to identify the most suitable distribution model for each month

throughout the year, utilizing monthly rainfall data from 34 stations distributed across seven regencies in the Lake Toba region, according to the data availability at each station.

3.1 Goodness of fit from all distribution

The goodness-of-fit of ten probability distribution models to the monthly rainfall data in the Lake Toba region was evaluated using the p-values from the one-sample Kolmogorov–Smirnov (KS) test. The results of these tests are visualized in Figure 3, which displays the distribution of p-values for each distribution across all months and stations in the study area. Through this analysis, it is possible to identify which distributions are statistically best able to represent the variation in monthly rainfall data in the Lake Toba region.

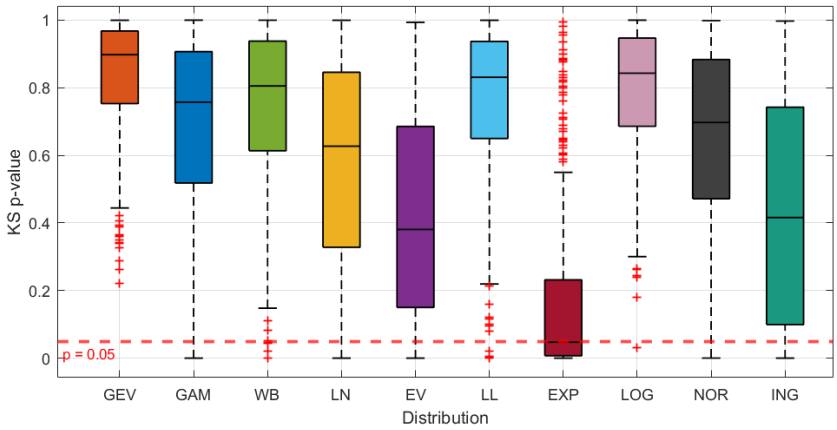

Figure 3. Boxplots of Kolmogorov–Smirnov (KS) p-values for ten candidate probability distributions fitted to monthly rainfall data from all stations in the Lake Toba region. A higher p-value indicates a better agreement between the theoretical distribution and the observed rainfall data, while the red dashed line at p = 0.05 marks the conventional significance threshold.

Figure 3 presents boxplots of the p-values from the Kolmogorov–Smirnov (KS) test for the ten probability distributions fitted to monthly rainfall data across all stations and months. The p-values from the KS test are used to determine whether the null hypothesis (H_0) can be accepted, i.e., that the observed data are drawn from the same distribution as the tested model. In this study, the significance threshold used is 0.05; thus, if the p-value exceeds this threshold, the distribution is considered not significantly different from the observed data. It showed in red dashed line in Figure 3. The higher the p-value, the better the distribution represents the empirical data, as it reduces the likelihood of rejecting the null hypothesis.

From Figure 3, it can be seen that three main distributions—Generalized Extreme Value (GEV), Gamma, and Weibull—consistently display high median p-values with relatively narrow spreads. This indicates that these three distributions are, in general, highly suitable for representing the monthly rainfall data characteristics in the Lake Toba region. Conversely, the Exponential and Inverse Gaussian distributions exhibit the lowest median p-values, often falling below the common significance threshold in red dashed line ($p < 0.05$), indicating that these models frequently fail to represent the data adequately.

These findings are consistent with the study by [25] in Japan, which also found that the GEV, Gamma, and Weibull distributions (including three-parameter variants) are the main candidates providing the best fit for monthly and annual rainfall data, while the Exponential distribution is generally inadequate, especially due to its inability to capture data variability during dry seasons and extremes during wet seasons. Additionally, the study by [26] confirmed that the Gamma and related distributions are physically capable of explaining the characteristics of daily and monthly rainfall intensity distributions across various latitudes, both tropical and subtropical, thus supporting the high goodness-of-fit observed for Gamma-type distributions in Lake Toba.

The low p-values for the Exponential distribution can be explained by the fact that this model has only a single scale parameter and lacks the flexibility to capture the diversity of empirical distributions, whether at the lower tail (dry days) or upper tail (extreme events), making it too simplistic for complex monthly rainfall data. In contrast, GEV, Gamma, and Weibull distributions offer greater flexibility (with two or three parameters), allowing them to accommodate the various forms of data distributions produced by climatic variability and meteorological processes in tropical mountainous regions such as Lake Toba.

3.2 Monthly Distribution

After obtaining the p-values from the Kolmogorov–Smirnov test, the next step was to determine the most suitable probability distribution for each month, based on the highest p-value at each station. The results of the

best-fit distribution selection for each month are presented in Figure 4, which displays the five distributions with the highest frequencies for each month throughout the year across all observation stations.

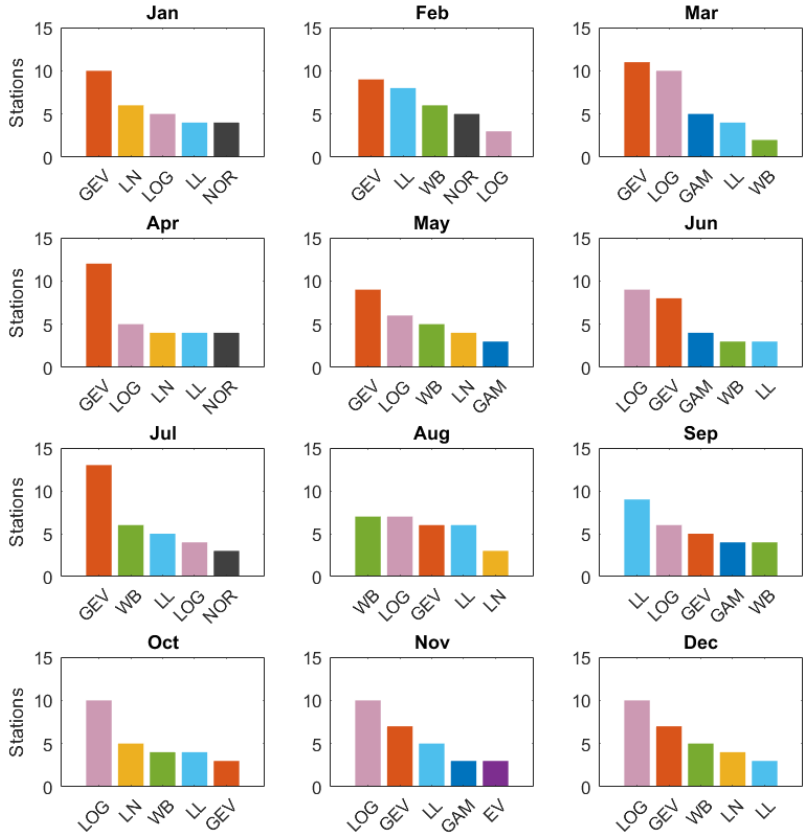

Figure 4. Frequencies of the five best-fit probability distributions (based on the highest Kolmogorov–Smirnov p-value) for each month across all stations in the Lake Toba region.

Figure 4 reveals a clear seasonal pattern in the dominance of certain probability distribution models. During the main wet season, particularly from January to May, the Generalized Extreme Value (GEV) distribution consistently emerges as the most dominant best-fit model at most stations. This pattern is consistent with the findings reported by [12], [13], where the GEV distribution is widely used to represent extreme monthly rainfall data in tropical and subtropical regions. In addition to GEV, the Logistic and Log-Normal distributions also frequently appear as the best-fit models during the early part of the season, reflecting the flexibility of these models in capturing the variability during periods of high and variable rainfall.

During the dry season months (June to September), there is a shift in dominance towards the Logistic (LOG), Log-Logistic (LL), and Weibull (WB) distributions, while GEV becomes less frequent as the best-fit model. This indicates that the form and spread of monthly rainfall data during the dry season are better represented by distributions that can capture lower and more dispersed data patterns (such as Log-Logistic and Weibull), as also identified in the study by [13] in North-eastern Brazil. In transition months such as October and November, the Logistic distribution once again dominates, indicating a seasonal cycle in the preference of probability models for rainfall in the Lake Toba region. A more detailed summary of the monthly frequency of best-fit probability distributions across all stations is provided in Table 3, which complements the seasonal patterns illustrated in Figure 4.

Table 2. Monthly frequency of best-fit probability distributions fitted to rainfall data from all stations in the Lake Toba region. The values indicate how many times each distribution was selected as the best fit in a given month, with the bottom row showing the overall totals across all months.

Month	GEV	GAM	WB	LN	EV	LL	EXP	LOG	NOR	ING
January	10	3	1	6	1	4	0	5	4	0
February	9	0	6	0	1	8	0	3	5	2
March	11	5	2	1	0	4	0	10	1	0
April	12	2	3	4	0	4	0	5	4	0
May	9	3	5	4	2	2	0	6	2	1
June	8	4	3	1	1	3	3	9	2	0

July	13	1	6	1	1	5	0	4	3	0
August	6	2	7	3	0	6	0	7	2	1
September	5	4	4	2	2	9	0	6	1	1
October	3	3	4	5	2	4	0	10	3	0
November	7	3	2	1	3	5	0	10	3	0
Total	100	31	48	32	15	57	3	85	32	5

3.3 Spatial Variation of Best-Fit Rainfall Distributions Across Regencies

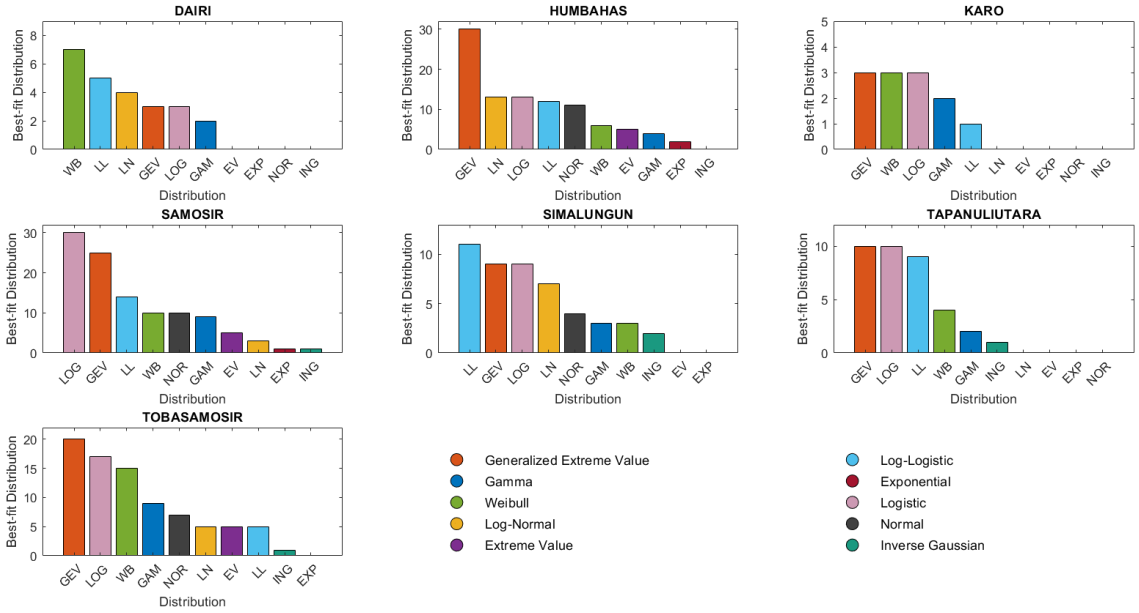

Figure 5. Best-fit probability distributions for monthly rainfall data in each regency surrounding Lake Toba.

Figure 5 displays the frequency distribution of probability models most often selected as the best-fit based on the evaluation results across all stations and months for each regency around Lake Toba. In this figure, each bar represents the number of times a probability distribution was chosen as the best model to represent monthly rainfall data for all combinations of months and stations within the regency. This visualization clearly shows that no single distribution absolutely dominates across all regencies. Instead, each regency exhibits unique patterns in best-fit distribution tendencies, indicating significant spatial variability in the statistical characteristics of monthly rainfall in the Lake Toba region.

For example, Humbang Hasundutan Regency stands out with the Generalized Extreme Value (GEV) distribution most frequently selected as the best-fit model, far surpassing other models, followed by Log-Normal (LN), Logistic (LOG), and Normal (NOR) distributions. Similar patterns are observed in North Tapanuli and Toba Samosir Regencies, where GEV consistently emerges as the primary model of choice, indicating that this distribution is highly suitable for representing rainfall data in areas dominated by mountainous terrain and high levels of rainfall extremes.

On the other hand, Samosir Regency demonstrates a different tendency, with the Logistic (LOG) distribution becoming the most dominant best-fit model, even surpassing GEV and other models. Interestingly, Log-Logistic (LL) and Normal (NOR) distributions are also frequently chosen in Samosir. The frequent selection of the Normal (NOR) distribution as a best-fit model in Samosir is noteworthy, as the normal distribution is inherently symmetric and is typically not expected to fit rainfall data, which are generally positively skewed. This phenomenon may be attributed to the unique climatic and geographic setting of Samosir Regency. Due to the moderating effect of Lake Toba, as well as the relatively consistent seasonal rainfall patterns and the potential dampening of extreme precipitation events by the surrounding water body, the monthly rainfall data at certain stations may exhibit lower skewness and a more symmetrical distribution. Additionally, aggregation over monthly timescales can reduce the influence of extreme events and the central limit theorem may act to further normalize the data, making the normal distribution statistically adequate in some cases. Simalungun also exhibits a more diverse best-fit pattern, with Log-Logistic (LL), GEV, Logistic (LOG), and Log-Normal (LN) each frequently serving as the best-fit model, indicating the presence of diverse rainfall characteristics in the regency.

Dairi Regency shows characteristics distinct from the others, with the Weibull (WB) and Log-Normal (LN) distributions more frequently selected as the best-fit models. This may be related to topographic conditions,

geographic location, and the more limited distribution of observation stations, resulting in more uniform or less extreme rainfall patterns being recorded. Conversely, Karo Regency exhibits a more "even" distribution of best-fit models among several distributions, such as GEV, Weibull, Logistic, and Gamma, with no single model truly dominating, suggesting that rainfall variability in Karo is strongly influenced by a combination of topographic factors, relative position to Lake Toba, and local wind dynamics.

In addition to the dominant models, some distributions such as Exponential (EXP) and Inverse Gaussian (ING) are rarely selected as the best-fit in any regency, consistent with earlier p-value analysis results showing poor fit of these models for monthly rainfall data in tropical mountainous areas like Lake Toba.

This variation in probability distribution model selection among regencies is closely related to topographical characteristics and local rainfall patterns, as described in [27]. Their study highlighted that topographical diversity in North Sumatra which includes coastal lowlands, slopes, mountains, and island areas—creates significant spatial variability in rainfall. Their spatial analysis showed, for example, that mountainous and sloped areas (which cover much of Lake Toba and its surroundings) have rainfall intensities different from those of lowland or coastal areas. The complex topography results in more complicated wind circulation and cloud formation patterns, thus causing the statistical characteristics of monthly rainfall data to vary among regencies.

Statistically, the tendency for best-fit models to favor the GEV, Logistic, and Weibull distributions in this region is consistent with the findings of Prasetyo et al. [23], who found that spatial variation in monthly and annual rainfall is greatly influenced by relative position to the Bukit Barisan, distance to the western coast, and elevation. The GEV and Weibull models are known to be flexible for capturing extreme rainfall events and the bimodal patterns frequently encountered in tropical highlands, while the dominance of Logistic and Log-Logistic distributions in some regencies reflects the high inter-station and inter-seasonal variability within those areas.

3.4 Discussion

This study makes an important contribution to monthly rainfall modeling in the Lake Toba region by evaluating the goodness-of-fit of ten probability distributions using observational data from 34 BMKG stations. The selection of appropriate distributions is crucial, particularly for water resource planning, disaster mitigation, and the development of agriculture based on local climate characteristics. The correct probability distribution can improve the accuracy of extreme rainfall predictions, return period calculations, and the design of hydrometeorological risk-based infrastructure.

The results show that the Generalized Extreme Value (GEV), Gamma, and Weibull distributions consistently provide the best fit for monthly rainfall data around Lake Toba, as indicated by high KS p-values and frequent selection as the best distribution in the top-five analysis for each month and regency. These findings are consistent with similar studies in other regions, such as in Bangladesh [12], which identified GEV, Pearson Type III, and Log-Pearson Type III as the best distributions for maximum monthly rainfall at most stations. Similarly, in Brazil, the study by [13] found Gamma and Weibull superior for monthly rainfall in semi-arid regions, while research in Europe has recommended GEV for flood and hydrological characteristics at various scales [28].

Meanwhile, the Exponential (EXP) distribution consistently performed poorly, both in terms of p-value and frequency as the best-fit model. This result aligns with various international studies that have shown the limitations of two-parameter distributions, especially the Exponential, in capturing the empirical nature of rainfall data, which is often highly variable and "heavy-tailed." The main reason for the Exponential distribution's inadequacy is its lack of flexibility in describing data variation and its inability to model the high skewness and kurtosis found in rainfall data. A study in Malaysia [11] even found that the Exponential distribution is only suitable for data dominated by light rainfall, but tends to fail when the data contain many extreme events, making it preferable to use mixture distributions or those with more parameters.

When compared to studies in other tropical and subtropical regions, the tendency for GEV, Gamma, and Weibull to emerge as best-fit distributions appears to be quite universal for monthly rainfall data, particularly in equatorial regions with high year-round rainfall. These three distributions are widely used and are the preferred choices in many studies [14], [23], [29], [30]. A study in India [29] demonstrated the superiority of the GEV distribution in representing extreme rainfall and temperature, both in observational data and climate model outputs, due to its ability to fit a variety of rainfall distribution shapes in diverse environments, from lowlands to mountainous areas. In Europe, research by [30] confirmed that GEV is the primary choice for extreme rainfall frequency analysis, especially for data with heavy distribution tails, and that Weibull and Gamma also perform very well on monthly and seasonal scales. Overall, these findings reinforce that the use of flexible distributions capable of capturing extremes is highly recommended for monthly rainfall analysis in both tropical and subtropical regions.

Furthermore, accurate selection of probability distributions is particularly important for bias correction in model or reanalysis rainfall data, as demonstrated by [15]. That study emphasized that statistically selecting the most appropriate distribution model prior to bias correction directly influences the quality of the correction results. A good probability distribution allows bias correction methods such as Quantile Mapping, Quantile Delta Mapping, or similar approaches to function optimally, thus minimizing systematic errors and improving the

accuracy of the corrected data. Overall, the findings of this research are not only useful for the Lake Toba and North Sumatra regions but are also relevant to other tropical regions with similar hydrometeorological conditions. The methodology and results can serve as a reference for the development of climate adaptation strategies and statistically based hydrometeorological disaster risk management.

Beyond their statistical relevance, the findings of this study have practical implications for regional water management. The identification of GEV, Gamma, and Weibull as the most suitable distributions supports their application in flood risk modeling, where reliable estimation of return periods for extreme rainfall is essential for the design of drainage systems, embankments, and other flood control measures. In agriculture, these distributions can improve planning for cropping calendars by identifying the likelihood of wet and dry months, thereby reducing vulnerability to climate variability. Similarly, in reservoir management, accurate rainfall probability models are critical for anticipating inflows, optimizing storage, and ensuring water availability for both irrigation and domestic use in the Lake Toba basin. Thus, the selection of appropriate distributions is not merely a statistical exercise but a key step toward improving resilience in climate-sensitive sectors.

A limitation of this study is that the evaluation of distribution fit relied solely on KS test p-values compared against a fixed threshold. This approach, while widely used, may risk overgeneralization. Incorporating confidence intervals for parameter estimates or test statistics could provide a more nuanced assessment of model suitability, which is recommended for future research.

4. CONCLUSION

This study comprehensively evaluated ten probability distribution models to determine the best-fit models for monthly rainfall data across 34 stations in the Lake Toba region. Using the Kolmogorov–Smirnov goodness-of-fit test, the results consistently showed that the Generalized Extreme Value (GEV), Gamma, and Weibull distributions provided superior performance in representing the statistical characteristics of monthly rainfall, both overall and at the regency level. Spatial analysis revealed significant variability in best-fit distributions between regencies, closely related to differences in topography and local rainfall patterns.

Beyond methodological contributions, the results have strong practical value for water resources and agriculture. Station- and regency-specific best-fit distributions support return-period estimation for flood risk, seasonal water allocation and reservoir operation, and planning of cropping calendars under bimodal rainfall. Importantly, our findings can be operationalized for bias correction of climate model outputs. Quantile mapping or related methods that use distribution can be aligned with the empirically selected monthly families and parameters to correct CMIP6 precipitation projections available for the Lake Toba area, thereby improving the credibility of downscaled projections used in adaptation and resource planning.

Future research should move beyond stationarity by considering non-stationary models with time-varying parameters, and exploring mixture or copula-based approaches to capture joint behavior, such as amount–duration–intensity at a site and multi-site dependence across stations. In addition, a promising direction is integration with machine-learning rainfall generators (e.g., LSTM/sequence models, GANs, diffusion models), potentially conditioned on climate indices (ENSO/IOD) and informed by the best-fit distributional families identified here.Application of the identified best-fit models in operational hydrological forecasting, infrastructure design, and disaster risk management would ensure that the statistical advances achieved here translate into tangible societal benefits, ultimately supporting resilience to climate variability in Indonesia and other tropical regions.

5. REFERENCES

[1] C. A. Chesner, "The Toba Caldera Complex," *Quat. Int.*, vol. 258, pp. 5–18, 2012, doi: 10.1016/j.quaint.2011.09.025.

[2] E. Aldrian and D. R. Susanto, "Identification of three dominant rainfall regions within Indonesia and their relationship to sea surface temperature," *Int. J. Climatol.*, vol. 23, pp. 1435–1452, 2003, doi: 10.1002/joc.950.

[3] E. Hermawan, "Pengelompokkan Pola Curah Hujan Yang Terjadi Di Beberapa Kawasan P. Sumatera Berbasis Hasil Analisis Teknik Spektral," *J. Meteorol. dan Geofis.*, vol. 11, no. 2, 2010, doi: 10.31172/jmg.v11i2.67.

[4] S. Nurdiati, E. Khatizah, M. K. Najib, and R. R. Hidayah, "Analysis of rainfall patterns in Kalimantan using fast fourier transform (FFT) and empirical orthogonal function (EOF)," *J. Phys. Conf. Ser.*, vol. 1796, no. 1, p. 12053, Feb. 2021, doi: 10.1088/1742-6596/1796/1/012053.

[5] H. Sihotang, M. Y. J. Purwanto, W. Widiatmaka, and S. Basuni, "Model for Water Conservation of Lake Toba," *J. Nat. Resour. Environ. Manag.*, vol. 2, no. 2, pp. 65–72, 2012, doi: 10.19081/jpsl.2012.2.2.65.

[6] Z. Kundzewicz *et al.*, "Freshwater Resources and their Management," 2007, pp. 173–210.

[7] J. Shortridge, "Observed trends in daily rainfall variability result in more severe climate change impacts to agriculture," *Clim. Change*, vol. 157, no. 3, pp. 429–444, 2019, doi: 10.1007/s10584-019-02555-x.

[8] W. Zhang, J. Yang, L. Yang, and D. Niyogi, "Impacts of City Shape on Rainfall in Inland and Coastal Environments," *Earth's Futur.*, vol. 10, no. 5, p. e2022EF002654, 2022, doi: https://doi.org/10.1029/2022EF002654.

[9] M. Jehanzaib, M. Ajmal, M. Achite, and T.-W. Kim, "Comprehensive Review: Advancements in Rainfall-Runoff Modelling for Floo d Mitigation," *Climate*, vol. 10, no. 10, 2022, doi: 10.3390/cli10100147.

[10] Z. ŞEN and A. L. I. G. ELJADID, "Rainfall distribution function for Libya and rainfall prediction," *Hydrol. Sci. J.*, vol. 44, no. 5, pp. 665–680, 1999, doi: 10.1080/02626669909492266.

[11] F. Yusof, Z. Mohd Daud, V.-T.-V. Nguyen, S. S. Syed Jamaludin, and Z. Yusop, "Fitting the best-fit distribution for the hourly rainfall amount in the Wilayah Persekutuan," 2007.

[12] M. A. Alam, K. Emura, C. Farnham, and J. Yuan, "Best-Fit Probability Distributions and Return Periods for Maximum Monthly Rainfall in Bangladesh," *Climate*, vol. 6, no. 1, 2018, doi: 10.3390/cli6010009.

[13] P. Ximenes, A. Silva, F. Ashkar, and T. Stosic, "Best-fit probability distribution models for monthly rainfall of Northeastern Brazil," *Water Sci. Technol.*, vol. 84, no. 6, pp. 1541–1556, 2021, doi: 10.2166/wst.2021.304.

[14] H. Irwandi, M. S. Rosid, and T. Mart, "Effects of Climate change on temperature and precipitation in the Lake Toba region, Indonesia, based on ERA5-land data with quantile mapping bias correction," *Sci. Rep.*, vol. 13, no. 1, pp. 1–11, 2023, doi: 10.1038/s41598-023-29592-y.

[15] S. A. Rafhida, S. Nurdiati, R. Budiarti, and M. K. Najib, "Bias correction of lake Toba rainfall data using quantile delta mapping," *CAUCHY*, vol. 9, no. 2, pp. 297–309, 2024, doi: 10.18860/ca.v9i2.29124.

[16] S. Nurdiati, A. Sopaheluwakan, and M. K. Najib, "Statistical Bias Correction for Predictions of Indian Ocean Dipole Index With Quantile Mapping Approach," *Int. MIPAnet Conf. Sci. Math.*, no. April 2021, 2019, doi: 10.31219/osf.io/7dq2j.

[17] M. K. Najib and S. Nurdiati, "Koreksi Bias Statistik Pada Data Prediksi Suhu Permukaan Air Laut Di Wilayah Indian Ocean Dipole Barat Dan Timur," *Jambura Geosci. Rev.*, vol. 3, no. 1, pp. 9–17, 2021, doi: 10.34312/jgeosrev.v3i1.8259.

[18] Tukidi, "Karakter Curah Hujan Di Indonesia," *J. Geogr.*, vol. 7, no. 2, pp. 136–145, 2010, [Online]. Available: http://journal.unnes.ac.id/nju/index.php/JG/article/view/84

[19] S. E. Fienberg and A. Rinaldo, "Maximum likelihood estimation in log-linear models," *Ann. Stat.*, vol. 40, no. 2, pp. 996–1023, Apr. 2012, doi: 10.1214/12-AOS986.

[20] A. Justel, D. Pefia, and R. Zamar, "˜!. STATISTICS& PROBABILITY LETTERS A multivariate Kolmogorov-Smimov test of goodness," *Stat. Probab. Lett.*, vol. 35, pp. 251–259, 1997.

[21] X. Zeng, D. Wang, and J. Wu, "Evaluating the Three Methods of Goodness of Fit Test for Frequency Analysis," *J. Risk Anal. Cris. Response*, vol. 5, p. 178, Oct. 2015, doi: 10.2991/jrarc.2015.5.3.5.

[22] S. Engmann and D. Cousineau, "Comparing distributions: the two-sample Anderson–Darling test as an alternative to the Kolmogorov–Smirnov test," *J. Appl. Quant. Methods*, vol. 6, pp. 1–17, Sep. 2011.

[23] A. J. Cannon, S. R. Sobie, and T. Q. Murdock, "Bias correction of GCM precipitation by quantile mapping: How well do methods preserve changes in quantiles and extremes?," *J. Clim.*, vol. 28, no. 17, pp. 6938–6959, 2015, doi: 10.1175/JCLI-D-14-00754.1.

[24] B. Gumus, S. Oruc, I. Yucel, and M. T. Yilmaz, "Impacts of Climate Change on Extreme Climate Indices in Türkiye Driven by High-Resolution Downscaled CMIP6 Climate Models," *Sustain.*, vol. 15, no. 9, 2023, doi: 10.3390/su15097202.

[25] S. YUE and M. HASHINO, "Probability distribution of annual, seasonal and monthly precipitation in Japan," *Hydrol. Sci. J.*, vol. 52, no. 5, pp. 863–877, Oct. 2007, doi: 10.1623/hysj.52.5.863.

[26] C. Martinez-Villalobos and J. D. Neelin, "Why Do Precipitation Intensities Tend to Follow Gamma Distributions?," *J. Atmos. Sci.*, vol. 76, no. 11, pp. 3611–3631, 2019, doi: https://doi.org/10.1175/JAS-D-18-0343.1.

[27] B. Prasetyo, H. Irwandi, and N. Pusparini, "Karakteristik Curah Hujan Berdasarkan Ragam Topografi Di Sumatera Utara," *J. Sains Teknol. Modif. Cuaca*, vol. 19, no. 1, p. 11, 2018, doi: 10.29122/jstmc.v19i1.2787.

[28] J. L. Salinas, A. Castellarin, S. Kohnová, and T. R. Kjeldsen, "Regional parent flood frequency distributions in Europe – Part 2: Climate and scale controls," *Hydrol. Earth Syst. Sci.*, vol. 18, no. 11, pp. 4391–4401, 2014, doi: 10.5194/hess-18-4391-2014.

[29] K. Pangaluru *et al.*, "Estimating changes of temperatures and precipitation extremes in India using the Generalized Extreme Value (GEV) distribution," *Hydrol. Earth Syst. Sci. Discuss.*, vol. 2018, pp. 1–33, 2018, doi: 10.5194/hess-2018-522.

[30] Z. Rulfová, A. Buishand, M. Roth, and J. Kyselý, "A two-component generalized extreme value distribution for precipitation frequency analysis," *J. Hydrol.*, vol. 534, pp. 659–668, 2016, doi: https://doi.org/10.1016/j.jhydrol.2016.01.032.