Distribution fitting on rainfall data in Jakarta

V Kurniawan¹
¹Department of Civil Engineering, Universitas Tarumanagara, Indonesia

*vkurniawan@ft.untar.ac.id

Abstract. The estimation of design flood or rainfall for various return periods is not easy due to the difficulty in finding the fitting distribution for the corresponding data. This study focuses on selecting the best distribution in representing rainfall data from Kemayoran rain gauge (Jakarta, Indonesia) for the last 20 years. The objective is to probe the most fitting distribution for Jakarta. This study applies goodness of fit tests to Kemayoran rainfall data. The result is quite unexpected as the best fitting distributions are not commonly used in flood frequency analysis (FFA) subject and most of them are distributions with three parameters. The referred distributions in Indonesia national standard for flood estimation (SNI 2415:2016) produce mixed results. The classic distributions in FFA e.g. Lognormal 3P and Log-Pearson 3 show good compatibility to the corresponding data while the less prevalent distributions such as Lognormal and Normal are not as compatible.

1. Introduction
Flood frequency analysis (FFA) is an important practice in hydraulic engineering and water resources engineering as the determination of design flood is one of the key factors in designing hydraulic structures e.g. dam, flood control structure, urban drainage system etc [1] [2]. Underestimating design flood could inflict heavy financial damage and at worst, it could claim lives. On the other hand, overestimating design flood means excessive resources being wasted needlessly. Therefore, it is crucial to determine design flood as accurate as possible in order to achieve effective and efficient design.

Design flood determination is a perennial issue. The process starts from determining rainfall for specific return period or extreme rainfall which later will translate into design flood. However, the determination of extreme rainfall hitherto remains a difficult practice although it has been researched for decades. The studies can be traced back to the work of Das in 1952 [3] or Buishand in 1977 [4]. Despite of such long period of research on stochastic aspect on rainfall, the estimation of extreme rainfall is still an ongoing subject especially in distribution fitting aspect. The accurate distribution fitting means more accurate design flood estimation hence it compels many researchers to perfect it.

This paper will focus on finding the most representative distribution for a particular rain gauge in Jakarta, Indonesia. The nation has a national standard (will be referred as SNI 2415:2016 henceforth) to estimate the design flood [5]. The standard mentions 8 (eight) distributions which are likely to describe the behaviour of the series of maximum annual rainfall in Indonesia. The distributions are:

1. Gumbel;
2. Pearson3;
3. Log-Pearson 3;
4. Gamma;
5. Log-Gamma;
6. Normal;
7. Log-Normal (2 parameters); and
8. Log-Normal 3 parameters.

Although there are eight distributions referred in SNI 2415:2016, there are only four distributions which are customarily used in Indonesia i.e. Normal, Lognormal, Gumbel, and Log-Pearson 3. Gumbel and Log-Pearson 3 are the most frequently selected distributions to model rainfall data while Normal and Lognormal often serve as the benchmarks for other distributions in goodness of fit test.

The main objective of this study is to investigate the most fitting distributions in representing the corresponding rainfall data. It will assess if the referred distributions in SNI 2415:2016 fit the assessed rainfall data. Furthermore, it is plausible if the pattern of rainfall data in Indonesia follows other distribution apart from the distributions mentioned in SNI 2415:2016. Therefore, the study also assesses the possibility if there are feasible alternative distributions in modelling Jakarta rainfall.

2. Case Study
The source of the rainfall data is from Kemayoran meteorology station which is located in Jakarta, Indonesia. Its geographical position is at -6.16 latitude and 106.84 longitude and 4 meter above mean sea level. It is registered in World Meteorological Organization with the ID number 96745. The station is positioned in the midst of the city with the prevailing tropical climate. There are two seasons i.e. dry season and wet season. The latter roughly lasts 4 months in a year.

Kemayoran rain gauge is chosen because it is located in the headquarters of Badan Meteorologi, Klimatologi, dan Geofisika (the national body of meteorology, climatology, and geophysics; abbreviated as BMKG). Therefore, it grants the accuracy, the reliability, and the availability of the retrieved data.

The assessed rainfall data comes from the observation of Kemayoran rain gauge in the last 20 years (1999-2018). The data is retrieved online from BMKG official website. The maximum rainfall in a day for each year are displayed in table 1 (abbreviated as R).

3. Methodology
This study utilises MathWave software to perform the goodness of fit analysis. It examines the data sample against three sorts of goodness of fit test i.e. Kolmogorov-Smirnov, Anderson-Darling, and Chi-Squared. The test will determine which distribution is the most fitting for the given data sample from the available 61 distributions.

Kolmogorov-Smirnov is a test which calculates the maximum difference between an empirical and a hypothetical cumulative distribution [6]. Anderson-Darling examines if the data sample conforms to a population from particular distribution [7]. Lastly, Chi-Square divides the sample data into several intervals and compare the factual number of data to the expected number of data in each interval [8]. These test of goodness of fit methods are very commonly employed in statistics calculation including rainfall analysis. The result of implementing these tests to the Kemayoran rainfall data will be displayed in the next section.
Table 1. The annual maximum rainfall in Jakarta in the last 20 years

| Year | R (mm/day) |
|------|------------|
| 1999 | 147.2      |
| 2000 | 94.8       |
| 2001 | 82.2       |
| 2002 | 168.5      |
| 2003 | 199.7      |
| 2004 | 129.3      |
| 2005 | 124.1      |
| 2006 | 72.0       |
| 2007 | 234.7      |
| 2008 | 192.7      |
| 2009 | 122.5      |
| 2010 | 93.0       |
| 2011 | 119.2      |
| 2012 | 105.2      |
| 2013 | 193.4      |
| 2014 | 147.9      |
| 2015 | 277.5      |
| 2016 | 124.5      |
| 2017 | 179.7      |
| 2018 | 104.6      |

4. Result and Discussion
The brief descriptive statistics of the data in table 1 is shown in table 2. Meanwhile, the result of the implementation of the aforementioned goodness of fit tests on Kemayoran rainfall data is displayed in table 3.

Table 2. The annual maximum rainfall in Jakarta in the last 20 years

| Statistic        | Value |
|------------------|-------|
| Sample size      | 20.0  |
| Mean             | 147.2 |
| Standard deviation | 94.8  |
| Variance         | 82.2  |
| Skewness         | 168.5 |
| Median           | 199.7 |

The meaning of each column in table 3 is as follows:

- column (1) : the rank of the most fitting distribution, sorted from the value of column (8);
- column (2) : distribution type;
- column (3) : the result of Kolmogorov-Smirnov test;
- column (4) : the distribution’s rank in Kolmogorov-Smirnov test;
- column (5) : the result of Anderson-Darling test;
- column (6) : the distribution’s rank in Anderson-Darling test;
Table 3. The rank of the most fitting distribution for Jakarta rainfall 1999-2018

| Rank | Distribution          | KS Stat. | Rank | AD Stat. | Rank | CS Stat. | Rank | Rank Sum |
|------|-----------------------|----------|------|----------|------|----------|------|----------|
| (1)  | Gamma (3P)            | 0.083    | 1    | 0.165    | 2    | 0.189    | 1    | 4        |
| (2)  | Weibull (3P)          | 0.089    | 6    | 0.169    | 5    | 0.206    | 2    | 13       |
| (3)  | Johnson SB            | 0.092    | 8    | 0.145    | 1    | 0.513    | 7    | 16       |
| (4)  | Fatigue Life          | 0.085    | 2    | 0.167    | 3    | 0.745    | 12   | 17       |
| (5)  | Inv. Gaussian         | 0.086    | 4    | 0.169    | 4    | 0.753    | 11   | 26       |
| (6)  | Lognormal (3P)        | 0.089    | 5    | 0.174    | 10   | 0.735    | 11   | 26       |
| (7)  | Pearson 6 (4P)        | 0.086    | 3    | 0.172    | 9    | 0.791    | 15   | 27       |
| (8)  | Log-Logistic (3P)     | 0.089    | 7    | 0.198    | 18   | 0.5      | 6    | 31       |
| (9)  | Pearson 5             | 0.093    | 9    | 0.183    | 12   | 0.901    | 18   | 39       |
| (10) | Pearson 5 (3P)        | 0.093    | 10   | 0.183    | 13   | 0.905    | 19   | 42       |
|      | Gen. Extreme Value    | 0.097    | 13   | 0.171    | 8    | 1.265    | 34   | 55       |
|      |                       |          |      |          |      |          |      |          |
| (13) | Log Pearson 3         | 0.103    | 17   | 0.170    | 6    | 1.364    | 40   | 63       |
| (16) | Log Gamma             | 0.104    | 18   | 0.171    | 7    | 1.362    | 39   | 64       |
| (26) | Gumbel                | 0.113    | 22   | 0.193    | 16   | 1.441    | 42   | 80       |
| (30) | Gamma                 | 0.126    | 28   | 0.223    | 20   | 1.270    | 36   | 84       |
| (32) | Lognormal             | 0.115    | 23   | 0.200    | 19   | 1.476    | 44   | 86       |
| (35) | Normal                | 0.169    | 38   | 0.466    | 30   | 0.973    | 26   | 94       |

- column (7) : the result of Chi-Square test;
- column (8) : the distribution’s rank in Chi-Square test; and
- column (9) : the sum of the rank in each goodness of fit test i.e. (4) + (6) + (8).

The top ten most fitting distribution as well as the rank of the distributions referred in SNI 2415:2016 are revealed in table 3. The top ten most fitting distributions are dominated by the distributions which are not prevalent in FFA subject and most of them feature three parameters.

The discussion of the top ten most fitting distribution is as follow:

1 Gamma (3P)
Three parameter Gamma or Generalized Gamma distribution fits the rainfall data better than other distributions in this case. It resonates with the finding of Chen [10] which concluded that Generalized Gamma distribution fits the flood data for various return period reasonably well. It also agrees with USGS [11] which referred this distribution to model rainfall data. However, Generalized Gamma distribution is still less popular than its counterpart i.e. Gamma distribution (with 2 parameter) as the latter is widely implemented in FFA [12].

2 Weibull (3P)
Three parameter Weibull distribution closely shares the same account with Gamma (3P) distribution. So far there is no previous study which supports the reliability of Weibull (3P) distribution in representing extreme rainfall. However, Weibull distribution (2 parameter) is a commonly used in hydrology field of study.

3 Johnson SB
This distribution is commonly used in biology and nature field of study. There are only few rainfall-related studies which utilizes this distribution such as D’Adderio et al who estimated rainfall rate, reflectivity, and diameter [9]. However, so far there is no study which connects Johnson SB distribution with extreme rainfall or FFA.
Fatigue Life (3P)
It is usually employed to model the lifetime of material before it suffers failure. Its applicability in FFA or in hydrology field of study is unheard. It is doubtful that this distribution can be universally used to represent extreme rainfall.

Inv. Gaussian (3P)
Inverse Gaussian or Wald distribution is widely used in analysing lifetime data [13] or in bioscience [14]. Judging from its characteristics and its past usage, it is unlikely that this distribution is able to extensively fit rainfall data.

Lognormal (3P)
Three parameter Lognormal distribution shares almost similar characteristics with Lognormal distribution (2P). The difference is Lognormal (3P) introduces an additional parameter on which the value of data is reduced. Lognormal (3P) distribution is less popular than Lognormal distribution. But, Sangal and Biswas [15] proved that the former produced reasonable success when projected on the data of flood flow (note that the data derives from the flood flow observation, not from the rainfall data). In this case, Lognormal (3P) displays better fitting than other distributions referred in SNI 2415:2016.

Pearson distribution family
Four parameter Pearson 6, Pearson 5, and three parameter Pearson 5 appear in the top ten list while Log-Pearson 3, which is very prevalent in FFA, surprisingly fares marginally worse. Although Pearson 6 4P, Pearson 5, and Pearson 5 (3P) outperform Log-Pearson 3 in this case, there is no proof which supports that the former would be accurate for other FFA-related cases.

Log-Logistic
Log-Logistic distribution is employed for various purposes such as medic [16], economy [17], and FFA as conducted by Ahmad [18]. In that particular case, Ahmad proved that Log- Logistic distribution performed well and even better than the more traditionally prevalent distribution such as Generalized Extreme Value (GEV) distribution. Here, in this study, Log-Logistic distribution again performs better than GEV although more extensive studies are required to prove if the former can be universally used in FFA.

Gen. Extreme Value
Although Generalized Extreme Value (GEV) sits outside of top ten list, but the distribution worths mentioning as it is one of the most generally used distribution in modelling any extreme events including natural extreme events [19] and FFA [11] [20] [21] [22] [23]. There are plenty of studies which prove its accuracy in modelling the extreme flood or rainfall [24] [25]. Its extensive use in the world raises an expectation for GEV to perform better in modelling rainfall data in Jakarta, however the result here still shows satisfactory reliability.

The following points are the elaboration about the distributions referred in SNI 2415:2016 which yield mixed result in the goodness of fit tests:

1 Log-Pearson 3
Log-Pearson 3 has a strong history in representing the distribution of flood discharge in
USA [11] [26] and it also shows good fit at other parts of the world such as Nigeria [2] or Australia [27]. From the conducted test above, the distribution is also compatible with Jakarta rainfall data as well.

2 Log-Gamma
While Gamma or Generalized Gamma distribution are used quite frequently in FFA, there is no study which supports the implementation of Log-Gamma in this field. Here in Jakarta, the distribution fits the rainfall data well, however, the question is raised if it can be universally applied to other regions.

3 Gumbel
Gumbel or Extreme Value (EV) Type 1 or Fisher-Tippett distribution is also adopted in FFA [28] [29] although it is less popular than Generalized Extreme Value (GEV) distribution. Nevertheless the former is referred in SNI 2415:2016 while the latter is not. EV Type 1 is inferior compared to GEV in fitting Jakarta rainfall data which is quite expected as GEV is more prevalent in FFA.

4 Lognormal
Lognormal distribution is quite often employed in goodness of fit test in FFA although it often loses to other distribution (e.g. GEV, Log-Pearson 3) in fitting the rainfall or flood data. Although there are two cases from Malaysia which resulted in Lognormal distribution’s prevailing over GEV or Log-Pearson 3 [30] [31] in fitting the rainfall data, GEV and Log-Pearson 3 remain the more favourable distributions in FFA. It agrees with the finding of this study as Lognormal distribution is only positioned at rank 32nd while GEV and Log-Pearson 3 are at 13rd and 15th.

5 Gamma
Gamma distribution is one of the most prevalent distribution in FFA [11] [21], but the outcome of the goodness of fit test here does reflect its prevalence. It is placed at rank 30th and even losing to Log-Gamma which are much less popular. However, similar to the discussions of the other distributions, perhaps the distribution might be incompatible for Jakarta rainfall while it works well in other data sets.

6 Normal
Normal distribution is probably the most popular stochastic distribution, but its applicability in FFA is not as popular. Normal distribution is often used in FFA, but it is merely as a comparison benchmark with other distributions. It rarely works as compatible as GEV or Log-Pearson 3 in modelling rainfall or runoff data except for rare occasion [32] [33]. The goodness of fit tests on Jakarta rainfall data again reflects the incompatibility of Normal distribution in FFA as it is only positioned at rank 35th, placed below all the distributions referred in SNI 2415:2016.

5. Conclusions
The list of the most fitting distributions for Jakarta in table 3 produces quite an unexpected result as it is dominated by the distributions which are uncommon in FFA. The appearance of Johnson SB, Fatigue Life, Inverse Gaussian etc in fitting rainfall data is very rare if not unheard of. Further studies on different data sets on different locations are required to figure out whether those distributions are generally applicable for FFA.

The top ten list is dominated by distribution with 3 (three) parameters. It is uncertain if distributions with three (or more) parameters could fit rainfall or flood data accurately or it only applies to Jakarta’s rainfall only. However, distribution with two parameters are more practical to work with while the additional parameters tend to be relatively difficult to determine. Practicality is also as important as accuracy especially when it serves as a code or standard which should hinder the usage of tedious mathematical processes. As finding the balance between accuracy and practicality is
important in code, it would be better to employ distributions with two parameters instead of distributions with more than two parameters although the latter might be more accurate.

The distributions referred in SNI 2415:2016 generate mixed outcome. Log-Pearson 3 is one of the most common distribution for FFA in Indonesia and it produces satisfactory result. Lognormal (3P) and Log-Gamma also show positive outcome although they are rarely used. Gumbel distribution is employed as frequently as Log-Pearson 3 in modelling Indonesia rainfall data, but the former is less compatible than the latter in fitting Jakarta rainfall data. Gamma distribution is quite popular in FFA subject, but it is not as popular in Indonesia. Here it does not show excellent result either. Lognormal and Normal distributions does not produce exceptional outcome and it sort of justifies its practice in FFA as they often serve as benchmarks for other popular distributions (e.g. GEV, Log-Pearson 3), but not as main distributions which represent the corresponding data.

References

[1] Tao D Q, Nguyen V T V, Bourque A 2002 Proc. Ann. Conf. of the Canadian Soc. for Civil Eng.
[2] Olofinmoye O O, Sule B F, Salami A W 2009 New York Science J. 2(3) 1-12
[3] Das S C 1952 Australian J. Physics 8 298
[4] Buishand T A 1977 Stochastic Modelling of Daily Rainfall Sequences (Wageningen : Com. Agric. Univ.Wageningen)
[5] Badan Standardisasi Nasional 2016 SNI 2415:2016 Tata cara perhitungan debit banjir rencana (Jakarta)
[6] Massey F J M 1951 J. Amer. Stat. Assn 46 68-78
[7] Anderson T W and Darling D A 1954 J. Amer. Stat. Assn 49(268) 765-769
[8] Cochran W G 1952 Ann. Math. Stat. 23(3) 315-345
[9] D’Adderio L P, Cugenerone K, Porcu’ F, Michele C D, Tokay A 2016 Adv. Water Resources 23(3) 315-345
[10] Chen L, Singh V P, and Xiong F 2017 Entropy 19(6) 239
[11] USGS 1982 Guidelines For Determining Flood Flow Frequency (Virginia)
[12] Yue S 2001 Hydrological Processes 15(6) 1033-1045
[13] Basak P and Balakrishnan 2010 J. Statistical. Computation and Simulation 82(7) 1055-1072
[14] Wise M E, Osborn S B, Anderson J, Tomlinson R W S 1968 Math. Biosci 2 199-224
[15] Sangal B P and Biswas A K 1970 Water Resources Research 6(2) 505-515
[16] Bennett S 1983 Royal Stats. Soc. 32(2) 165-171
[17] Kantam R R L, Rao G S, and Sriram B 2007 J. Appl. Stats. 33(3) 291-296
[18] Ahmad M I, Sinclair C D, and Werritty 1988 J. Hydrology. 98(3-4) 205-224
[19] Smith R L 1986 J. Hydrology. 86 27-43
[20] Smith J A 1987 Water Resour. Res. 23(8) 1657-1666
[21] Onoz B and Bayazit M 1995 J. Hydrology. 167 195-208
[22] Morisson J E and Smith J A 2002 Water Resour. Res. 38(12)
[23] National Engineering Research Council 1975 Flood Studies Report Vols. 1-5 (London)
[24] Zalina M D, Desa M N M, Nguyen V T V, and Kassim A H M 2002 Water Sci. Tech. 45(2) 63-8
[25] Alam M A, Emura K, Farnham C, and Yuan J H 2018 Climate 6(9)
[26] Chow V T, Maidment D R, and Mays L W 1988 Applied Hydrology (New York : McGraw Hill)
[27] McMahon T A and Srikanthan R 1981 J. Hydrology 52 139-147
[28] Todorovic P 1978 Water Resour. Res. 14(2) 345-356
[29] Onen F and Bagatur T 2017 Arab J Sci Eng 42 3895-3906
[30] Suhaila J and Jemain A A 2007 J. Phys.: Conf. Series 890
[31] Deraman W H A W, Mutalib N J A, and Mukhtar N Z 2017 J. Phys.: Conf. Series 890
[32] Markovic R D 1965 Probability Functions of Best Fit to Distributions of Annual Precipitation
and Runoff (Colorado : Hydr. Papers Colorado State Univ.) [33] Yue S 1999 Water Int. 24(3) 248-254