Selecting the probability distribution of annual maximum temperature in Malaysia

Nurfatini Mohd Supian[1*], and *Husna* Hasan[1]

[1]School of Mathematical Sciences, Universiti Sains Malaysia, 11800 Pulau Pinang, Malaysia

Abstract. The issues on global warming have become very popular and been discussed both locally and internationally. This phenomenon due to the temperature rises will increase the variability of climate and more natural disasters were expected to occur. Increasing of global temperature will affect the agricultural sector, increase some of the infectious diseases that may lead to high mortality rates in humans, high demand for electricity, water and food which eventually affecting the economy of Malaysia. Hence, this work aims to study the best fitted probability distribution that describes the annual maximum temperature recorded at seventeen meteorological stations in Malaysia. The Normal, Lognormal, Gamma, Weibull and Generalized Skew Logistic distributions are considered using the maximum likelihood estimation method to estimate the parameters. The goodness of fit test and model selection criteria such as Kolmogorov-Smirnov and Anderson-Darling tests, Corrected Akaike Information Criterion and Bayesian Information Criterion are used to measure the accuracy of the predicted data using theoretical probability distributions. The results show that most of the stations favour the Generalized Skew Logistic distribution as the best fitted probability distribution. Also, some stations favour the Normal, Lognormal as well as Weibull distribution as the best fitted distribution to describe the annual maximum temperature.

1 Introduction

Recently, the issues of global warming are becoming more common due to the high temperature. Although the changes in mean temperature are small, it is actually due to the large changes in the frequency. According to Intergovernmental Panel Climate Change (IPCC), based on observations of increases in global average air temperatures, the warming of the climate system is now “unequivocal” and the surface temperatures is predicted to rise over the 21st and the heat waves are expected to occur more frequently and last longer [1, 2]. This phenomenon due to the temperature rises eventually will increase the variability of climate and also the occurrence of natural disasters.

Malaysia consists of two regions, namely, Peninsular Malaysia and East Malaysia and is located within the equatorial region where it is characterized as being hot and humid throughout the year. The average annual temperature is about 27°C where the daytime temperature rises above 30°C and the night temperature rarely drops below 20°C. The

*Corresponding author: nurfatini8532@gmail.com temperature in Malaysia is predicted to continue on an increasing trend [3] and compared to the other region in Malaysia, the rising temperature in Western Peninsular Malaysia is more significant [4].

The anthropogenic or human factors such as land conversion, industrialization and transportation release greenhouse gasses which amplify the air temperature [4]. Several studies have shown that extreme temperatures will increase heat waves [5]. Stress from excess heat may lead to blood pressure and heart diseases. Increased temperatures and changes in precipitation patterns may cause a rise in malaria, cholera and dengue. This problem is already observed with malaria in Southeast Asia. Malaysia is one of the countries located in Southeast Asia faces potential threats to population health and development due to the changes in temperature.

Based on the research conducted by various researchers, the increase in global temperature gives negative effects on the agricultural sector, health, food and water supply as well as the environment. The rate of evaporation becomes faster and thus lead to drought. The agricultural sector will be affected by the impacts of rising temperatures where the soil moisture tends to reduce more. The low humidity will increase the risk of wildfires and open burning which eventually speeding up the warming in air temperature.

Mori [6] found that there is an increased demand for electricity during periods of extremely hot temperatures. Moreover, extremely high temperatures even have an indirect impact on the economy of Malaysia. As we all know, floods are one of the inevitable accidents that frequently happen in Malaysia. From 2001 to 2005, a total of RM1.79 billion was spent on structural flood mitigation measures [7]. The increment of extreme weather such as drought and heavy floods is also associated with the influence of the El Nino phenomenon [4].

Several analysts from all over the world have conducted this statistical analysis of maximum temperature in different locations. Araújo et al. [8] have shown that the Normal distribution as the best distribution to describe the daily series of maximum temperature in Iguatu City, Ceara. In 2011, de Araújo et al. [9] further their research in Iguatu City, Ceara where they focused on fitting the probabilities of occurrence of maximum temperature in the scale of fifteen days for each month of the year by using the Lognormal distribution. Torsen et al. [10] revealed that the Johnson Sb distribution is the best fitted distribution for the maximum temperature in Adamawa state in Nigeria. The Generalized Skew Logistic distribution is selected as the best fit to observe the monthly maximum temperature of Dhaka station [11]. In 2018, Abdulla and Hossain [12] found the Generalized Skew Logistic distribution was the favoured distribution for Cox’s Bazar station while the Weibull distribution describes the best for Patuakhali station.

In Malaysia, several similar analyses have been done for the rainfall [13-15] and wind speed data [16, 17]. Daud et al. [13] discussed the comparative assessment of eight candidate distributions in providing accurate and reliable maximum rainfall estimates for Malaysia. Among the eight distributions, the Generalized Extreme Value (GEV) distribution is chosen as the best fitted probability distribution to describe the annual rainfall in Malaysia. Dan’azumi et al. [14] also studied the statistical distribution of hourly rainfall depth for twelve representative stations spread across Peninsular Malaysia. It is observed that the Generalized Pareto distribution (GPD) fits well compared to the Exponential and Gamma distribution. Meanwhile, the Weibull distribution is widely used and chosen as the best fits for describing the wind speed data [16, 17].

Aside from analyzing the rainfall and wind speed distribution, it is also important to identify the behaviour of maximum temperature as it is necessary to decrease the impact of climate change in this country. Most studies on temperature emphasize the use of the Generalized Extreme Value [3, 18, 19] and Generalized Pareto distribution [20] only. Hence, this work aims to find the maximum temperature distribution by comparing several distributions and determine the probability distribution that describes the best for the annual maximum temperature in Malaysia.

1.1 Data description

The daily maximum temperature data are recorded at seventeen meteorological stations in Malaysia over the period of January 1994 to December 2017, in which the annual maximum temperature is obtained. Fourteen of the stations, namely, Chuping (CP), Alor Setar (AS), Bayan Lepas (BL), Sitiawan (ST), Subang (SBG), Kuala Lumpur International Airport (KLIA), Seremban (SR), Malacca (MC), Senai (SN), Mersing (MSG), Muadzam Shah (MS), Kuantan (KN), Kuala Terengganu (KT) and Kota Bharu (KB) are located in Peninsular Malaysia. The other three stations that are Kuching (KCG), Labuan (LB) and Kota Kinabalu (KK) are located in East Malaysia. The datasets obtained from the Malaysian Meteorological Department are measured in degree Celsius (°C). All the data are recorded from 1994 to 2017 except for KLIA station which is observed from 1999 to 2017.

2 Methodology

In order to describe the annual maximum temperature, it is important to identify the distribution that fits well with the data. As mentioned, most studies in Malaysia are focusing on the analysis of maximum temperature data by using the GEV and GPD distributions. Hence, this study aims to cover other distributions other than GEV and GPD to analyze the maximum temperature. The Normal, Lognormal, Gamma, Weibull and Generalized Skew Logistic distributions are used and the parameters are estimated using the maximum likelihood estimation method since it provides a consistent approach to the parameter estimation problems. Table 1 shows the considered probability distributions.

Table 1. The probability density function of five distributions and its parameter.

Distribution	**Probability density function**	**Parameter**
Normal (N)	$f(x)=\dfrac{1}{\sigma\sqrt{2\pi}}\exp\left[-\dfrac{1}{2}\left(\dfrac{x-\mu}{\sigma}\right)^2\right];-\infty<x<\infty$	$\mu=$ location $\sigma=$ scale $\gamma=$ shape
Lognormal (LG)	$f(x)=\dfrac{1}{x\sqrt{2\pi\sigma^2}}\exp\left[\dfrac{-\left(\ln(x)-\mu\right)^2}{2\sigma^2}\right];0<x<\infty,\mu\in R,\sigma>0$	
Gamma (G)	$f(x)=\dfrac{x^{\gamma-1}}{\sigma^{\gamma}\Gamma(\gamma)}\exp\left(-\dfrac{x}{\sigma}\right);x>0,\gamma>0,\sigma>0$	
Weibull (W)	$f(x)=\dfrac{\gamma}{\sigma}\left(\dfrac{x}{\sigma}\right)^{\gamma-1}\exp\left[-\left(\dfrac{x}{\sigma}\right)^{\gamma}\right];0<x<\infty,\gamma>0,\sigma>0$	
Generalized Skew Logistic (GSL)	$f(x)=\dfrac{\gamma}{\sigma}\dfrac{\exp\left(-\left[\dfrac{x-\mu}{\sigma}\right]\right)}{\left(1+\exp\left(-\left[\dfrac{x-\mu}{\sigma}\right]\right)\right)^{\gamma+1}};-\infty<x<\infty$	

2.1 Goodness of fit tests and model selection criterion

2.1.1 Kolmogorov-Smirnov (KS) and Anderson-Darling (AD) tests

The Kolmogorov-Smirnov statistic computes the largest difference between the empirical distribution function of the sample and the cumulative distribution function of the selected distribution. The test statistic is defined as:

$$D = \max_{1 \le i \le n}\left(F(x_i) - \frac{i-1}{n}, \frac{1}{n} - F(x_i) \right)$$

where F is the theoretical cumulative distribution of the tested distribution that must be continuous, and the parameter is fully specified. The null hypothesis will be rejected if D is greater than the critical value computed from the statistical table.

The Anderson-Darling test is a modification of the Kolmogorov-Smirnov test and gives more weight to the tails than does the KS test. The test statistic for the AD test is

$$A^2 = n - \sum_{i=1}^{n} \frac{(2i-1)}{n}\left[\ln F(x_i) + \ln\left(1 - F(x_{n+1-i})\right)\right]$$

and F is the cumulative distribution function of the specified function while x_i's are the ordered data and n is the sample size. From the statistical table, the null hypothesis will be rejected if A^2 is greater than the critical value. The null hypothesis of both tests is that the data follow the specified distribution.

2.1.2 Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC)

The AIC and BIC are used to check the accuracy of the predicted data using theoretical probability distributions. The AIC is defined as

$$\text{AIC} = 2k - 2\ln(L)$$

where n is the number of observations contributes to the model and $\ln(L)$ is the log-likelihood function for the statistical model with k parameters. For a small sample size, a corrected AIC is developed and defined as

$$\text{AICC} = \text{AIC} + \frac{2k(k+1)}{n-k-1}.$$

The BIC is computed as

$$\text{BIC} = -2\ln(L) + k\ln(n)$$

The distribution which provides the smallest value of AIC and BIC is preferable.

2.2 Model Validation

2.2.1 Quantile-Quantile (Q-Q) Plot

The Q-Q plot is obtained for the stations which show an unclear decision on selecting the best fitted probability distribution. It is a graphical tool that help us to assess if a set of data plausibly came from some theoretical distribution. A sample of $x_1, x_2, .., x_n$ are used to construct the plot by plotting the theoretical quantiles against the sample quantiles, x_i, where x refers to the annual maximum temperature data. If the empirical distribution is consistent with the theoretical distribution, the points in the Q-Q plot should lie along the 45-degree reference line.

3 Results and discussion

The daily maximum temperatures that are covered up for 19 to 24 years are observed. The annual maxima are used as the selection period to study the characteristic of the maximum temperature in seventeen meteorological stations mentioned.

Table 2 shows a descriptive analysis of the annual maximum temperature. The longitude and latitude are also listed. Chuping station records the highest mean annual maximum temperature with 37.28°C followed by Alor Setar station with 36.95°C. These two stations are located in Western Peninsular Malaysia. Meanwhile, the mean of Kuala Terengganu station, which is located in the east is observed as the lowest annual maximum temperature. This is consistent with the fact that the temperature in Western Peninsular Malaysia experiences a more significant rise compared to other regions in Malaysia [4].

The bigger standard deviations of Chuping and Alor Setar stations indicate that the annual maximum temperature deviates far from the average maximum temperature while the smallest standard deviation by Kuala Terengganu station indicates that the maximum temperatures are mostly close to the mean.

Table 2. Summary of the maximum temperature for seventeen meteorological stations in Malaysia.

Station	Latitude	Longitude	Minimum	Maximum	Mean	Standard Deviation
AS	6.1263° N	100.3672° E	35.40	39.10	36.9583	1.05456
BL	5.2945° N	100.2593° E	33.30	36.00	34.5625	0.73710
CP	6.4985° N	100.2580° E	35.90	40.10	37.2875	1.02674
ST	4.2168° N	100.6978° E	34.20	36.40	35.1625	0.59331
KLIA	2.7424° N	101.7062° E	34.50	37.40	35.8368	0.84275
SBG	3.0567° N	101.5851° E	35.00	37.90	36.1792	0.67694
SR	2.7259° N	101.9378° E	34.40	38.30	35.8417	1.04420
MC	2.1896° N	102.2501° E	34.20	38.00	35.5583	0.96365
KN	3.7634° N	103.2202° E	34.60	37.80	35.7083	0.76948
MS	3.0562° N	103.0852° E	35.30	37.70	36.1667	0.62043
SN	1.6020° N	103.6444° E	34.10	37.20	35.3167	0.67352
MSG	2.4309° N	103.8361° E	33.50	38.20	35.0833	0.99637
KB	6.1248° N	102.2544° E	34.20	36.40	34.9833	0.65784
KT	5.3296° N	103.1370° E	33.50	35.80	34.4542	0.58680
KK	5.9804° N	116.0735° E	34.00	36.50	35.2125	0.76970
LB	5.2831° N	115.2308° E	33.60	36.60	34.6208	0.62761
KCG	1.5535° N	110.3593° E	34.60	37.30	35.5958	0.68110

The test statistics for the KS and AD tests along with the AICC and the BIC for the annual maximum temperature were calculated for each of the considered distributions. The distribution favoured by each station are counted based on the smallest value produced from the goodness of fit tests and the model selection criterion. The most preferable distribution for each of the goodness of fit tests and the model selection criterion are shown in Table 3.

Table 3. The most preferable probability density function by each model selection tools.

Station	KS test		AD test		AICC		BIC	
	Dist	Statistic	Dist	Statistic	Dist	Statistic	Dist	Statistic
CH	GSL	0.07168	GSL	0.1593	GSL	71.6914	GSL	74.02557
AS	GSL	0.09134	GSL	0.1949	GSL	73.3393	LG	75.55297
BL	GSL	0.11223	LG	0.2779	LG	56.9659	LG	58.75009
ST	GSL	0.10694	GSL	0.2150	LG	46.4717	LG	48.25594
SBG	N	0.13754	LG	0.3604	LG	52.7760	LG	54.56025
KLIA	GSL	0.09745	GSL	0.2337	LG	50.9545	LG	52.09345
SR	GSL	0.13004	GSL	0.3143	GSL	72.0374	GSL	74.3716
MC	GSL	0.11006	GSL	0.2079	GSL	66.9848	GSL	69.31898
SN	GSL	0.12441	GSL	0.3076	LG	52.3618	LG	54.14604
MSG	GSL	0.11965	GSL	0.3460	GSL	67.9769	GSL	70.3111
KN	GSL	0.11362	GSL	0.2775	GSL	56.1932	GSL	58.52737
MS	GSL	0.12288	GSL	0.2805	GSL	46.3503	GSL	48.68446
KT	LG	0.10473	LG	0.3153	LG	45.8718	LG	47.65601
KB	GSL	0.16046	GSL	0.6270	LG	51.3353	LG	53.11952
KK	W	0.11031	W	0.4443	N	59.0956	N	60.87982
LB	GSL	0.09300	GSL	0.2683	GSL	48.1181	GSL	50.45234
KCG	GSL	0.11807	GSL	0.3823	GSL	52.7368	LG	54.69403

Based on the KS test, the Generalized Skew Logistic distribution provides a good fit to the annual maximum temperature data at most of the stations except for Subang, Kuala Terengganu and Kota Kinabalu stations which favour the Normal, Lognormal and Weibull distributions, respectively. The AD test reveals that the results are almost the same as the KS test. Only for Bayan Lepas and Subang stations where the Lognormal distribution fit well to the annual maximum temperature data.

Meanwhile, the AICC and BIC give quite a similar result except for Alor Setar and Kuching stations, where there is a different result between the Generalized Skew Logistic and Lognormal distributions. In general, eight stations provide a clear decision while nine other stations show an unclear decision on selecting the best fitted probability distribution.

The best fitted probability distribution for the annual maximum temperature for seventeen meteorological stations are presented in Table 4, along with the estimated parameters. Based on the comparison from the goodness of fit tests, the model selection criterion as well as the validation from the Q-Q plot, it is observed that the Generalized Skew Logistic distribution provides the best fitted probability distribution for twelve stations such as Chuping, Alor Setar, KLIA, Seremban, Melaka, Kuantan, Muadzam Shah, Senai, Mersing, Kota Bharu, Labuan and Kuching stations. This result is also consistent with Hossain et al. [11] who conducted a study using maximum temperature for Dhaka stations.

Meanwhile, three stations, namely, Bayan Lepas, Sitiawan and Kuala Terengganu stations favour the Lognormal distribution. The Weibull distribution provides the best fit for Kota Kinabalu station. This result is also consistent with Abdulla and Hossain [12], who conducted a study using maximum temperature for Patuakhali stations. Lastly, Subang station favours both of the Normal and Lognormal distributions as the best distributions. This is also consistent with the results obtained by Araújo et al. [8] and de Araújo et al. [9].

Table 4. The best fitted probability distribution for seventeen meteorological stations in Malaysia.

Station	Best fit	Parameters
CH	GSL	μ = 31.0743, γ = 1550.011, σ = 0.7831
AS	GSL	μ = 30.2124, γ = 2338.526, σ = 0.8081
BL	LG	μ = 3.54255, σ = 0.02086
ST	LG	μ = 3.5598, σ = 0.016476
KLIA	GSL	μ = 25.9569, γ = 103601, σ = 0.68517
SBG	N	μ = 36.1792, σ = 0.66268
	LG	μ = 3.58832, σ = 0.01826
SR	GSL	μ = 29.0555, γ = 3401.995, σ = 0.7769
MC	GSL	μ = 29.6669, γ = 2441.064, σ = 0.7014
KN	GSL	μ = 31.053029, γ = 2140.39, σ = 0.5634
MS	GSL	μ = 31.53876, γ = 14252.92, σ = 0.455
SN	GSL	μ = 34.913235, γ = 1.9256, σ = 0.420902
MSG	GSL	μ = 25.9812, γ = 154432.5, σ = 0.726
KB	GSL	μ = 30.325165, γ = 5148.54, σ = 0.5095
KT	LG	μ = 3.5395, σ = 0.0166
KK	W	μ =52.3396, σ = 35.577
LB	GSL	μ = 34.1092874, γ = 2.49478, σ = 0.3955
KCG	GSL	μ = 31.180608, γ = 2301.26, σ = 0.5305

4 Conclusion

In this study, the annual maximum temperature recorded at seventeen meteorological stations in Malaysia are analyzed using the Normal, Lognormal, Gamma, Weibull and Generalized Skew Logistic distributions. The parameters are estimated using the maximum likelihood estimation method. The selection of best fitted probability distribution is determined using the comparison between the goodness of fit test and the model selection criterion namely the Kolmogorov-Smirnov and Anderson-Darling tests along with the Corrected Akaike Information Criterion and Bayesian Information Criterion. It has been observed that most of the stations favour the Generalized Skew Logistic distribution and some of the stations favour the Lognormal, Normal as well as Weibull distributions as the best fitted probability distribution to describe the annual maximum temperature. It is also noticed that most stations having a right-skewed distribution as shown in Fig. 1, where the tail of the distribution is longer to the right-hand side compared to the left-hand side. It is observed that the number of maximum temperature peaks around 34°C to 36°C and the distribution extend further into the higher maximum temperature than to the lower maximum temperature. Hence, the stations that follow the Generalized Skew Logistic distribution have a higher annual maximum temperature compared to the stations that follow the Normal, Lognormal and Weibull distributions. The results can be improved for future works by using the longer-term period as it would give more accurate information about the behaviour of annual maximum temperature. More three-parameter distributions can be included since the distribution might provide more fit to the data. Plus, the more parameters a probabilistic model has, the more flexible it becomes in adjusting the data. The analysis on the maximum temperature allow the scientists to study the behaviour of maximum temperature and its impacts and later make

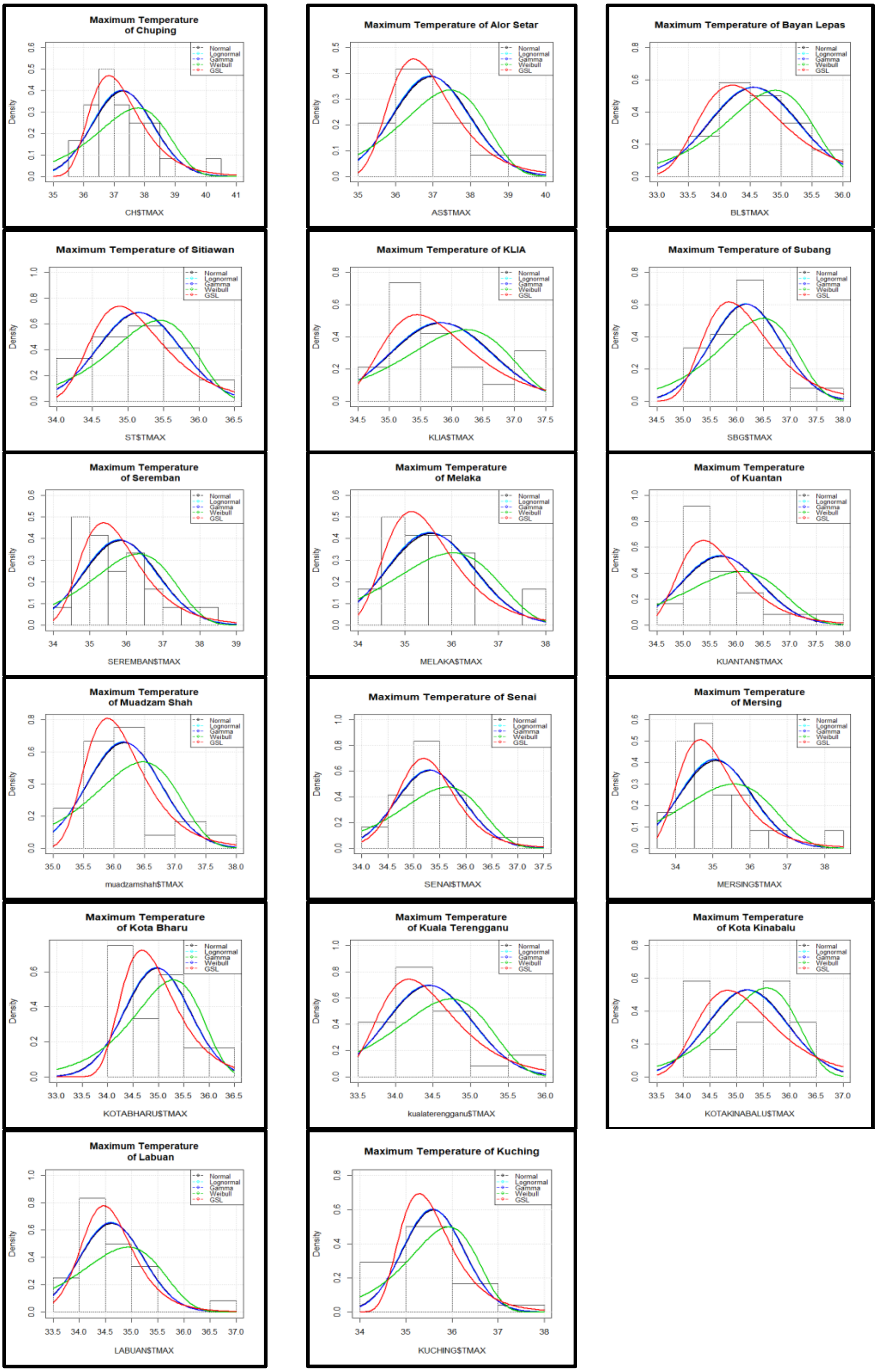

Fig. 1. The graphical comparison between the selected distributions for each station.

a prediction. The results from this study will give benefits to the society to build a better explanation on the maximum temperature and help to bring awareness for the local people about the maximum temperature. We hope that this study on the maximum temperature will be useful in understanding the events of extreme temperatures in Malaysia.

We would like to thank the Malaysian Meteorological Department for providing the data and the School of Mathematical Sciences, Universiti Sains Malaysia for the support.

References

1. B. Raggad, Env. Model. Assess. **23**(1)**,** 99 – 116 (2018)
2. IPCC. Climate Change 2014 : Synthesis Report. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, Pachauri, R.K., Allen, M.R., Barros, V.R., Broome, J., Cramer, W., Christ, R., Church, J.A., Clarke, L., Dahe, Q., Dasgupta, P. & Dubash, N.K.]. IPCC, Geneva, Switzerland, 151 (2014)
3. N.H. Mohd Salleh, H. Hasan, ESTEEM Academic J. **13**, 107 – 117 (2017)
4. Malaysian Meteorological Department, Climate change scenarios for Malaysia 2001 – 2099. Malaysia Meteorological Department, Kuala Lumpur (2009)
5. N.H. Mohd Salleh, H. Hasan, F. Yunus, J. Math. Stat. **8**(2A), 28 – 35 (2020)
6. K.W. Mori, Modelling extreme temperature behaviour in Upper East Region, Ghana (Doctoral dissertation, 2016)
7. H.A. Rahman, IJoM-NS **1**(2), 55 – 77 (2018)
8. E.M. Araújo, I.N Silva, J.B. de Oliveira, E.G. Junior, B.M. de Almeida, Rev. Ciênc. Agron. **41**(1), 36 – 45 (2010)
9. E.M. de Araújo, E.M. Araújo, J.B. de Oliveira, E.R.F. Lêdo, P. C. Viana, M.G. Silva, Revista Brasileira de Agricultura Irrigada **5**(1)**,** 48 – 53 (2011)
10. E. Torsen, A.A. Akinrefon, B.Z. Rueben, Y.V. Mbaga, IOSR-JM **11**(4), 1 – 6 (2015)
11. M.M. Hossain, F. Abdulla, M.H. Rahman, Jahangirnagar Univ. J. Stat. Stud. **33**, 33 – 45 (2016)
12. F. Abdulla, M.M. Hossain, J. Env. Stat. **8,** 020010 (2018)
13. Z.M. Daud, M.N.M. Desa, V.T.A. Nguyen, A.H.M. Kassim, Water Sci. Technol. **45**(2), 63 – 68 (2002)
14. S. Dan'azumi, S. Shamsudin, A.A. Rahman, Int. J. Env. Ecol. Eng. **4**(12)**,** 670 – 674 (2010)
15. J.L. Ng, S. Abd Aziz, Y.F. Huang, A. Wayayok, M.K. Rowshon, *Analysis of annual maximum rainfall in Kelantan, Malaysia*, in Proceedings of the 3rd International Conference on Agricultural and Food Engineering, 23-25 August 2016, Kuala Lumpur, Malaysia (2016)
16. M.S. Kamil & A.M. Razali, *Assessing distributions for monthly mean wind speed data*, in AIP Conference Proceedings, 13-14 April 2016, Selangor, Malaysia (2016)
17. N. Sanusi, A. Zaharim, & S. Mat, *Separate analysis of wind speed and direction for Mersing, Malaysia,* in Proceedings of Mechanical Engineering Research Day, 90 – 91 (2016)
18. H. Hasan, N.A. Radi, S. Kassim, *Modeling of extreme temperature using generalized extreme value (GEV) distribution: a case study of Penang,* in Proceedings of the World Congress on Engineering 2012, 4-6 July 2012, London (2012)

19. H. Hasan, N. Salam, M.B. Adam, Int. J. Math. Comput. Phys. Elec. Comput. Eng. **7**(6), 983 – 989 (2013)
20. N.H. Mohd Salleh, H. Hasan, Sci. Int. (Lahore) **30**(1), 63 – 67 (2018)