Estimating the Best-Fitted Probability Distribution for Monthly Maximum Temperature at the Sylhet Station in Bangladesh

Md. Rashidul Hasan

Department of Public Health, Leading University, Sylhet-3112, Bangladesh

✉ **Corresponding Author:** Md. Rashidul Hasan, **E-mail**: rashidulhasan067@gmail.com

ARTICLE INFORMATION	ABSTRACT

The estimation of a suitable probability model depends mainly on the features of available temperature data at a particular place. As a result, existing probability distributions must be evaluated to establish an appropriate probability model that can deliver precise temperature estimation. The study intended to estimate the best-fitted probability model for the monthly maximum temperature at the Sylhet station in Bangladesh from January 2002 to December 2012 using several statistical analyses. Ten continuous probability distributions such as Exponential, Gamma, Log-Gamma, Beta, Normal, Log-Normal, Erlang, Power Function, Rayleigh, and Weibull distributions were fitted for these tasks using the maximum likelihood technique. To determine the model's fit to the temperature data, several goodness-of-fit tests were applied, including the Kolmogorov-Smirnov test, Anderson-Darling test, and Chi-square test. The Beta distribution is found to be the best-fitted probability distribution based on the largest overall score derived from three specified goodness-of-fit tests for the monthly maximum temperature data at the Sylhet station.

1. Introduction

Increasing temperature is the burning issue at the present time. High temperatures are one of the most studied severe occurrences with agriculture, water resources, energy consumption, and human mortality among the fields in which they have an impact (Watson et al., 1996). Several studies on the trend of climatic parameters such as temperature in Bangladesh have been conducted in the past. The temperature has been altered, according to Chowdhury and Debsharma (1992) and Mia (2003), by analyzing historical data from a few specified meteorological sites. According to Parthasarathy et al. (1987) and Mehrotra et al. (1995), the average yearly temperature of Bangladesh has climbed by 0.31^0C between 1895 and 1980. Karmakar and Shrestha (2000) predicted that the annual maximum temperature in Bangladesh would rise 0.40^0C and 0.73^0C by the year 2050 and 2100 respectively, based on data from 1961 to 1990. According to Ahmad, Warrick, Ericksen, and Mirza (1996), the vast geographical area containing Bangladesh has become warmer by around 0.5^0C over the last 100 years.

Several researchers from around the world and in the country have looked into using probability distributions to analyze temperature data. But it is necessary to find the appropriate probability model for the temperature data on a small scale. According to Hossain, Abadulla, and Rahman (2016), Generalized Skew Logistic (GSL) distribution offered the best match for the recorded monthly maximum temperature data in Dhaka station. Hughes et al. (2007) analyzed the minimum and maximum temperatures of the Antarctic Peninsula using the Generalized Extreme Value (GEV) distribution and time series models. Hasan et al. (2013) used the GEV distribution to forecast the yearly maximum temperature reported at twenty-two climatic zones in Malaysia. For the Kanyakumari region, Vivekananda (2015) used EV1 and EV2 distributions to represent extreme temperature, rainfall, and wind speed. They revealed that the EV1 probability distribution is particularly applied for modeling the extreme yearly rainfall and yearly maximum or minimum temperature, whereas the EV2 probability distribution is better equipped for modeling the annual hourly maximum wind speed for Kanyakumari.

Sylhet is one of the districts of natural beauty in Bangladesh. It is an attractive tourist place. Tourists nowadays strive to make guesses about a destination's weather before going there. So to identify the weather condition of Sylhet, it is essential to fit the best probability model for the temperature at the Sylhet station. As a result, the study aims to find the best-fitted probability distribution for the monthly maximum temperature at the Sylhet station. The resulting information is crucial for a variety of research projects, climatology, and agricultural studies in Sylhet as well as in Bangladesh.

2. Study Area

Bangladesh is a low-lying, riverine country in South Asia with a coastline of 580 km. on the northern littoral of the Bay of Bengal (Wikipedia, n.d.). The climate is tropical monsoon, with large seasonal variations in rainfall, high temperatures, and high humidity (Wikipedia, n.d.). From March to June, there is a hot, muggy summer; from June to November, there is a hot, humid, and rainy monsoon season; and from December to February, there is a warm-hot, dry winter. In most areas of this country, April is the hottest month, while January is the coldest. Sylhet is a large city in north-eastern Bangladesh. It is located on the bank of the Surma River and is surrounded by the Jaintia, Khasia, and Tripura hills (LGED Sylhet Menu, n.d.). Tea gardens and tropical woodlands are well-known in the Sylhet region. The study is based on monthly maximum temperatures recorded at the Sylhet station.

3. Materials and Methods

3.1 Sources of Data

The study considers a secondary data set of the monthly maximum temperature of Sylhet station from January 2002 to December 2012. The relevant data were collected from the website of the Bangladesh Meteorological Department (BMD). The data covers the maximum temperature of 132 months in the Sylhet station over 11 years.

3.2 Probability Distributions

To describe the maximum temperature data, it is required to select the probability distribution that best fits the data. Several authors have proposed various probability distributions to characterize the features of maximum temperature data in the recent past. To examine the goodness-of-fit of maximum temperature data for Sylhet station, ten continuous probability distributions were considered: Exponential, Gamma, Log-Gamma, Beta, Normal, Log-Normal, Erlang, Power Function, Rayleigh, and Weibull. The probability density function, parameters, and the range of the above-selected distributions are exhibited in the following Table 1:

Table 1: Probability density function, parameters, and the range of selected distributions

Distributions	Probability density function (pdf)	Parameters	Range
Exponential	$f(x) = \theta \exp(-\theta x)$	θ = inverse scale parameter ($\theta > 0$)	$0 \leq x < +\infty$
Gamma	$f(x) = \dfrac{x^{\alpha-1}}{\beta^\alpha \Gamma(\alpha)} \exp(-x/\beta)$	α = shape parameter ($\alpha > 0$) β = scale parameter ($\beta > 0$)	$0 \leq x < +\infty$
Log-Gamma	$f(x) = \dfrac{(\ln(x))^{\alpha-1}}{x\beta^\alpha \Gamma\alpha} \exp(-\ln(x)/\beta)$	α = shape parameter ($\alpha > 0$) β = scale parameter ($\beta > 0$)	$0 < x < +\infty$
Beta	$f(x) = \dfrac{1}{B(\alpha_1,\alpha_2)} \dfrac{(x-a)^{\alpha_1-1}(b-x)^{\alpha_2-1}}{(b-a)^{\alpha_1+\alpha_2-1}}$	α_1 = shape parameter ($\alpha_1 > 0$) α_2 = shape parameter ($\alpha_2 > 0$) a,b = boundary parameters ($a < b$)	$a \leq x \leq b$
Normal	$f(x) = \dfrac{1}{\sigma\sqrt{2\pi}} \exp\left(-\dfrac{1}{2}\left(\dfrac{x-\mu}{\sigma}\right)^2\right)$	μ = location parameter σ = scale parameter ($\sigma > 0$)	$-\infty < x < +\infty$ $-\infty < \mu < +\infty$
Log-Normal	$f(x) = \dfrac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\dfrac{1}{2}\left(\dfrac{\ln(x)-\mu}{\sigma}\right)^2\right)$	μ = location parameter σ = scale parameter ($\sigma > 0$)	$-\infty < x < +\infty$ $-\infty < \mu < +\infty$
Erlang	$f(x) = \dfrac{x^{m-1}}{\beta^m \Gamma(m)} \exp(-x/\beta)$	m = shape parameter (positive integer) β = scale parameter ($\beta > 0$)	$0 \leq x < +\infty$
Power Function	$f(x) = \dfrac{\alpha(x-a)^{\alpha-1}}{(b-a)^\alpha}$	α = shape parameter ($\alpha > 0$) a,b = boundary parameters ($a < b$)	$a \leq x \leq b$

Rayleigh	$f(x) = \dfrac{x}{\sigma^2} \exp\left(-\dfrac{1}{2}\left(\dfrac{x}{\sigma}\right)^2\right)$	σ = scale parameter ($\sigma > 0$)	$0 \le x < +\infty$
Weibull	$f(x) = \dfrac{\alpha}{\beta}\left(\dfrac{x}{\beta}\right)^{\alpha-1} \exp\left(-\left(\dfrac{x}{\beta}\right)^{\alpha}\right)$	α = shape parameter ($\alpha > 0$) β = scale parameter ($\beta > 0$)	$0 \le x < +\infty$

3.3 Maximum Likelihood Estimate (MLE)

Let $x_1, x_2, \ldots\ldots, x_n$ is a random sample drawn from a population X with probability density function $f(x; \theta)$, where θ is an unknown parameter. The likelihood function $L(\theta)$ of a random sample $x_1, x_2, \ldots\ldots, x_n$ is the joint density of random variables. Thus the likelihood function $L(\theta) = \prod\limits_{i=1}^{n} f(x_i; \theta)$. The θ that maximize the likelihood function $L(\theta)$ is called the maximum likelihood estimator of θ, and it is denoted by $\hat{\theta}$. Hence $\hat{\theta} = \dfrac{\text{Arg sup}}{\theta \in \Theta} L(\theta)$, where Θ is the parameter space of θ. The maximum likelihood technique is a sense that selects all the possible values of θ the one most likely to have generated the given observations $x_1, x_2, \ldots\ldots, x_n$ (Sahoo, 2013).

3.4 Goodness-of-Fit Tests (GOF)

The goodness-of-fit test is applied to determine the accuracy of the predicted values using the theoretical probability model. It aids in the selection of the best model from among the available distributions. The goodness-of-fit test is applied to test the following hypothesis:

H_0 : The maximum temperature data follow the specified probability distribution.

H_1 : The maximum temperature data do not follow the specified probability distribution.

For selecting the best probability distribution, several goodness-of-fit tests were applied, including the Kolmogorov-Smirnov test, Anderson-Darling test, and Chi-square test at a ($\alpha = .05$) level of significance (Sharma et al., 2010).

3.4.1 Kolmogorov-Smirnov (K-S) Test

The Kolmogorov-Smirnov test (Chakravarti et al., 1967) is applied to determine if a sample comes from a population with a particular distribution. This test is based on the empirical cumulative distribution function (ECDF). The ECDF for *N* ordered data points $X_1, X_2, \ldots\ldots, X_N$ is defined as

$$E_N = \frac{n(i)}{N}$$

Where, $n(i)$ is the number of points less than X_i and the X_i's are ordered from smallest to the largest value. This is a step function that increases by 1/*N* at the value of each ordered data point. The Kolmogorov-Smirnov test statistic is defined as:

$$D = \frac{\max}{1 \le i \le N}\left(F(X_i) - \frac{i-1}{N}, \frac{i}{N} - F(X_i)\right)$$

Here, *F* denotes the theoretical cumulative distribution of the distribution under test, which must be continuous and fully specified (i.e., the location, scale, and shape parameters cannot be estimated from the data). If the test statistic *D* is greater than the critical value determined from a table, the hypothesis regarding the distributional form is rejected. For the goodness-of-fit test, the K-S test is preferred to the Chi-square test if the sample size is small.

3.4.2 Anderson-Darling (A-D) Test

The Anderson-Darling test (Stephens, 1974) is used to determine if a sample of data is drawn from a population with a particular distribution. This test is a modification of the K-S test in which the tails are assigned more weight than the K-S test. The critical

values of the K-S test are distribution-agnostic, meaning that they are independent of the distribution being examined. To calculate critical values, the A-D test uses the specified distribution. The Anderson-Darling test statistic is defined as follows:

$$A^2 = -N - \frac{1}{N}\sum_{i=1}^{N}(2i-1)\big[\ln F(X_i) + \ln(1 - F(X_{N-i+1}))\big]$$

Here, F is the cumulative distribution function of the specified distribution. It is worth noting that the X_i's are the ordered data. The critical values for the A-D test can be determined by using the distribution under consideration. This test is one-sided, and if the test statistic A is greater than the critical value, the hypothesis that the distribution of a given form is rejected.

3.4.3 Chi-square Test
The Chi-square test is used to see whether a sample originates from a population that has a particular distribution (Sharma and Singh, 2010). This test assumes that there are enough observations to approximate the distribution of the test statistic with the Chi-square distribution. Because this test is performed on grouped data, the value of the test statistic is determined by how the data is grouped. It is to be noted that this test can only be used with continuous sample data. The following formula can be used to calculate the number of groups k based on the sample size N:

$$k = 1 + \log_2(N)$$

The Chi-square test statistic is defined as

$$\chi^2 = \sum_{i=1}^{k}\frac{(O_i - E_i)^2}{E_i}$$

Here, O_i is the observed frequency and E_i is the expected frequency calculated by

$$E_i = F(X_2) - F(X_1)$$

Here, F is the CDF of the probability distribution being tested.

3.5 Determination of the Best-Fitted Probability Distribution
The goodness-of-fit tests listed above were fitted to the maximum temperature data at Sylhet station. The test statistic of each test was computed and tested at a ($\alpha = .05$) significance level. The minimum test statistic value was used to rank distinct probability distributions from 1 to 10. A maximum score of 10 was given to the rank first probability distribution, and fewer scores were given to the distributions with a rank of 2 to 10. For each of the three tests, the distribution with the highest rank was chosen independently. The total score from all three tests was added together to get the best-fitted probability distribution based on the highest score.

4. Statistical Software Used
The analysis of monthly maximum temperature data at the Sylhet station has been completely done with the help of Easy Fit 5.6 Standard and Microsoft Excel software.

5. Results and Discussion
The descriptive statistics for the monthly maximum temperature data at the Sylhet station are summarized in the following Table 2.

Table 2: Summary Statistics for the monthly maximum temperature data at Sylhet station

Statistics	Value
Observations	132
Minimum (in ⁰C)	24.1
Maximum (in ⁰C)	34
Range	9.9
Average	30.62
Standard Deviation (SD)	2.39
Coefficient of Variation (CV)	0.08
Median	31.4
Mode	32.6

Skewness	-0.97
Kurtosis	0.10

From Table 2 it is found that the minimum, maximum, and the average monthly maximum temperature at Sylhet station during the study period were 24.1^{0}C, 34^{0}C, and 30.62^{0}C, respectively. The distribution of monthly maximum temperature data was found to be negatively skewed and platykurtic. It is also observed that the coefficient of variation of temperature for the selected station was around 8.0%.

The average monthly maximum temperature calculated in different years for the Sylhet station is presented in the following Figure 1.
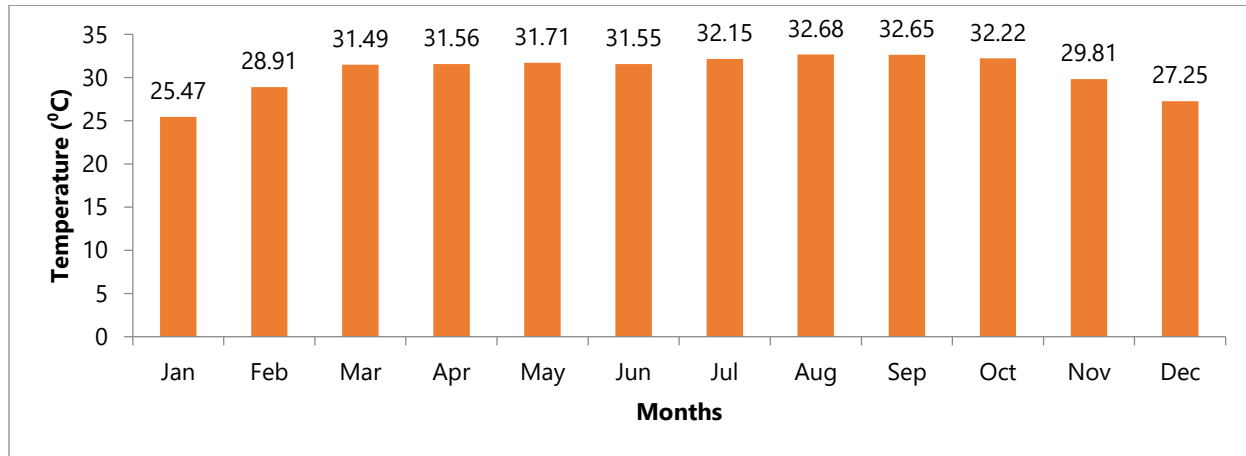

Figure 1: Average monthly maximum temperature of Sylhet station

Figure 1 shows that the average monthly maximum temperature was recorded 32.68^{0}C in August, whereas the minimum was recorded 25.47^{0}C in January over the study period.

The probability distributions along with their test statistics, *P*-value, and critical value for three GOF tests, are presented in Table 3.

Table 3: Different probability distributions and the summary of goodness-of-fit tests

Distributions	Summary of different tests						
	Kolmogorov-Smirnov		Anderson-Darling		Chi-Square		
	Statistics	*P*-Value	Statistics	CV	DF	Statistics	*P*-Value
Exponential	0.5448	.00	51.902	2.5018	6	936.92	.00
Gamma	0.1512	.00	4.9162	2.5018	6	34.019	.00
Log-Gamma	0.1531	.00	5.2919	2.5018	6	42.471	.00
Beta	0.0689	.54	0.7432	2.5018	7	11.592	.11
Normal	0.1424	.00	4.2264	2.5018	6	23.41	.00
Log-Normal	0.1509	.00	5.09	2.5018	6	32.936	.00
Erlang	0.1563	.00	5.1058	2.5018	6	36.578	.00
Power Function	0.0997	.14	1.8939	2.5018	7	18.075	.01
Rayleigh	0.3852	.00	37.297	2.5018	7	562.48	.00
Weibull	0.0985	.14	2.0081	2.5018	7	23.895	.00

*CV = Critical Value at $\alpha = .05$ Significance Level, DF = Degrees of Freedom

From Table 3 the *P*-value for the Kolmogorov-Smirnov test is greater than .05 for Beta, Power Function, and Weibull distributions. So the null hypothesis cannot be rejected at the $\alpha = .05$ significance level and it may be concluded that the monthly maximum temperature data of Sylhet station fit the Beta, Power Function, and Weibull distributions.
Again, the test statistics for the Anderson-Darling test is less than the critical value at the $\alpha = .05$ significance level for Beta, Power Function, and Weibull distributions. So the null hypothesis cannot be rejected at the $\alpha = .05$ significance level and it may be concluded that the monthly maximum temperature data of Sylhet station fit the Beta, Power Function, and Weibull distributions.

Also, the *P*-value for the Chi-square test is greater than .05 for the Beta distribution. Hence the null hypothesis cannot be rejected at the $\alpha = .05$ significance level and it may be concluded that the monthly maximum temperature data of Sylhet station fit the Beta distribution.

Thus the three probability distributions viz. Beta, Power Function, and Weibull were identified as the fitted to the maximum temperature data based on three tests independently.

Now the total score of ten probability distributions based on the rank of each GOF test was summarized. This was done to find the best-fitted probability distribution in addition to the three previously found probability distributions. The maximum overall score based on the sum of individual point scores obtained from three selected goodness-of-fit tests was used to identify the best-fitted probability distribution (Sharma et al., 2010). Table 4 shows the probability distributions that were found to have the highest overall score.

Table 4: Score wise probability distributions

Distributions	Kolmogorov-Smirnov		Anderson-Darling		Chi-Square		Total
	Rank	Scores	Rank	Scores	Rank	Scores	score
Exponential	10	1	10	1	10	1	3
Gamma	6	5	5	6	6	5	16
Log-Gamma	7	4	8	3	8	3	10
Beta	1	10	1	10	1	10	30
Normal	4	7	4	7	3	8	22
Log-Normal	5	6	6	5	5	6	17
Erlang	8	3	7	4	7	4	11
Power Function	3	8	2	9	2	9	26
Rayleigh	9	2	9	2	9	2	6
Weibull	2	9	3	8	4	7	24

Comparing the scores of different distributions, it has been found that the Beta distribution having the highest score of 30 offers the best fit to the monthly maximum temperature data at the Sylhet station in Bangladesh.
Table 5 shows the fitted distributions for monthly maximum temperature data, along with the estimated parameters. The estimated parameters were used to generate random numbers, and the least square approach was employed for temperature analysis (Ghosh et al., 2016).

Table 5: The estimated values of the parameter for the fitted distributions

Distributions	Parameters	Estimated value
Beta	α_1	16.566
	α_2	1.836
	a	-0.436
	b	34.063
Power Function	α	2.131
	a	23.258
	b	34.004
Weibull	α	14.767
	β	31.684

6. Conclusion

The efficiency of different probability distributions was evaluated using a systematic assessment approach to find the best-fitted probability distribution for monthly maximum temperature data at the Sylhet station in Bangladesh. According to the findings, the monthly maximum temperature at the Sylhet station ranged from 24.1^{0}C to 34^{0}C over the study period. The Beta distribution has been found best fit to the monthly maximum temperature data at the Sylhet station in Bangladesh. Identification of the best-fit probability distribution of monthly maximum temperature data could be useful in agriculture and climate studies.

References

[1]   Ahmad, Q. K., Warrick, R. A., Ericksen, N. J., & Mirza, M. Q. (1996). *The Implications of Climate Change for Bangladesh*. Kluwer Academic Publishers.

[2]   Chakravarti, I. M., Laha, R. G., & Roy, J. (1967). *Handbook of Methods of Applied Statistics*. John Wiley and Sons.

[3]   Chowdhury, M. H. K., & Debsharma, S. K. (1992). *Climate change in Bangladesh- A statistical Review*. Report on IOC-UNEP Workshop on Impacts of Sea Level Rise Due to Global Warming, NOAMI, held in Bangladesh.

[4]   Ghosh, S., Roy, M. K., & Biswas, S. C. (2016). Determination of the Best Fit Probability Distribution for Monthly Rainfall Data in Bangladesh. *American Journal of Mathematics and Statistics*, *6*(4), 170-174. DOI: 10.5923/j.ajms.20160604.05

[5]   Hasan, H., Salam, N., & Adam, M. B. (2013). Modeling Extreme Temperature in Malaysia Using Generalized Extreme Value Distribution. *International Journal of Mathematical, Computational, Physical, Electrical and Computer Engineering*, *7*(6), 983-989.

[6]   Hossain, M. M., Abadulla, F., & Rahman, M. H. (2016). Selecting the Probability Distribution of Monthly Maximum Temperature of Dhaka (Capital City) in Bangladesh. *Jahangirnagar University Journal of Statistical Studies*, 33-45.

[7]   Hughes, G. L., Rao, S. S., & Rao, T. S. (2007). Statistical analysis and time-series models for minimum/maximum temperatures in the Antarctic Peninsula. *Proceedings of the Royal Society, Series A*, 463(2077), 241–259. http://doi.org/10.1098/rspa.2006.1766

[8]   Karmakar, S., & Shrestha, M. L. (2000). Recent Climate Change in Bangladesh. SMRC Series, No. 4. Dhaka, *SAARC Meteorological Research Centre*.

[9]   LGED Sylhet Menu. (n.d.). *About Sylhet*. https://oldweb.lged.gov.bd/DistrictLGED.aspx?DistrictID=61.

[10]   Mehrotra, D., & Mehrotra, R. (1995). Climate change and hydrology with emphasis on the Indian Subcontinent. *Hydrological Sciences Journal*, 40(2), 231-242. https://doi.org/10.1080/02626669509491406

[11]   Mia, N. M. (2003, December 10-12). *Variations of Temperature of Bangladesh*. Proceedings of SAARC Seminars on Climate Variability in the South Asian Region and Its Impacts, Dhaka.

[12]   Parthasarathy, B., Sontakke, N. A., Monot, A. A., & Kothawale, D. R. (1987). Drought-flood in the summer monsoon season over different meteorological subdivisions of India for the period 1871-1984. *Journal of Climatology*, 7(1), 57–70. https://doi.org/10.1002/joc.3370070106

[13]   Sahoo, P. (2013). *Probability and Mathematical Statistics*. Louisville, KY 40292 USA.

[14]   Sharma, M. A., & Singh, J. B. (2010). Use of Probability Distribution in Rainfall Analysis. *New York Science Journal*, 3(9), 40-49.

[15]   Stephens, M. A. (1974). EDF Statistics for Goodness of Fit and Some Comparisons. *Journal of the American Statistical Association*, 69(347), 730-737.

[16]   Vivekanandan, N. (2015). Modeling of Annual Extreme Rainfall, Temperature, and Wind Speed Using OSA of EV1 and EV2 Distributions. *International Journal of Innovative Research in Computer Science & Technology*, 3(4), 57-60.

[17]   Watson, R. T., Zinyowera, M. C., Moss, R. H., & Dokken, D. J. (1996). *Impacts, Adaptations, and Mitigation of Climate Change: Scientific-Technical Analyses*. Published for Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge.

[18]   Wikipedia. (n.d). *Climate*. https://en.wikipedia.org/wiki/Geography_of_Bangladesh#Climate.

[19]   Wikipedia. (n.d). *Geography of Bangladesh*. https://en.wikipedia.org/wiki/Geography_of_Bangladesh.