A COMPARATIVE ANALYSIS BETWEEN HISTORICAL RAINFALL DATA-BASED PREDICTION AND FINE PARTICULATE MATTER-BASED RAINFALL PREDICTION FOR BANGLADESH

Arnab Naha Ushna[1], Ishfaq Ahmad Chakladar*[2], Durjoy Datta Mazumder[2] and Ahmed Imtiaz Zamee[3]

[1] Student, Bangladesh University of Engineering and Technology, Bangladesh, e-mail: arnabnushna24@gmail.com
[2] Student, Bangladesh University of Engineering and Technology, Bangladesh, e-mail: ishfaqahmad1nahin000@gmail.com
[2] Student, Rajshahi University of Engineering and Technology, Bangladesh, e-mail: durjoydmazumder1813058@gmail.com
[3] Student, Bangladesh University of Engineering and Technology, Bangladesh, e-mail: ahmedimtiazzamee04054@gmail.com

***Corresponding Author**

ABSTRACT

$PM_{2.5}$ exposure is subjected to various health and environmental issues that needs to be addressed more seriously. It also affects the natural cyclic order of seasonal variation and rainfall occurrences. This study aimed to estimate the mean annual rainfall depth (in mm) in 2024 for the Indian sub-continent region (7 countries: Bangladesh, India, Pakistan, Nepal, Bhutan, Sri Lanka, and Maldives) on the basis of the effect of mean annual $PM_{2.5}$ exposure on rainfall occurrence and compare the results with that of the historical rainfall data. For this purpose, 2 datasets of mean annual $PM_{2.5}$ exposure (in $\mu g/m^3$) and rainfall depth (in mm) in 2010-19 were used. These 2 datasets were analyzed through Random Forest Regression (RFR) and mean annual rainfall depth value (in mm) were estimated for 2024. Findings from this case scenario were, then, compared to that of historical mean annual rainfall depth (1965-2019) dataset. This dataset was analyzed using Linear Regression (LR) and mean annual rainfall depth values (in mm/year) for 2024 were predicted. Furthermore, LR-analyzed values in this case were validated against The World Bank dataset of mean annual rainfall depth (in mm/year) for 2024. Validated LR-analyzed values and RFR-analyzed values of mean annual rainfall depth were compared to check the validity of RFR analysis through ArcGIS plotting and histogram analysis. The result was satisfactory upon the analysis and data visualization, indicating good precision of RFR analysis. Also, the performance of the LR and RFR models were evaluated on the basis of Mean Squared Logarithmic Error (MSLE) and Root Mean Squared Logarithmic Error (RMSLE), and both of them yielded satisfactory values. According to the analyses, the predicted values of mean annual rainfall depth for Bangladesh were better than most of the sub-continental countries (LR prediction: 2105.11 mm; RFR prediction: 2052.03 mm). It indicates that Bangladesh has a good number of rainfall occurrences throughout the years. However, the increasing trend of $PM_{2.5}$ exposure worldwide might be a matter of concern in the future. So, planning, policymaking, interventions, and regulations need to be well-structured to avert the adverse situations. In this regard, inclusion of refined and rich datasets will help to train ML-oriented approaches to develop efficient models to predict on rainfall occurrences more precisely.

Keywords: *Rainfall, $PM_{2.5}$, mapping, prediction, forecast*

1. INTRODUCTION

Air pollution is one of the major concerns for the world. Several human activities, technologies, and structural infrastructures are worsening this issue day-by-day. There are many pollutants which affect the environmental and climatic conditions adversely, such as carbon dioxide (CO_2), carbon monoxide (CO), nitrogen oxides (NO_x), black carbon, and particulate matter (fine particulate matter, $PM_{2.5}$ and ultrafine particulate matter, $PM_{0.1}$). $PM_{2.5}$ is classified as a particulate matter having a diameter of less than 2.5 μm. Significant emphasis is being given to $PM_{2.5}$, one of the major air pollutants, due to its detrimental impacts on local and regional air quality, atmospheric visibility, and the global climate (Cesari et al., 2018; Fuzzi et al., 2015). It has major health implications for mankind and can cause fatal consequences. Not only does $PM_{2.5}$ enter the gas exchange zone of the lungs, but it can also enter the circulatory system (Feng et al., 2016). Its exposure is linked to a number of health problems, including airway inflammation, decrease in the incidence of lung function, an aggravation of asthma and COPD, and a higher risk of infection (Dąbrowiecki et al., 2021; Feng et al., 2016). Over 2.1 million persons in South Asia lost their lives to air pollution in 2020; 1,667,000 of the instances were in India, 235,700 were in Pakistan, 173,500 were in Bangladesh, and 42,100 were in Nepal (Faisal et al., 2022). $PM_{2.5}$ exposure has adverse effects on environmental aspects too. According to California Air Resources Board, $PM_{2.5}$ is associated with reduced visibility and adverse impacts on climatic conditions. Some of the particulate matter (PM) constituents promote climate warming which can hamper the nature cyclic order of seasonal variability. A number of studies on the relationship between $PM_{2.5}$ and meteorological variables were conducted across the world. Anusasananan et al. (2021) conducted a study on the relationship between $PM_{2.5}$ and 2 meteorological variables (i.e., rainfall and temperature) in Chiang Mai, Thailand. This study found that rainfall occurrence reduces $PM_{2.5}$ concentration due to rain's wet scavenging effect on $PM_{2.5}$. Also, for a period of 1 day, $PM_{2.5}$ concentration decreases in day time and increases at night. A similar study was conducted in Delhi, India by Chate et al. (2012). In this study, mass variation of particulate matters (PM_{10}, $PM_{2.5}$, and PM_1) was assessed during monsoon and winter seasons. Concentration of these 3 PMs was in 20-180 μg/m^3 range in monsoon and 100-500 μg/m^3 range in winter. This study also revealed higher mass concentration of PMs with higher relative humidity and lower ambient temperature.

However, there has not been any significant research on the rainfall prediction on the basis of $PM_{2.5}$ exposure in Bangladesh. In this study, the association between mean annual $PM_{2.5}$ concentration and mean annual rainfall depth was assessed and a relationship between these 2 variables was developed through Artificial Neural Network (ANN) to predict the future mean annual rainfall depth for the year 2024 in Bangladesh. To compare the findings from the developed ANN model, historical mean annual rainfall depth data was processed through Linear Regression (LR) for similar type of prediction.

2. METHODOLOGY

2.1 Study Area

Indian sub-continent is a physio-graphical region located in South Asia with diverse cultures and climatic conditions. There are 7 countries in this region, namely Bangladesh, India, Pakistan, Nepal, Bhutan, Sri Lanka, and Maldives. This study aimed to analyze the climatic condition of each country on the basis of a meteorological parameter (rainfall) and an air pollutant (fine particulate matter, $PM_{2.5}$), evaluate the findings with respect to the historical mean annual rainfall depth dataset of each country, and highlight the finding from the dataset of Bangladesh to assess the climatic condition.

2.2 Dataset Description

2.2.1 Historical Mean Annual Rainfall Depth Dataset

This dataset consists of mean annual rainfall depth (in millimeters, mm) of the aforementioned 7 countries in the year range of 1965-2019. It was retrieved from Climate Change Knowledge Portal

(CCKP) of The World Bank (except India). Global data on past, present, and future vulnerabilities, and impacts are available on CCKP (https://climateknowledgeportal.worldbank.org). It is a collection of climate-related information for The World Bank Group (WBG) to assess their current and future policy and investment decisions. On the other hand, India's historical mean annual rainfall depth data were mainly retrieved from 114 Years of Rainfall in India - Interactive (n.d.). Along with that, some missing data were available at another website (https://tradingeconomics.com).

2.2.2 Historical Mean Annual $PM_{2.5}$ Exposure Dataset

It contains mean annual $PM_{2.5}$ exposure (in micrograms per cubic meter, µg/m^3) of the countries during 2010-19 period. This dataset was retrieved from World Development Indicators Databank of The World Bank (https://databank.worldbank.org). This databank provides vast datasets on various topics, such as economic indicators, educational aspects, and environmental indices of different countries for a certain period of time. Similar to the previous dataset, it is also free of cost and reliable.

2.3 Framework

2.3.1 Linear Regression (LR)

Linear regression (LR) is a supervised Machine Learning (ML) model, which is deployed to model the continuous variables and predict values for a certain input. It creates a straight line to fit the variables while minimizing the net deviation of the distance between the variables and line. LR analysis is a basic and simplest form of data analysis. However, its main disadvantage is that it does not perform well for the complex set of data. When the variables are too much scattered or dispersed, i.e., non-linearity issues arise, LR performs badly. In that case, polynomial regression can be used for a better performance. Its simplest formulation is:

$$y = a + bx \tag{1}$$

Here, 'y' is dependent variable, 'a' is the intercept (y-axis), 'b' is the slope of the straight line, and 'x' is independent variable.

2.3.2 Random Forest Regression (RFR)

Random Forest Regression (RFR) is an ensemble-learning algorithm that combines a large set of regression trees. Regression trees are hierarchically arranged sets of constraints or conditions that are applied one after the other from the root to the leaf of the tree. RFR starts with a large number of randomly selected bootstrap samples from the initial training dataset. For every bootstrap sample, a regression tree is fitted. A tiny subset of the whole set of input variables is randomly considered for binary partitioning for each node in the tree. The regression tree splitting criterion is based on choosing the input variable with the lowest Gini Index:

$$I_G\left(t_{X(x_i)}\right) = 1 - \sum_{j=1}^{m} f\left(t_{X(x_i)}, j\right)^2 \tag{2}$$

2.4 Performance Evaluation

To evaluate the performance of the aforementioned 2 models, Mean Squared Logarithmic Error (MSLE) and Root Mean Squared Logarithmic Error (RMSLE) were used. MSLE is used to assess how accurate a regression model is, especially when the predicted values take the form of logarithmic scales. On the other hand, RMSLE is used when it is preferable to refrain from overly penalizing significant discrepancies between the observed and projected values, particularly when those values are quite high.

$$MSLE = \frac{1}{n}\sum_{i=1}^{n}(\log(Y_a) - \log(Y_p))^2 \tag{3}$$

Here, 'n' is the complete count of the data elements, 'Y_a' is actual y values, and 'Y_p' is the predicted y values.

$$RMSLE = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(\log(p_i + 1) - (\log(a_i + 1))^2} \tag{4}$$

Here, 'n' is the total number of observations in the dataset, 'p_i' is the target prediction, and 'a_i' is the actual target for i.

3. RESULTS AND DISCUSSIONS

3.1 Historical Rainfall Data-based Prediction

3.1.1 Preliminary Data Analysis

Historical mean annual rainfall depth datasets retrieved from the aforementioned sources were pre-processed through missing data correction and outlier removal. Pre-processed data were then manually inputted into Excel (Microsoft, Seattle, WA, USA) for preliminary analysis and scatter plots were developed for each country in the same year range (1965-2019).

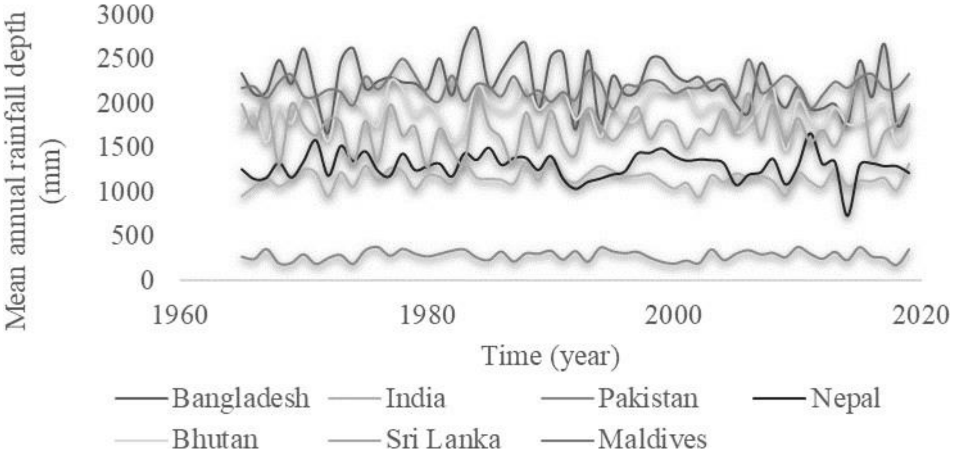
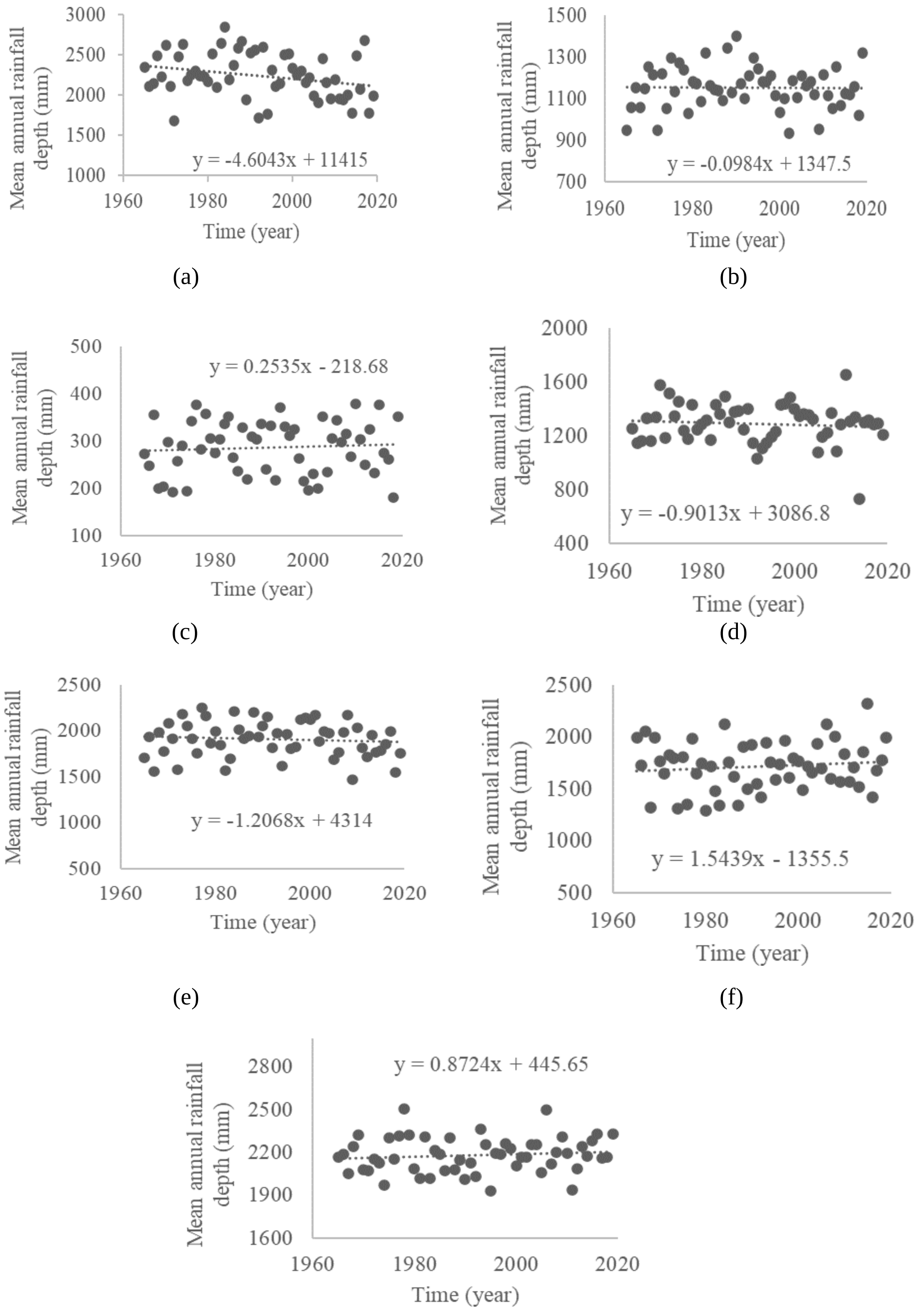


Figure 1: Scatter plot of the historical mean annual rainfall depth (mm) datasets of the Indian sub-continental countries (1965-2019)

3.1.2 LR Analysis

After the preliminary analysis of the datasets through scatter plotting, LR analysis were considered to estimate the mean annual rainfall depth (mm) for the period of 2020-24. LR analysis yielded an equation to set up a best-fit straight line, in which timespan (year) and mean annual rainfall depth (mm) were considered to be explanatory or independent variable (x) and dependent variable (y), respectively. The performance of LR models were evaluated through MSLE and RMSLE, which would be discussed in 'Performance Evaluation' section. Upon finding the equations for each of the country, mean annual rainfall depth values (mm) were predicted for the following 5 years (2020-24). Again, the predicted values were averaged to convert the values in mm/year unit and yield a constant value for further analysis for 2024.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

Figure 2: Model development through LR analysis for (a) Bangladesh, (b) India, (c) Pakistan, (d) Nepal, (e) Bhutan, (f) Sri Lanka, and (g) Maldives (1965-2019)

Table 1: Estimation of mean annual rainfall depth (mm/year) from the developed LR models

Year	**ean annual rainfall depth (mm)**						
	Bangladesh	**India**	**Pakistan**	**Nepal**	**Bhutan**	**Sri Lanka**	**Maldives**
2020	2114.314	1148.732	293.39	1266.174	1876.264	1763.178	2207.898
2021	2109.7097	1148.6336	293.6435	1265.2727	1875.0572	1764.7219	2208.7704
2022	2105.1054	1148.5352	293.897	1264.3714	1873.8504	1766.2658	2209.6428
2023	2100.5011	1148.4368	294.1505	1263.4701	1872.6436	1767.8097	2210.5152
2024	2095.8968	1148.3384	294.404	1262.5688	1871.4368	1769.3536	2211.3876
Mean (mm/year)	2105.11	1148.54	293.90	1264.38	1873.86	1766.27	2209.65

Furthermore, estimated mean annual rainfall depth (mm/year) values were compared with that of World Development Indicators Databank of The World Bank for validation check and overall comparative analysis was satisfactory.

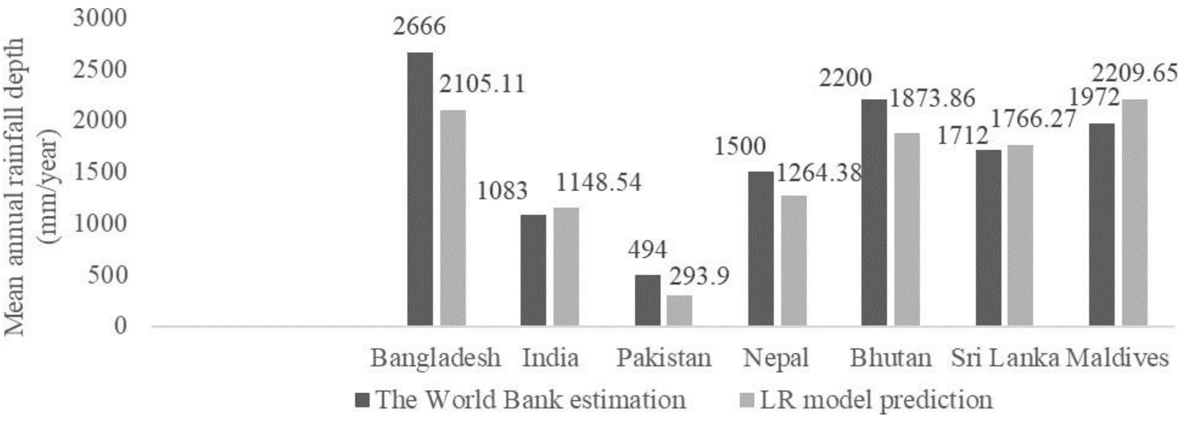


Figure 3: Comparative histogram visualization (2024)

3.2 Particulate Matter-based Rainfall Prediction

3.2.1 Preliminary Data Analysis

Preliminary analysis of historical mean annual $PM_{2.5}$ exposure datasets were accomplished similar to that of historical mean annual rainfall depth and scattered plots were developed.

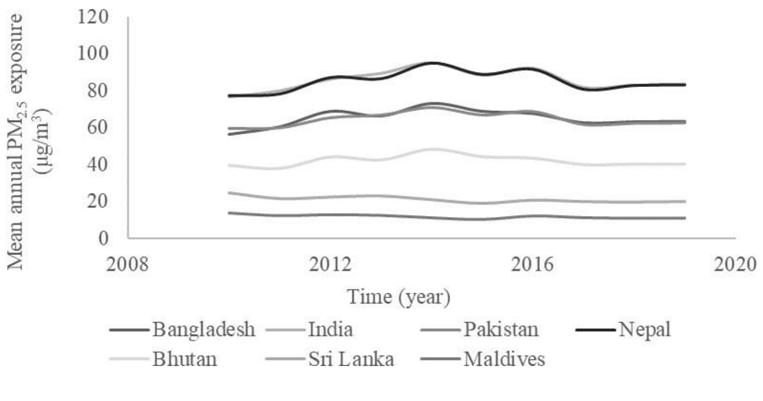


Figure 4: Scatter plot of the historical mean annual $PM_{2.5}$ exposure (µg/m^3) datasets of the Indian sub-continental countries (1965-2019)

3.2.2 RFR Analysis

After the preliminary analysis of the datasets through scatter plotting, LR analysis were considered to estimate the mean annual rainfall depth (mm) for the period of 2020-24. LR analysis yielded an equation to set up a best-fit straight line, in which timespan (year) and mean annual rainfall depth (mm) were considered to be explanatory or independent variable (x) and dependent variable (y), respectively. The performance of LR models were evaluated through MSLE and RMSLE, which would be discussed in 'Performance Evaluation' section. Upon finding the equations for each of the country, mean annual rainfall depth values (mm) were predicted for the following 5 years (2020-24). Again, the predicted values were averaged to convert the values in mm/year unit and yield a constant value for further analysis for 2024.

Table 2: Estimation of mean annual rainfall depth (mm) from the developed RFR models (2024)

Country	**Mean annual PM2.5 exposure ($\mu g/m^3$)** * Estimated through LR	**Mean annual rainfall depth (mm)** * Estimated through RFR
Bangladesh	68.12	2052.03
India	89.03	1237.62
Pakistan	66.06	310.20
Nepal	88.63	1246.25
Bhutan	42.21	1722.41
Sri Lanka	16.76	2193.94
Maldives	9.17	2278.37

3.3 Performance Evaluation

In this study, 2 ML models were deployed for the purpose of model development and value prediction. The accuracy to predict the values, i.e., performance of these models was evaluated through MSLE and RMSLE.

Table 3: Performance evaluation of LR and RFR models

Countries	**LR**		**RFR**	
	MSLE	**RMSLE**	**MSLE**	**RMSLE**
Bangladesh	0.013	0.115	0.004	0.063
India	0.007	0.085	0.002	0.045
Pakistan	0.038	0.196	0.012	0.063
Nepal	0.014	0.118	0.008	0.089
Bhutan	0.009	0.098	0.0009	0.031
Sri Lanka	0.017	0.13	0.003	0.055
Maldives	0.003	0.053	0.0006	0.024

From this table, it can be observed that the values of MSLE and RMSLE are remarkably small, i.e., approximately zero, indicating the models' good efficiency for the designated purpose. Forexample, in the case of Bangladesh, LR model yielded MSLE and RMSLE values 0.013 and 0.115, respectively. On the other hand, RFR model performed better, yielding 0.004 (MSLE) and 0.063 (RMSLE). Hence, the performance of these 2 models can be remarked as satisfactory.

3.4 Comparative Analysis

Predictions for the both case scenarios have been made in the data analysis along with the performance evaluation of the models for validation purpose. To visualize the prediction comparison between the 2 cases, map plotting and histogram were used. Plotting of the maps was performed using ArcGIS (esri, Redlands, CA, USA), whereas histogram analysis was done by Excel (Microsoft, Seattle, WA, USA).

3.4.1 Prediction Mapping

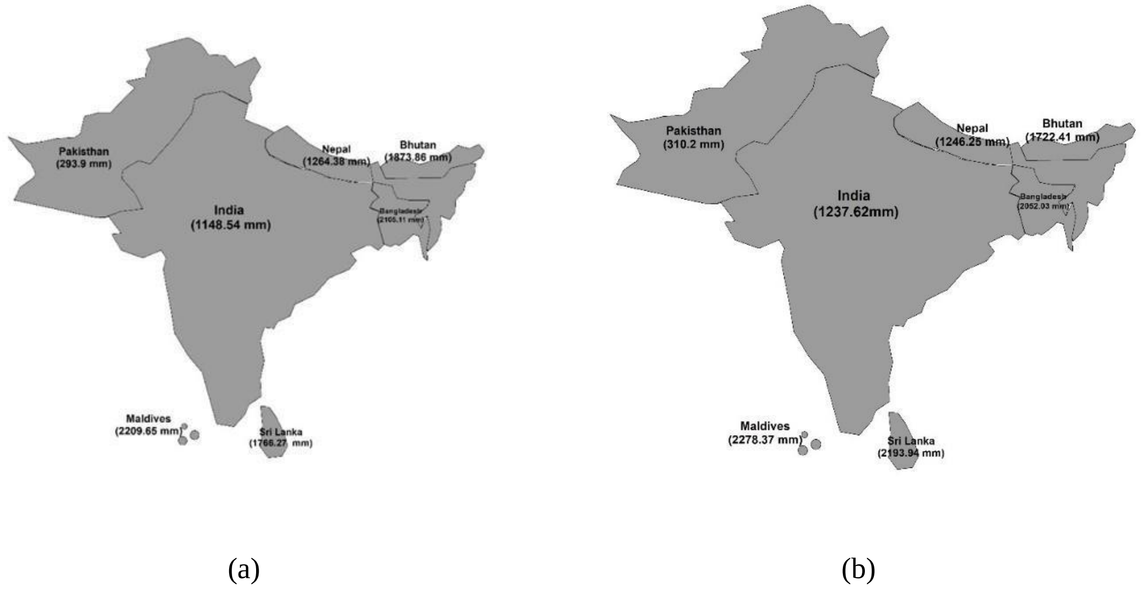


Figure 5: Comparative ArcGIS mapping visualization of mean annual rainfall depth (mm) between (a) LR and (b) RFR analyses (2024)

3.4.2 Histogram Analysis

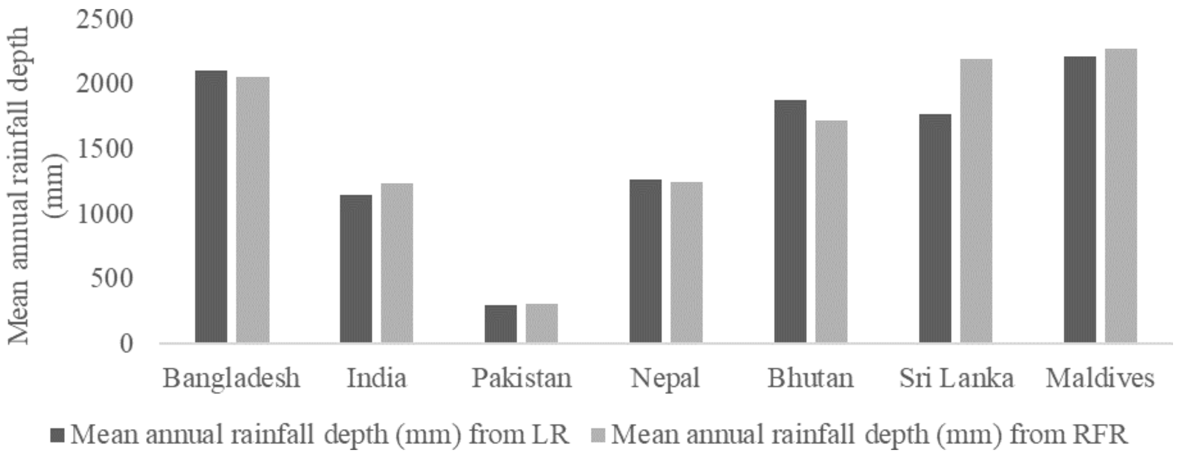


Figure 6: Comparative histogram visualization of mean annual rainfall depth (mm) between LR and RFR analyses (2024)

4. CONCLUSIONS

According to the LR analysis of historical mean annual rainfall depth dataset and The World Bank estimations of mean annual rainfall depth, Bangladesh will likely to have experiencing more rainfall occurrences in 2024 than most of the Indian sub-continental countries (LR: 2105.11 mm/year; The World Bank: 2666 mm/year). On the other hand, even after being one of the most polluted countries in that region on the basis of mean annual $PM_{2.5}$ exposure (68.12 μg/m^3 in 2024), mean annual rainfall depth value of Bangladesh ranked 3rd according to the RFR analysis (2052.03 mm for the year 2024). However, the developed models here are basic and give a rough visualization of the reality as there is not enough $PM_{2.5}$ exposure datasets on the basis of year and country. It can be made more efficient and precise with the inclusion of more datasets of $PM_{2.5}$ exposure and then, more ML models can be deployed to predict the effects of $PM_{2.5}$ exposure on the meteorological variables. Till now, there has not been much datasets in this regard as the research on this topic is relatively new and records date back only to the dawn of the 21st century. Nevertheless, upon considering all the analyses, it is safe to say that rainfall occurrences in Bangladesh have not reduced to a significant level due to the air pollution issue. But things might get worse as the concentration of several air pollutants like $PM_{2.5}$ has been on an increasing trend, which can influence the meteorological variables and change the natural cyclic order of seasonal variability. To tackle this issue and minimize the adverse and detrimental impacts of air pollution, the Government of Bangladesh (GoB) and the relevant regulatory bodies need to impose strict interventions to lessen the emission of air pollutants and implement the specified rules for a sustainable and eco-friendly future, which will ensure the living quality of the citizens of Bangladesh.

ACKNOWLEDGEMENTS

The authors want to express their cordial gratitude to The World Bank Group (WBG) for providing free access to retrieve relevant datasets for this study. Also, the authors want to cordially thank Hemel Sarkar, Undergrad Senior (Department of Water Resources Engineering), Bangladesh University of Engineering and Technology (BUET), Dhaka (Bangladesh) for his supportive assistance and guidelines for ArcGIS plotting.

REFERENCES

Anusasananan, P., Morasum, D., Suwanarat, S., & Thangprasert, N. (2021). Correlation between PM2.5 and meteorological variables in Chiang Mai, Thailand. *Journal of Physics: Conference Series*, *2145*(1), 012045. https://doi.org/10.1088/1742-6596/2145/1/012045

Cesari, D., De Benedetto, G. E., Bonasoni, P., Busetto, M., Dinoi, A., Merico, E., Chirizzi, D., Cristofanelli, P., Donateo, A., Grasso, F. M., Marinoni, A., Pennetta, A., & Contini, D. (2018). Seasonal variability of PM2.5 and PM10 composition and sources in an urban background site in Southern Italy. *Science of The Total Environment*, *612*, 202–213. https://doi.org/10.1016/j.scitotenv.2017.08.230

Chate, S. T., D. M., Ali, P. P., Kaushar, & Bisht, D. S. (2012). Variations in Mass of the PM10, PM2.5 and PM1 during the Monsoon and the Winter at New Delhi. *Aerosol and Air Quality Research*, *12*(1), 20–29. https://doi.org/10.4209/aaqr.2011.06.0075

Dąbrowiecki, P., Adamkiewicz, Ł., Mucha, D., Czechowski, P. O., Soliński, M., Chciałowski, A., & Badyda, A. (2021). Impact of Air Pollution on Lung Function among Preadolescent Children in Two Cities in Poland. *Journal of Clinical Medicine*, *10*(11), Article 11. https://doi.org/10.3390/jcm10112375

Faisal, A.-A.-, Kafy, A.-A., Abdul Fattah, Md., Amir Jahir, D. Md., Al Rakib, A., Rahaman, Z. A., Ferdousi, J., & Huang, X. (2022). Assessment of temporal shifting of PM2.5, lockdown effect, and influences of seasonal meteorological factors over the fastest-growing megacity, Dhaka. *Spatial Information Research*, *30*(3), 441–453. https://doi.org/10.1007/s41324-022- 00441-w

Feng, S., Gao, D., Liao, F., Zhou, F., & Wang, X. (2016). The health effects of ambient PM2.5 and potential mechanisms. *Ecotoxicology and Environmental Safety*, *128*, 67–74. https://doi.org/10.1016/j.ecoenv.2016.01.030

Fuzzi, S., Baltensperger, U., Carslaw, K., Decesari, S., Denier van der Gon, H., Facchini, M. C., Fowler, D., Koren, I., Langford, B., Lohmann, U., Nemitz, E., Pandis, S., Riipinen, I.,Rudich, Y., Schaap, M., Slowik, J. G., Spracklen, D. V., Vignati, E., Wild, M., … Gilardoni,

S. (2015). Particulate matter, air quality and climate: Lessons learned and future needs. *Atmospheric Chemistry and Physics*, *15*(14), 8217–8299. https://doi.org/10.5194/acp-15- 8217-2015

114 Years of Rainfall in India—Interactive. (n.d.). Retrieved January 3, 2024, from http://www.indiaenvironmentportal.org.in/media/iep/infographics/Rainfall%20in%20India/112%20years%20of%20rainfall.html