Int. J. Nonlinear Anal. Appl. 13 (2022) 1, 2897–2909
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2021.21639.2284

Analysis of thunderstorms in Bangladesh using ARIMA model

Kalyan Das[a,*], Md Lutfor Rahman[b], M N Srinivas[c], Anisha Das[d], Vijay Kumar[a]

[a]*Department of Basic and Applied Sciences, National Institute of Food Technology Entrepreneurship and Management, HSIIDC Industrial Estate, Kundli – 131028, Haryana, India.*
[b]*Institute of Statistical Research & Training (ISRT),University of Dhaka,Dhaka - 1000, Bangladesh.*
[c]*Department of Mathematics, School of Advanced Sciences, Vellore Institute of Technology, Vellore-632014, India.*
[d]*Department of Statistics and Biostatistics, Florida State University, USA.*

(Communicated by Madjid Eshaghi Gordji)

Abstract

In this paper our main goal is to study the climatology and variability of the frequency of thunderstorm days over Bangladesh region throughout the year. It has been found that the mean thunderstorm days increase significantly from March to May, i.e. during the pre-monsoon season, although the graphical devices show that there does not seem to be much deviation from the occurrences of thunderstorms each year. The mean monthly and seasonal thunderstorm days were maximum in 1993, followed by that in 1997; whereas it was a minimum in the year 1980, with an extension in its frequency in the subsequent years 1981 and 1982. The coefficient of variation of both annual and seasonal thunderstorm days is minimum over the areas of maximum frequency of mean thunderstorm days and vice-versa. The time-domain analysis confirms that the occurrence happened to be maximum in the year 1991, although each and every state did not witness thunderstorms every year. Also some other time-domain models like autocorrelation and seasonal integrated moving average provide adequate evidence for exploring the number of thunderstorms which happen to confirm the trend of occurrence of thunderstorm over the years.

Keywords: Thunderstorm, frequency, differential equation, interpolation, extrapolation, autocorrelation function, ARIMA.
2010 MSC: Primary: 62M10, 65B05, 65L06; Secondary: 03C40, 32E30, 97N50

*Corresponding author
Email addresses: daskalyan27@gmail.com (Kalyan Das), lutfor@isrt.ac.bd (Md Lutfor Rahman), mnsrinivaselr@gmail.com (M N Srinivas), anisha.das14@gmail.com (Anisha Das), vijay.niftem@gmail.com (Vijay Kumar)

Received: October 2020 *Accepted:* September 2021

1. Introduction

Thunderstorms are the meso-scale phenomena, which develop from cumulonimbus clouds and are characterised by lightning discharges. They usually occur in the form of strong gusts, hail and heavy rainfall, thus causing a huge amount of hazard to aviation and river navigation, as well as damage to standing crops, especially in a coastal country like Bangladesh, although in some regions it is found that despite heavy destruction, the shower has a positive impact on agriculture[1].These winds or thunderstorms are locally called "Nor'westers" or *kalbaisakhi*, and occur during the pre-monsoon season, i.e. between March and May when the temperature in these areas is very high. Here we try to picture out the occurrence of thunderstorms in Bangladesh in the form of time series analysis. We have used here Minitab 17 and SPSS 20 to serve our purpose.

Time series modelling is a vast research area which has attracted the attention of researchers over the last few decades. The main aim of time series modelling is to carefully collect and rigorously study the past observations of a time series to develop an appropriate model which can best describe the inherent structure of the series; this model can then be used to generate future values for the series, i.e. to make forecasts[2].Generally a time series $\{x(t),\ t = 0,\ 1,\ 2,\ \dots\}$ is assumed to follow certain probability model[3] which describes the joint distribution of the random variable x_t and the sequence of observations of the series is actually a sample realization of a stochastic process that produced it. A usual assumption is that the time series variables x_t are independent and identically distributed (i.i.d.) following the normal distribution. While building a proper time series model, the principle of parsimony is accounted for [4, 5, 7, 8]. The idea of model parsimony is similar to the famous Occam's razor principle[6]. As discussed by Hipel and McLeod[6], one aspect of this principle is that when faced with a number of competing and adequate explanations, the simplest one is picked – this forms a basis to logical analysis.

The objective of this project is to present a systematic and comprehensive mathematical analysis of the total frequency of the thunderstorms occurring in Bangladesh annually. We observe that in some years, there is zero occurrence of thunderstorm for a particular state; whereas in the same year, the meteorologists record a huge number of thunderstorms for another state. So, we ignore any fluctuations occurring in different districts and take into account the entire country for a period of 37 years i.e. from 1980-2016. Here we have incorporated mathematical tools to serve our purpose, focussing particularly on the use of stochastic models and differential equations. The pre-monsoon season includes the months of March-May when most of the thunderstorms occur in Bangladesh. The space and time distribution of the number of thunderstorms and their variability together with their probabilistic frequency are very essential especially for aviation and navigation purposes. It is important to note that the thunderstorms' days are actually the days which include only thunderstorm but no precipitation at the time of observation. Slight or moderate thunderstorm occurs without hail but may be accompanied by rain and/ or snow at the time of observation. Heavy thunderstorm may be combined with dust storm or sandstorm as well as hail at the time of observation[9].

2. Preliminaries

ARIMA Methodology:
Auto-Regressive Integrated Moving Average (ARIMA) Model introduced by Box and Jenkins (1976) includes autoregressive as well as moving average parameters, and explicitly includes differencing in the formulation of the model. Specially, the three types of parameters in the model are: the autoregressive parameters (p), the number of differencing passes (d), and moving average parameters (q).

ARIMA models are summarized as ARIMA (p, d, q).

Identification: Input series of ARIMA needs to be stationary, that is, it should have a constant mean, variance and autocorrelation throughout the time period. Therefore the series need to be differenced until it is stationary. The number of times the series need to be differenced is reflected in the d parameter. The appropriate level of differencing can be determined by examining the plot of the data and autocorrelogram. At identification stage we also need to decide about the appropriate number of autoregressive (p) and moving average (q) parameters, which is necessary to yield an effective but still parsimonious model of the process. Generally, the number of the p and q parameters very rarely needs to be greater than two.

Estimation and Forecasting: Parameters are estimated using function minimization procedure, so that the sum of squared residuals is minimized. The estimates of the parameters are used in the last stage to calculate new predicted values of the series and confidence intervals for the predicted values. The estimation process is performed on transformed or differenced data and before the forecasts are generated, the series needs to be integrated, so that the forecasts are expressed in values compatible with the input data. This automatic integration feature is represented by the letter I in the name of the methodology ARIMA – Auto-Regressive Integrated Moving Average.

Identification of Number of parameters to be estimated: Before estimation, for ARIMA, one needs to identify specific number and type of ARIMA parameters to be estimated. For identification of parameters, plot of the series, correlograms of auto correlation (ACF) and partial autocorrelation (PACF). An empirical time series patterns can be sufficiently approximated using one of the five basic models that can be identified based on the shape of ACF and PACF. Also, the number of parameters of each kind is almost never greater than two; it is often practical to try alternative models on the same data.

Parameter	**ACF**	**PACF**	**Correlation**
One autoregressive (p)	Exponential decay	Spike at lag 1	No correlation for other lags
Two autoregressive (p)	A sine-wave shape pattern or a set of exponential decays	Spikes at lags 1 and 2	No correlation for other lags
One moving average (q)	Spike at lag 1	Damps out exponentially	No correlation for other lags
Two moving average (q)	Spikes at lags 1 and 2	A sine-wave shape pattern or a set of exponential decays	No correlation for other lags
One autoregressive (p) and One moving average (q)	Exponential decay starting at lag 1	Exponential decay starting at lag 1	

Materials and Methods: The data of thunderstorms is collected for 37 years from Bangladesh region from 1980 to 2016. For analyzing the data, time series based model named as ARIMA (Auto-Regressive Integrated Moving Average Model) is used using SPSS software. ARIMA models are a very general class of time series models. Their construction is based on the phenomenon of autocorrelation. These models form the basis for the forecasting method known as Box–Jenkins method [7]. They can be used to model stationary time series and those non-stationary time series which can be transformed into stationary ones.

There are three basic models of this class such as autoregression models (AR), moving average models (MA) and mixed autoregression and moving average models (ARMA). The symbol I used in the model name indicates that a time series was subject to differencing. In order to formulate an ARIMA model (p, d, q) a notation stipulating the row of individual model components is used: autocorrelation – p, differencing – d, moving average – q. The process of model construction consists of parts relating to: identification, estimation, diagnostic checking [1, 3]. The general form of ARIMA (p,d,q) model can be written as follows:

$$Y_t = \mu + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} \tag{2.1}$$

Where, Y_t is the observed value of the time series at the stage of t, e_t is the deviation of time series at the stage of t, $\{\varphi_1, \varphi_2, \ldots., \varphi_p\}$ is the autoregressive coefficients, and$\{\theta_1, \theta_2, \ldots., \theta_q\}$ is the moving average coefficient and μ is constant. When p=0, the model is called the moving average model, denoted by MA(q); when q=0, the model is called autoregressive model, denoted by AR(p).

3. Modelling Process

Modelling process is as follows

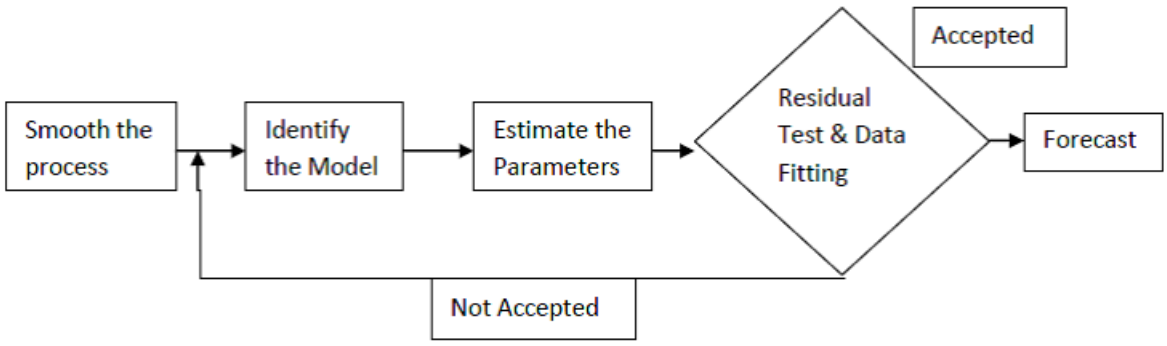

Figure-1

Figure-1: The process of building ARIMA model [10].

In the first stage, the initial identification of a time series in terms of stationarity is made through checking the function of autocorrelation (ACF) and partial autocorrelation (PACF). The fact that in stationary processes the autocorrelation function decreases (as a rule, quite rapidly) is used to

determine whether the process is stationary or non-stationary. If the decrease of the autocorrelation function is very slow, it means that the time series is non-stationary and should be reduced to the stationary form using differencing. During the second phase the parameters of given models are estimated.

The identification phase provides information on the possible variants of the process. The final choice is based on an analysis of several criteria: relevance/validity of (the) model parameters, mean squared error, information criterion. Then the model is subject to the analysis of properties of the model residuals. If the model residuals are a white noise process and there are no significant values of functions ACF or PACF of the model residuals, the model can be used in forecasting. Otherwise, another model should be chosen or the model should be identified again.

Having estimated the model parameters and their statistical significance checked, an assessment of model fit should be made [3, 9]. From the goodness of fit tests based on the analysis of correlation of residuals, the Q test should be used as

$$Q = n\,(n+2)\sum_{\tau=1}^{m}(n-\tau)^{-1}r_{\tau}^{2}(a)$$

where $r_\tau(a)$ is the function of autocorrelation of residuals, and m is the maximum delay of this function. The Q statistics has the distribution of (m-p-q) degrees of freedom. For testing the significance of parameters as well as for building confidence intervals for the forecasts, it is important that a white noise has normal distribution. Finally, the model should be used to make a forecast. The basic difficulty in the use of ARIMA models is the fact that there is no way to automate the procedure for their construction. The thunderstorm data were used for forecasting the thunderstorms by ARIMA models using Box-Jenkins methodology. The Box-jenkins procedure is concerned with fitting a mixed Auto Regressive Integrated Moving Average (ARIMA) model to a given set of data. The main objective of fitting the ARIMA model is to identify the stochastic process of the time series and predict the future values accurately. First appropriate values of different parameters of ARIMA model i.e. p, q and d were found. The tools used for identification are the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) and the resulting correlograms and partial correlograms through SPSS 20.0.

4. Data Smoothing

The sequence plot of frequency of Thunderstorm from 1980 to 2006 in Bangladesh is given below: The below figure captioned as **Figure-2** which indicates that the time series is having increase trend. So the time series is not stationary. To apply time series forecasting model (ARIMA, i.e. Auto-Regressive Integrated Moving Average) series should be stationary, i.e. it should have the following three aspects: a zero average value, constant variance and correlation coefficient only related to time interval and independent of specific time. So to achieve stationary condition for the series, it need to be differenced once, i.e. d=1. After differencing once the result is as follows. Sequence plot of Frequency of Thunderstorm from 1980 to 2006 in Bangladesh.

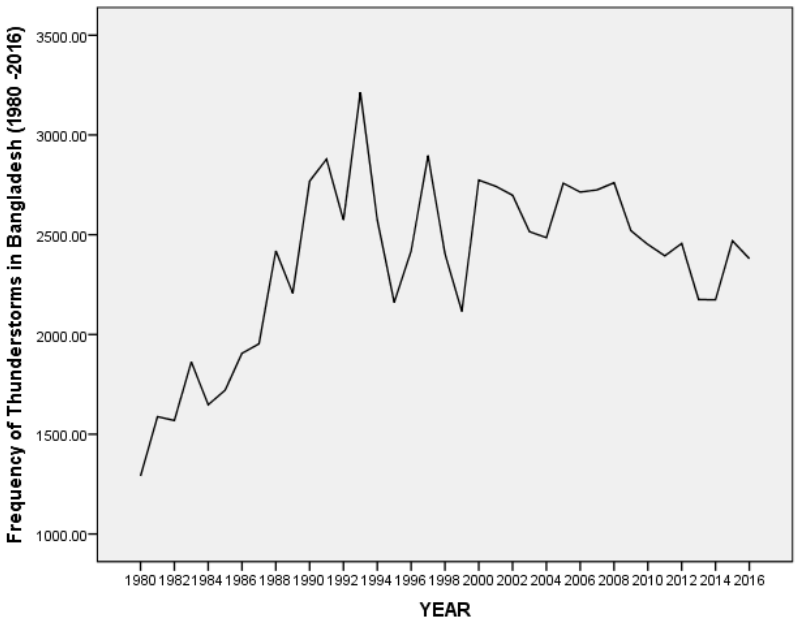

Figure-2

Figure-2: The process of building ARIMA model [10].

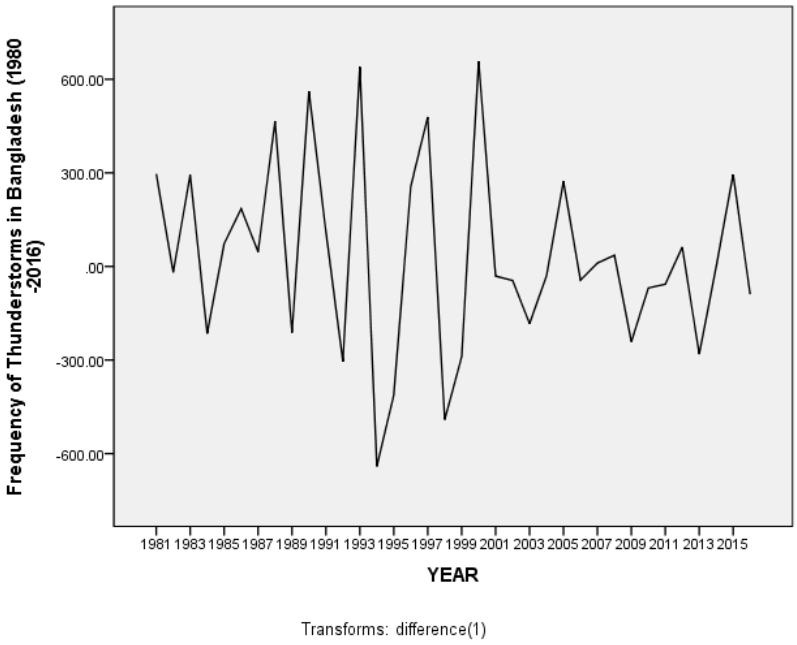

Figure-3

Figure-3: Lag (1) time series which show the stationarity.Figure (3) shows the stationarity, so ARIMA model can be used.

5. Model Identification

Model identification involved the following steps : (1) determine the which model should be adopted, AR (p), MA (q) or ARIMA (p,d,q) and (2) determine the value of p, d and q. The above identification depends on the ACF & PACF properties. The table given below indicate the criteria of order determination.

ACF	PACF	Model Determination
Tailing	Order p tailing	AR(p)
Order q tailing	Tailing	MA (q)
Tailing	Tailing	ARIMA (p,d,q)

Table 1: Order determination of ARIMA model

The analysis of Fig 4 & Fig 5 indicated that the ARIMA (p,d,q) model is appropriate.
Now to finalize appropriate ARIMA (p,d,q) model, we need to decide how many autoregressive (p) and moving average (q) parameters are necessary to develop the effective model. To identify the number of parameters we need to use plot of series, correlograms of auto correlation (ACF) and partial autocorrelation (PACF)

Autocorrelations of Frequency of Thunderstorms in Bangladesh (1980 -2016)

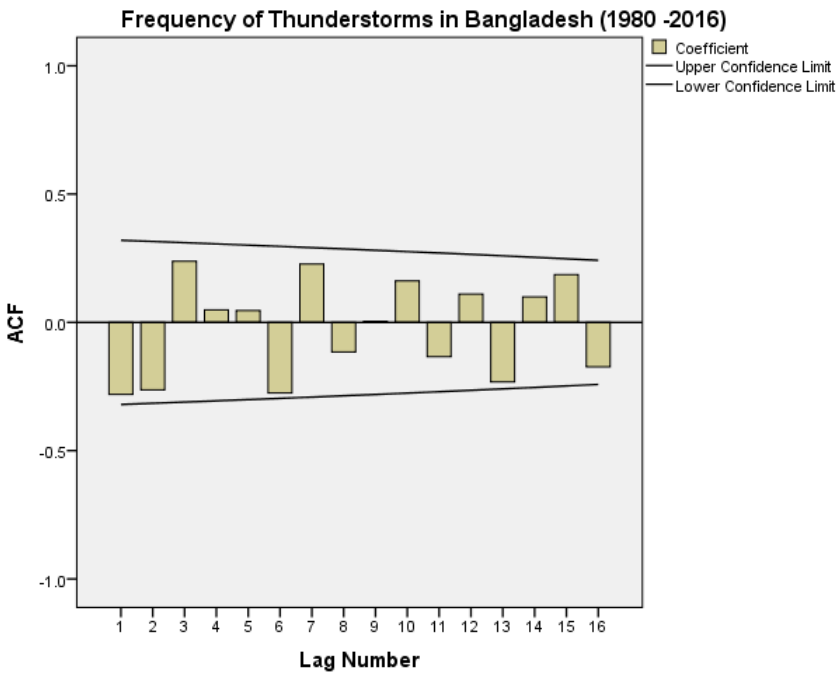

Figure-4

Figure-4: Auto-correlation of series.

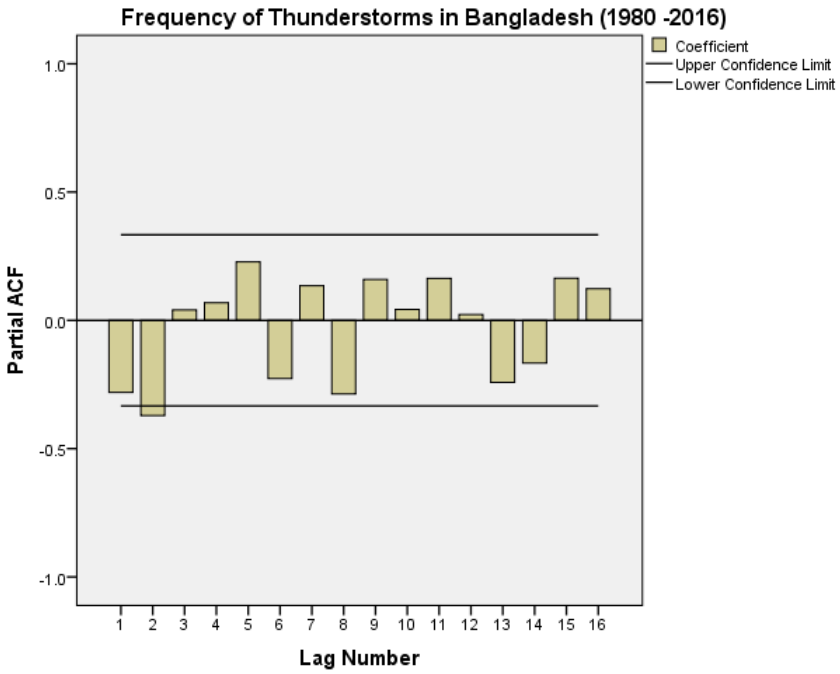

Figure-5

Figure-5: Partial autocorrelations of the series.

ACF (Fig.4) and PACF (Fig. 5) plots indicate that there are no visible patterns. Plots of Auto correlation function (ACF) and Partial Auto correlation function (PACF) are based on first difference, i.e. d=1. So, we get the following four possible models: ARIMA(0,1,0), ARIMA(0,1,1) , ARIMA(1,1,0) ,ARIMA(1,1,1).

Comparison of these models using SPSS is given below in table 2

Model	Smooth R-Square	MAPE	Normalize BIC
ARIMA(0,1,0)	.405 ≈ .41	9.830 ≈ .98	11.559 ≈ 11.56
ARIMA(1,1,0)	.452 ≈ .45	9.631 ≈ .96	11.605 ≈ 11.61
ARIMA(0,1,1)	.489 ≈ .49	9.234 ≈ .92	11.534 ≈ 11.53
ARIMA(1,1,1)	.490 ≈ .49	9.191 ≈ .92	11.662 ≈ 11.66

Table 2: Comparison of ARIMA models

As per the results listed in Table 2, it can be seen that the smooth R-square represents the estimation value of the total variation explained by the models, and the larger the value, the better the fitting degree. MAPE represents the average absolute percentage error. The smaller the value of MAPE, the better model is. Also, one of the best criteria used for evaluation of model, the smaller the value of normalize BIC, the better it is. It is quite obvious that, the final model will be different with different evaluation criterion. So considering the model fitting degree of the history data, i.e. using the MAPE and normalize BIC, we choose ARIMA (0,1,1) model.

6. Model Parameter Estimation

The parameters of ARIMA (0,1,1) (simple exponential smoothing) model can be evaluated by SPSS and are represented under in table 3. The model can be as follows:

$Y_t(predicted) = Y_{t-1} - \theta_1 e_{t-1}$; Where $e_{t-1} = Y_{t-1} - Y_{t-1}(predicted)$

	Estimate	SE	t	Sig.
Difference	1			
MA Lag 1	.433	.156	2.773	.009
Numerator Lag 0	.013	.014	.967	.340

Table 3: ARIMA (0,1,1) model

7. Model Validation

For model validation, we need to test the residual series to determine whether the model has reached the optimum. If the residual series is a white noise series, which means all residuals are random and that it cannot be used to make further improvements to the model, then the model has reached its optimum and can be used for forecast. Fig. 6 shows that all correlation coefficients fall within the range randomly, and the residual series is a white noise series.

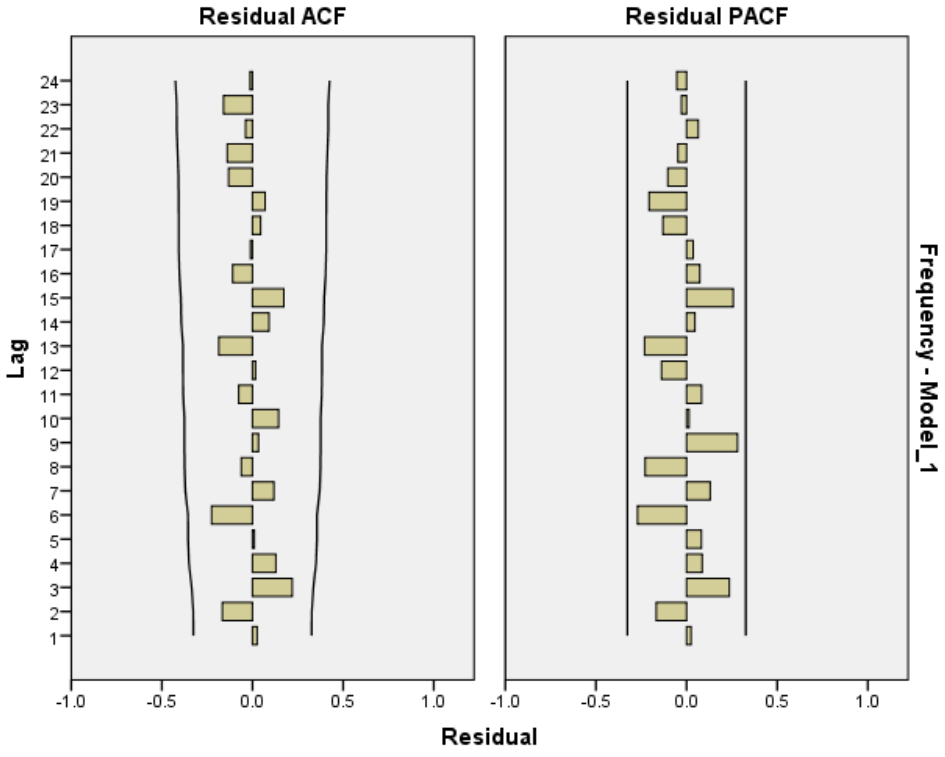

Figure-6

Figure-6: ACF & PACF of residual series in ARIMA (0,1,1) model.

Table 4 given below represents the thunderstorm forecasting data in Bangladesh from 2017 to 2022 with upper and lower control limits. Fig. 7, also represents graphically the forecasted / estimated values of frequency of thunderstorms in Bangladesh from 1980 to 2022.

Model	2017	2018	2019	2020	2021	2022
Forecast	2421.06	2448.28	2475.51	2502.76	2530.03	2557.30
UCL	3007.85	3122.83	3227.65	3325.20	3417.21	3504.82
LCL	1834.26	1773.73	1723.38	1680.32	1642.84	1609.78

Table 4: Thunderstorm forecasting in Bangladesh from 2017 to 2022

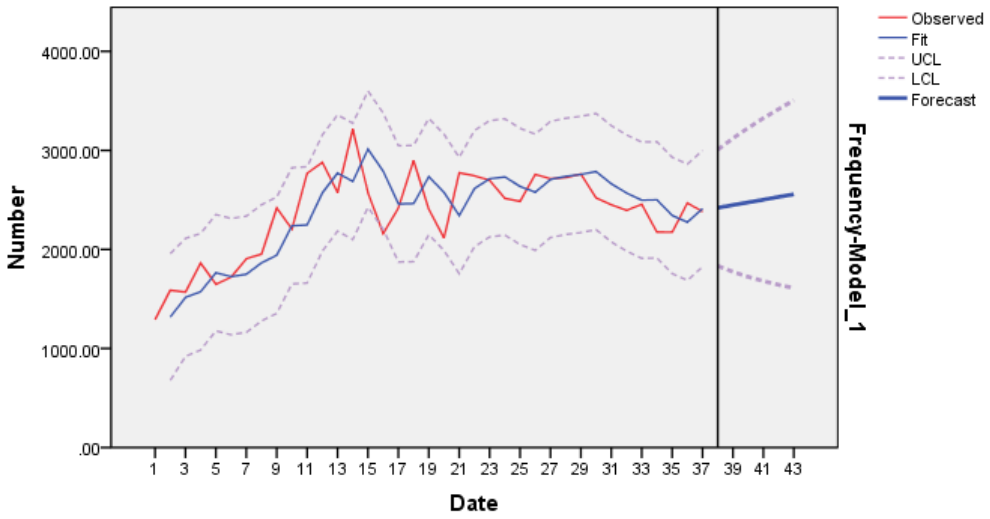

Figure-7

Figure-7: Plot of observed and predicted values of thunderstorms in Bangladesh(1980-2022).

Year	Observed Frequency of Thunderstorms	Predicted Frequency of thunderstorms	Error
1980	1291	.	.
1981	1588	1317.72	270.28
1982	1569	1516.18	52.82
1983	1863	1573.54	289.46
1984	1648	1765.10	-117.1
1985	1721	1725.43	-4.43
1986	1906	1749.71	156.29
1987	1953	1865.13	87.87
1988	2419	1941.77	477.23
1989	2207	2239.19	-32.19
1990	2769	2247.78	521.22
1991	2879	2570.17	308.83

1992	2574	2772.15	-198.15
1993	3214	2686.68	527.32
1994	2573	3012.57	-439.57
1995	2161	2790.25	-629.25
1996	2418	2460.39	-42.39
1997	2897	2463.30	433.7
1998	2406	2736.16	-330.16
1999	2116	2575.93	-459.93
2000	2774	2342.13	431.87
2001	2743	2613.99	129.01
2002	2698	2714.14	-16.14
2003	2515	2732.01	-217.01
2004	2485	2636.00	-151
2005	2758	2577.43	180.57
2006	2714	2706.87	7.13
2007	2725	2737.99	-12.99
2008	2761	2757.71	3.29
2009	2520	2786.68	-266.68
2010	2451	2662.59	-211.59
2011	2394	2569.75	-175.75
2012	2456	2497.24	-41.24
2013	2175	2501.01	-326.01
2014	2174	2343.33	-169.33
2015	2469	2274.50	194.5
2016	2380	2411.98	-31.98
2017		2421.06	
2018		2448.28	
2019		2475.51	
2020		2502.76	
2021		2530.03	
2022		2557.30	

Table 5: ARIMA (0, 1, 1) based predicted values

8. Discussions and Conclusions

This paper has mainly tried to find different ways to approach the analysis of the thunderstorms in Bangladesh. At the end of the work, we observe that a quadratic model can well define the trend of the occurrences of thunderstorms for nearly the last four decades; but an exponential model can more specifically highlight the trend of the thunderstorm pattern over the years. Two other techniques of mathematical curve fitting were tried through differential equation tools which also yield satisfactory results. Thus, it can be seen that probabilistic models and stochastic processes are giving satisfactory results when time series analysis is concerned.

The target was to highlight the years which are having some remarkable effects. Along with it, a somewhat subtle low-frequency movement buried in the larger trend and seasonal pattern was highlighted such as the years which are having lower or unexpectedly higher occurrences of thunderstorms

in some of the states. These degrees of refinement of models in this series of data can have some impact on the meteorology-oriented time series for decision making purposes[11]

Time series forecasting is a fast growing area of research and as such provides many scope for future works. One of them is the Combining Approach, i.e. to combine a number of different and dissimilar methods to improve forecast accuracy. A lot of works have been done towards this direction and various combining methods have been proposed by analysts; also there is a wide scope of finding an efficient combining model in future, where further studies on time series modelling and forecasting can be done[12], [13], [14]

An important decision-making tool can be obtained by making use of ARIMA models. The advantage of using this type of models is that, despite many problems connected with their construction and testing, we obtain information about the time series structure and mechanism of its creation. Creating models for thunderstorm is interesting and due to its complex structure and sensitivity to environmental conditions. The study shows that the ARIMA (0, 1, 1) model is not only stable but also the most suitable model to forecast the thunderstorms. Predicted results give signals to policy makers to do the necessary arrangement in advance to deal with thunderstorms.

9. Future Scope with open problems

In this article we have mainly focused the uses of the seasonal patterns of the Bangladesh landmass only. The forecasting tools used are particularly related to this kind of weather pattern. However, similar approaches can be tried out with other weather patterns as well with a bit modification. Since Bangladesh currently lacks detailed and reliable information on diurnal and spatio-temporal characteristics of thunderstorms, it is hoped that the database developed in this study will prove invaluable in efficient management of thunderstorm-related disasters. It is also possible to carry out a compare and contrast study of thunderstorms and lightning characteristics of Bangladesh. Association of thunderstorm with meteorological parameters for Bangladesh is therefore warranted. We suggest that similar types of research can be undertaken for other regions of the world where data is sparse[15] and can be studied.

Acknowledgment: This research did not receive any specific grant from funding agencies in the public, commercial, or non-for-profit sectors.

Conflict of interest: The authors declare no conflict of interest.

References

[1] R. Adhikari, R. Agarwal, *An Introductory Study on Time Series Modelling and Forecasting*, 2013 . `https://arxiv.org/ftp/arxiv/papers/1302/1302.6613.pdf`

[2] M. N. Ahasan, S. K. Debsarma, *Impact of data assimilation in simulation of thunderstorm (squall line) event over Bangladesh using WRF model, during SAARC–STORM Pilot Field Experiment 2011*, Nat. Hazards, 75(2)(Jan 1) (2015) 1009-22.

[3] J. S. Armstrong, *Findings from evidence-bases forecasting: Methods for reducing forecast error*, Int. J. Forecasting, 22 (2006) 583-598.

[4] D. Bikos , J. Finch , J. L. Case, *The environment associated with significant tornadoes in Bangladesh*, Atmos. Res., 167(Jan 1) (2016) 183-95.

[5] S. Karmakar, D. A. Quadir, M. A. Mannan, *Trends in maximum temperature and thunderstorms, their correlation and impacts on the livelihood of Bangladesh*, The Atmosphere, (2015)(Jul 5) 113-29.
[6] M. Khatun, M. A. Islam, M. A. Haque, *Studies of thunderstorms and lightning on human health, agriculture and fisheries in Mymensingh and Jamalpur district of Bangladesh*, Progressive Agric., 27(1) (2016)(Apr 29) 57-63.
[7] R. Mahmood, *Spatial and temporal analysis of a 17–year lightning climatology over Bangladesh with LIS data.*
[8] A. Mannan, N. Ahasan, S. Alam, *Study of Severe Thunderstorms over Bangladesh and Its Surrounding Areas During Pre-monsoon Season of 2013 Using WRF-ARW Model*, In High-Impact Weather Events over the SAARC Region, Springer, Cham., 2015, 3-22.
[9] M. A. Mannan, S. Karmakar, S. K. Devsarma, *Climate Feature of the Thunderstorm Days and Thunderstorm Frequency in Bangladesh*, Proc. SAARC Seminar on application of Weather and Climate forecasts in the Socio-economic Development and Disaster mitigation, 05-07 August(2007), Dhaka, Bangladesh, (2008).
[10] P. Paul, A. Imran, M. J. Islam, A. Kabir, S. Jaman, I. M. Syed, *Study of Pre-Monsoon Thunderstorms and Associated Thermodynamic Features Over Bangladesh Using WRF-ARW Model*, Dhaka Univ. J. Sci., 67(2) (2019) 155-6.
[11] A. Tyagi, D. R. Sikka, S. Goyal, M. Bhowmick, *A satellite based study of pre-monsoon thunderstorms (Nor'westers) over eastern India and their organization into mesoscale convective complexes*, Mausam, 63(1)(Jan 1) (2012) 29-54.
[12] M. J. Uddin, M. A. Samad, M. A. Mallik, *Impact of Horizontal Grid Resolutions for Thunderstorms Simulation over Bangladesh Using WRF-ARW Model*, Dhaka Univ. J. Sci., 69(1) (2021) 43-51.
[13] M. Wahiduzzaman, A. R. Islam, J. J. Luo, S. Shahid, M. Uddin, S. M. Shimul, M. A. Sattar, *Trends and Variabilities of Thunderstorm Days over Bangladesh on the ENSO and IOD Timescales*, Atmosphere, 11(11)(Nov) (2020) 1176.
[14] M. Wang, Y. Wang, X. Wang and Z. Wei, *Forecast and Analyze the Telecom Income based on ARIMA Model*, The Open Cbernetics & Syst. J., 9 (2015) 2559-2564.
[15] P. Zhang, *A neural network ensemble method with jittered training data for time series forecasting*, Inf. Sci., 177 (2007) 5329-5346.