Quantitative Prediction of Pre-Monsoon Rainfall in a Metro City of India

*Sweta Chakraborty [a,b,d] Sarbari Ghosh [a,c,d],Subrata Kumar Midya[a]

a Department of Atmospheric Sciences, University of Calcutta, Kolkata, (India)

b Jagadis Bose National Science Talent Search, Kolkata,(India)

c Department of Mathematics, Vidyasagar Metropolitan College, Kolkata, (India)

d Centre for Interdisciplinary Research and Education, Kolkata,(India)

*Corresponding author-

Sweta Chakraborty

Email-sweta.chakraborty23@gmail.com

Abstract

The present study mainly aims at the quantitative prediction of rainfall during the pre-monsoon season at Kolkata (22° 34' N, 88° 24' E), India, for the next 24 hours from the time of observation. For this purpose multiple linear regression models are developed separately for March, April, May and the Whole Pre-Monsoon Season. The total period of the study is 1974-2014 (41 years) among which data set (1974-2001) is used to select the important parameters for constructing the models and that of (2002-2014) is used for validation. The parameters are selected on the basis of stepwise linear regression, backward selection procedure and ANOVA.

It is interesting to note that the multiple linear regression models thus developed are capable of predicting the moderate range rainfall (7.5mm-35.55mm) almost accurately with the maximum error lying between ±10mm.

Though the residual plots indicate that there are no such obvious defects present in the models, the RMSE values reveal that the models for May and Whole Pre-Monsoon Season produce better results than the others.

Key words: Pre-monsoon rainfall; Multiple Linear Regression (MLR); Residual Plot; Root mean square error (RMSE); Analysis of variance (ANOVA);

1. Introduction

It is well known that rainfall during pre- monsoon months (March, April, and May) in India has both beneficial and hazardous effect on agriculture as well as urban life. From the socio-economic point of view early quantitative prediction of pre-monsoon rainfall is of utmost importance for food production plan, water resource management, disaster management of metro city etc. During pre-monsoon season occurrence of a large scale circulation is caused by the ascending motion of the air over the region of Chotanagpur plateu , Bihar, where a low pressure system prevails and the descending motion of the seasonal high pressure is present over bay of Bengal (Srinivasan *et al*. 1973, Weston *et al*.1972). Thus two different air masses, colder dry air mass coming from the north-west and moist air mass coming from the Bay of Bengal mixes. On mixing, the buoyant air mass goes up and becomes saturated with moisture in presence of a suitable triggering effect (Chatterjee *et al*. 2008). The thunderstorms formed under this condition are known as Nor'westers in West Bengal, India. The rainfalls associated with the pre-monsoon thunderstorms are very transient in nature with short duration (Saha *et al*.2014). Hence the prediction of the Indian pre-monsoon rainfall is still a highly challenging area of study.

Several methods are already employed for early prediction of amount of rainfall. These methods include Dynamical and Empirical approach. In the present study a dynamical technique is applied, since multiple linear regressions easily explain the relationship of multiple influential parameters to form a statistically significant rainfall. As the pre-monsoon rainfall is dependent on more than one parameter (independent) so the operational multiple linear regression (MLR) model has gradually become popular (Mandal *et al*.2006). In statistical analysis, regression models are frequently used for predicting the future values. Regression analysis includes non-parametric methodologies are also been applied to estimation and prediction problems (Holmström *et al.*1997). Singhrattna *et al*.(2005) presented the statistical forecasting method for summer monsoon rainfall

over Thailand. The study by Rajeevan *et al*.(2006) produced an improved result of the statistical models based on multiple linear regression and projection pursuit regression techniques for predicting the summer monsoon rainfall over India during the season (1981-2004). Zaw *et al*. (2008) worked on the multiple polynomial regression model (MPR) for monthly summer monsoon rainfall prediction over Myanmar. Selvaraj *et al*. (2011) developed multiple linear regression models to predict the monsoon rainfall by using outgoing long wave radiations, global temperatures and sunspot number over Tamilnadu.

2. Objective of the study

The synoptic systems in the pre-monsoon season are very diffuse and difficult in nature, hence predicting the amount of rainfall in this season is really challenging. Adequate objective prediction of amount of pre-monsoon rainfall may lead to more efficient planning to take preventative measures against the hazard caused by water logging in the urban metropolis e.g. Kolkata, especially in the relatively low area . If the amount of rainfall may be predicted earlier, the water available from the pre-monsoon rainfall can be utilized through proper planning for better agricultural, water resource management and other activity plans. The main objective of the study is to predict the amount of pre-monsoon rainfall within the range (7.6 mm-35.5 mm) which occur most often (Fig 1) in Kolkata (22° 34' N, 88°24' E), the capital of West Bengal. It has the third largest agglomeration in India by population (14,035,959) (wiki.org).

3. Data

The upper air RS/RW (Radio Sonde/ Radio Wind) observations1200 UTC at Kolkata (22° 34' N, 88° 24' E) during pre-monsoon season (March–May) of 1974–2014 are used in computation of the theoretically important parameters some of them are selected as predictors for the occurrence of pre-monsoon rainfall in Kolkata. The upper air sounding data are collected from Department of Atmospheric Science, Wyoming University

(http://weather.uwyo.edu).The daily rainfall data of Dumdum during the same season are collected from the Regional Meteorological Centre, Indian Meteorological Department, Alipore. This study is confined to moderate range rainfall (7.6 mm-35.5 mm).

4. Methodology:

- During the pre-monsoon season the rain is usually considered as thunderstorm rain. The rainfall recorded by Indian Meteorological Department, Alipore on a particular day represents the rainfall from 8 a.m. of the previous day to 8 a.m. on that day.
- The numerical values of physically important parameters as established by previous researchers for pre-monsoon rainfall are collected for 41 years considered for moderate range rainfall (7.6mm-35.5mm) which occur most often (fig 1).
- On the basis of R^2 values the different polynomial forms (linear, quadratic, polynomial degree 3) of the different parameters are selected respectively for March, April, May and Whole Pre-Monsoon Season to develop the primary multiple regression models. Here only the table for May is presented as an example (table 1). The other parameters have been selected in the similar way. Among all the parameters only seven [surface pressure, surface temperature, temperature at 850 hpa level , surface relative humidity, relative humidity at 850 hpa level, wind speed at 850 hpa level and precipitable water content for entire sounding] are selected as potential predictors for primary models.
- The standard technique, multiple linear regressions is used for the required prediction with the reduced number of parameters. To reduce the number of parameters backward selection procedure and ANOVA are applied. The multiple linear regression models are developed separately for March, April, May and the Whole Pre-Monsoon Season. The backward selection procedure is continued until and unless all the

coefficients associated with the parameters are 5% significant. Here only the ANOVA table of final model of May is given (table 2) as an example.

- In the present work the training data set used for developing the models cover the season of 28 years 1974-2001, and test data set used for validation covers 13 years from 2002-2014.
- The line plots of cross validation and validation of March, April, May and Whole Pre-Monsoon Season are drawn though only one plot is presented here (fig 2 a,b).
- Residuals are plotted for March, April, May and Whole Pre-Monsoon Season with number of observations along X axis and corresponding residuals along Y axis to detect if there is any error present in the models. The corresponding fig of May is presented only fig 3(a,b).
- RMSE (root-mean square error) values for March, April, May and Whole Pre-Monsoon Season for cross validation and validation are computed to check the validity of the models (table 3).
- To avoid multicollinearity the better models are restudied with their standardized forms (table 4).

5. Results and discussion:

5.1. For Non-standardized multiple linear regression model:

Depending on the R^2 ($R^2 \geq 0.03$) values the following parameters are chosen to develop the initial multiple regression models.

For March the selected parameters are: surface pressure (X_1), surface temperature (X_2), temperature at 850 hpa level (X_3) , relative humidity at 850 hpa level (X_5), wind speed at 850 hpa level (X_6) and precipitable water content for entire sounding (X_7).

For April the selected parameters are: surface pressure (X_1), surface temperature (X_2), temperature at 850 hpa level (X_3) and wind speed at 850 hpa level (X_6).

For May the selected parameters are: surface pressure (X_1), surface temperature (X_2), relative humidity at 850 hpa level(X_5), wind speed at 850 hpa level (X_6) and precipitable water content for entire sounding (X_7).

For Whole Pre-Monsoon Season the selected parameters are: surface pressure (X_1) and wind speed at 850 hpa level (X_6).

Here only the table for selection of the parameters of May is given as an example in the table 1. The selected parameters are highlighted.

The stepwise regression model is applied to develop the final models for March, April, May and the Whole Pre-Monsoon Season, which is based on back ward selection procedure in which one parameter is eliminated from the original model depending upon p- values (The p-value is the level of marginal significance within a statistical hypothesis test representing the probability of the occurrence of a given event), where $p \leq 0.05$ indicates the coefficients associated with the parameters are 5% significant. The procedure is continued until and unless all the coefficients are 5% significant (table 2).The corresponding parameters are used to develop the final model.

The plots of cross validation and validation of March, April, May and the Whole Pre-Monsoon season represent almost accurate prediction for the moderate range pre-monsoon rainfall over the area of the study [here only fig 2a,2b representing the cross validation and validation for May only are given].

Though the corresponding residual plots for March, April, May and Whole Pre-Monsoon Season give the assurance that there is no such obvious error present in the models since the residuals are randomly distributed and lies within a horizontal band, yet it is noted that the best result is obtained for the month of May and for the Whole Pre-Monsoon Season. Results of May has been presented in fig (3a,3b).

For **March,** RMSE of cross validation << RMSE of validation, which implies that the model is **over fitted** though there is no obvious defects according to the residual plot (Table 3).

For **April**, RMSE of cross validation < RMSE of validation which implies that the model is **slightly over fitted**, though the residual plot indicates no such obvious defect in the model.

5.2 . **Results and Discussion with the standardized parameters**

As the above best fitted models for May and for the Whole Pre-Monsoon Season involve the square and cubic powers of the same parameters, so to avoid multicollinearity the models are restudied with their standardized form. It is to be noted that in the final model for May and Whole Pre-Monsoon Season the parameters involved in standardized and non-standardized versions of the model remain same as expected. But the numerical values of the associated coefficients slightly differ (Table 4).

For May, RMSE of cross validation ≈ RMSE of validation which implies that the model is a **good fit** (Table 3). The coefficients in the model with standardized parameters indicate that during May the parameter, surface pressure(X_1) and precipitable water content(X_7) plays the relatively important role in predicting the amount of rainfall of moderate range in Kolkata, India.

For the Whole Pre-Monsoon Season, since the RMSE for cross validation = RMSE for validation, which depicts that the model is **best fitted** among the other models**.** The coefficients in the models with standardized parameters reveal that, surface pressure(X_1) and wind speed at 850 hpa level (m/s)(X_6), are relatively important than the other parameters for predicting the amount of rainfall during the Whole Pre-Monsoon Season (Table 4).

It is noteworthy that not only the tables 5.1 and 5.2 representing the actual and predicted pre-monsoon rainfall of May but also tables for the months of March, April and Whole Pre-monsoon season (not presented here) indicate that the magnitudes of the residuals lie between ±10mm.

6. Conclusion

The inhabitants of the urban metro city, Kolkata, are mainly engaged in non-agricultural jobs; not only that many people regularly come to Kolkata from outskirts to earn their livelihood. Due to pre-monsoon rain many roads in Kolkata get water logged. Sometimes the overhead wires of railway track get disconnected. Thus people often have to face trouble in this season especially in the evening on their way back home. The early prediction of the amount of pre-monsoon rainfall may help the common people as well as the State Government to take precautionary measures.

It is interesting to note that the present work reveals that the amount of the pre-monsoon rainfall of moderate range in Kolkata (India) may be predicted based on suitably developed multiple linear regression models, though those for May and Whole Pre-Monsoon Season produce better results than the other models.

It is found that during May, surface pressure(X_1) and precipitable water content(X_7) and for the Whole Pre-Monsoon Season, surface pressure(X_1) and wind speed at 850 hpa level (m/s)(X_6) play the relatively important role in predicting the amount of moderate range rainfall in Kolkata, India.

More parameters are yet to be explored to improve the models of March and April. But it may be concluded that similar multiple regression models may be developed for other metro cities too for operational purpose, provided a sufficient number of data is available.

Acknowledgements

Thanks to the Regional Meteorological Centre, India Meteorological Department, Alipore, for supplying the necessary rainfall data.Our sincere gratitude to Late Prof. Utpal Kumar De, Ex-emeritus Prof. of School of Environmental Studies, Jadavpur University, without whose active participation and encouragement this study could not have been possible.

References

Chatterjee P, Pradhan D and De U K, *Indian Journal of Radio & Space Physics*, 37 (2008) 419.

Helming K, Roth Ch H, Wolf R, and Diestel H, Soil Tech, 6 (1993) 273.

Holmström, L., Koistinen, P., Laaksonen, J. and Oja, E., Neural and Statistical Classifiers: Taxonomy and Two Case Studies, Jan. 1997, IEEE Trans. Neural Networks, vol. 8, No. 8, pp 5-15.

Mandal V and De U K, Weather and Forecasting, 22 (2006) 428.

Rajeevan M, Pai D S, Kumar R Anil, and Lal B, *Clim Dyn*.,28 (2006) 813.

Saha U, Maitra A, Midya S K and Das G K, *Atmospheric Research*, 138 (2014) 240.

Selvaraj R S and Aditya R, *Universal Journal Of Environmental Research And Technology*, 1(4), (2011) 557.

Singhrattna N, Rajagopalan B, Clark M and Kumar K K, *Int. J. Climatol.*, 25 (2005) 649.

Srinivasan V K, Ramamuthy Y R, and Nene, *India Meteorological Department, Forecasting Manual Part III.* (1973).

Weston K J, *Quart. J. Roy. Meteorol. Soc*, 98 (1972) 519.

Zaw W T and Naing T T, *International Journal of Computer and Information Engineering*, 2 (2008) 3418.

Tables

Dependent Parameter	Independent Parameter	Value of R^2 (Linear)	Value of R^2 (polynomial degree 2)	Value of R^2 (polynomial degree 3)	Value of R^2 (Exponential)	Value of R^2 (Log)
Y (Rainfall of May) [*Indicates the selected parameter]	Surface Pressure (hpa) (X_1)	0.023	***0.082**	0.082	0.019	0.027
	Surface Temperature (X_2)(^{0}C)	0	0.042	***0.046**	0	0
	850 hpa level Temperature (X_3)(^{0}C)	0.001	0.004	0.004	0	0.002
	Surface Relative Humidity (%)(X_4)	0.011	0.013	0.013	0.005	0.008
	Relative Humidity (850 hpa level)(%) (X_5)	0.001	0.011	***0.045**	0.001	0.002
	Wind Speed (850 hpa level) in kt (X_6)	0.005	0.005	***0.031**	0.001	0.003
	Precipitable water content for entire sounding(mm)(X_7)	0	0.004	***0.046**	0	0

Table 1. Selection of parameters for Month of May on the basis of coefficient of determination ($R^2 \geq 0.03$)

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.414431							
R Square	0.171753							
Adjusted R Square	0.115024							
Standard Error	7.207268							
Observations	79							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	5	786.3393	157.2679	3.027601	0.015405			
Residual	73	3791.964	51.94472					
Total	78	4578.304						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	136105.9	53896.33	2.525329	0.01373	28690.71	243521.2	28690.71	243521.2
X Parameter 1	-271.846	107.8597	-2.52037	0.013909	-486.81	-56.882	-486.81	-56.882
X Parameter 2	0.135796	0.053962	2.516499	0.01405	0.028249	0.243344	0.028249	0.243344
X Parameter 3	-2.30028	1.093478	-2.10364	0.038856	-4.47958	-0.12098	-4.47958	-0.12098
X Parameter 4	0.03974	0.018675	2.127934	0.036719	0.00252	0.07696	0.00252	0.07696
X Parameter 5	-0.0002	9.75E-05	-2.09259	0.039862	-0.0004	-9.7E-06	-0.0004	-9.7E-06

Final Model: $Y=136105.9-271.846X_1+0.135796X_1^2-2.30028X_7+0.03974X_7^2-0.0002X_7^3$

Table 2. ANOVA of final model of May

Month	RMSE values	
	Cross validation of training data set(1974-2001)	Validation of test data set(2002-2014)
March	6.4	9.6
April	4.7	6
May(Both standardized and non-standardized)	5.8	6
Whole Pre-Monsoon Season (Both standardized and non-standardized)	6	6

Table 3. RMSE values for March, April, May and Whole Pre-Monsoon Season

Sl no	Time	Non-standardized final Model	Standardized final Model
1	March	$Y=-538.5+0.000549X_1^2$	NA
2	April	$Y=-315573+629.9196X_1-0.31433X_1^2$	NA
3	May	$Y=136105.9-271.846X_1+0.135796X_1^2-2.30028X_7+0.03974X_7^2-0.0002X_7^3$	$Y=15.15794-0.8760X_1+1.332445 X_1^2+1.718509X_7+1.606459X_7^2-0.4501 X_7^3$
4	Whole Pre-Monsoon Season	$Y=95683.77-190.744X_1+0.095075X_1^2+0.122108X_6^2-0.00904X_6^3$	$Y=15.92624-0.61588X_1+1.385052 X_1^2+0.008833 X_6^2-0.11418 X_6^3$

[X_1=Surface Pressure (hpa), X_2=Surface temperature (^{0}C), X_3= Temperature at 850 hpa level(^{0}C),X_4= Surface Relative Humidity (%), X_5=Relative humidity at 850 hpa level(%),X_6=Wind speed at 850 hpa level (m/s)

Table 4. Table of standardized and standardized final model equations for March, April, May and Whole Pre-Monsoon Season.

Sl no	***Actual Rf***	***Predicted Rf***	***Absolute value of the difference between predicted and actual rf***	***Sl no***	***Actual Rf***	***Predicted Rf***	***Absolute value of the difference between predicted and actual rf***
1	8	14	6	**36**	13	17	4
2	15	19	4	**37**	17	20	3
3	13	16	3	**38**	10	19	9
4	15	19	4	**39**	11	16	5
5	25	17	8	**40**	15	17	2
6	34	25	9	**41**	23	18	5
7	16	17	1	**42**	10	15	5
8	17	15	2	**43**	17	16	1
9	10	17	7	**44**	14	18	4
10	8	17	9	**45**	10	16	6
11	16	17	1	**46**	33	26	7
12	19	15	4	**47**	17	15	2
13	11	16	5	**48**	24	15	9
14	26	31	5	**49**	10	15	5
15	17	15	2	**50**	13	17	4
16	8	16	8	**51**	19	17	2
17	23	15	8	**52**	12	16	4
18	14	15	1	**53**	8	16	8
19	10	16	6	**54**	12	14	2
20	24	17	7	**55**	11	18	7
21	13	21	8	**56**	11	19	8
22	23	18	5	**57**	15	15	0
23	10	19	9	**58**	11	17	6
24	15	18	3	**59**	17	15	2
25	10	15	5	**60**	22	19	3
26	22	15	7	**61**	28	26	2
27	15	25	10	**62**	30	28	2

28	25	15	10	**63**	16	16	0
29	18	19	1	**64**	10	16	6
30	26	17	9	**65**	15	17	2
31	18	20	2	**66**	13	17	4
32	11	15	4	**67**	9	16	7
33	10	18	8	**68**	29	20	9
34	12	21	9	**69**	25	21	4
35	20	15	5	**70**	9	19	10

Table 5.1.Training Data Set of May

Sl no	***Actual Rf***	***Predicted Rf***	***Absolute value of the difference between predicted and actual rf***	***Sl no***	***Actual Rf***	***Predicted Rf***	***Absolute value of the difference between predicted and actual rf***
1	21	18	3	**11**	19	15	4
2	21	19	2	**12**	9	16	7
3	11	18	7	**13**	17	18	1
4	13	15	2	**14**	23	17	6
5	26	16	10	**15**	19	26	7
6	20	16	4	**16**	21	18	3
7	19	17	2	**17**	11	16	5
8	13	16	3	**18**	9	16	7
9	17	16	1	**19**	11	20	9
10	22	15	7	**20**	27	18	9

Table 5.2.Validation Data Set of May

Figures

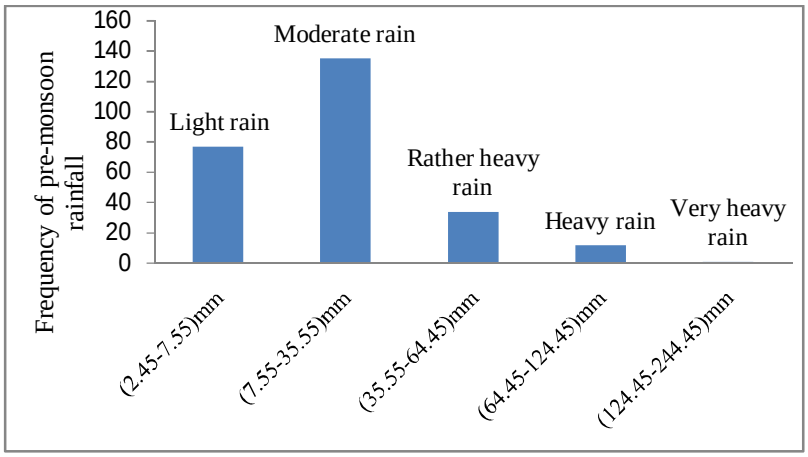

Fig 1. Bar diagram of distribution of pre-monsoon rainfall over Kolkata during the study period (1974-2014).

a) b)

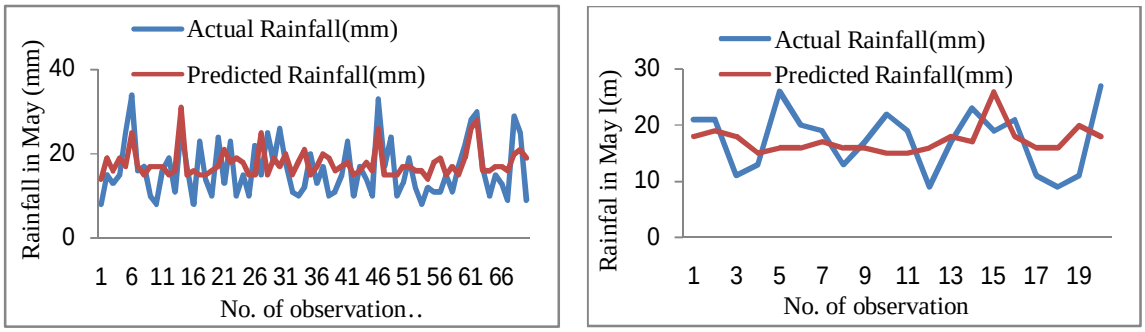

Fig 2. Line plots of cross validation and validation of May a) cross validation of training data set (1974-2001) of May, b) validation of test data set (2002-2014) of May.

a) b)

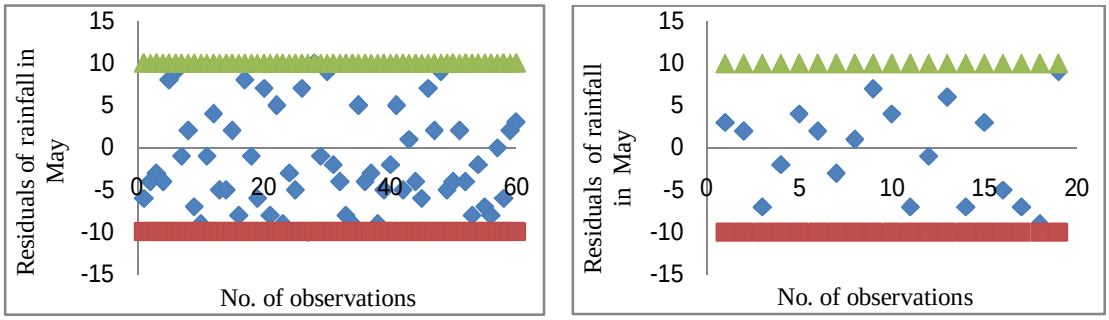

Fig 3. Residual plots of cross validation and validation of May a) Residual plots of cross validation of training data set (1974-2001) of May b) Residual plots of validation of test data set (2002-2014) of May.