# A novel approach to predict the onset of South West Monsoon in Kerala

T.K. Ajitha[1*] and P. Anil Kumar [2]

[1]*College of Horticulture, KAU P.O., Kerala Agricultural University, Thrissur 680656, Kerala, India.*; [2]*Department of statistics, Farook College, Calicut. Kerala, India.*

Received 29 May 2017; received in revised form 15 June 2017; accepted 30 June 2017

**Abstract**

A univariate statistical approach for prediction of onset of South West monsoon along with a model based on Pre Monsoon Rainfall Peak (PMRP) has been demonstrated in this study using the dates of onset for 147 years from 1870 – 2016 published by IMD and daily climatic variables viz; rainfall, maximum temperature, minimum temperature, relative humidity, sunshine hours and wind velocity for a period from 1984 – 2016 for the region Vellanikkara (bounded by $10^0$ 31'; $76^0$13'), Thrissur. The average date of onset was found to be June1with a S.D of 7 days. The occurrence of a PMRP was confirmed about 8 pentads before the onset of monsoon. A significant correlation coefficient of 0.59 between the dates of PMRP and onset dates has resulted in the formulation of a regression equation to predict onset dates with adequate degree of precision. Long term predictions with minimum possible errors have been realised using an emerging univariate statistical model. A non contiguous regression model by means of weighted least squares with heteroscedasticity correction has been applied to predict onset dates using its own significant auto correlated lags.

**Keywords**: Kerala, Onset, PMRP, Prediction.

## Introduction

The onset of South-West (SW) monsoon over Kerala signals the arrival of monsoon over the Indian subcontinent and represents beginning of rainy season over the region. A good monsoon always results in high crop productivity. Hence, in each year the farmers are eagerly waiting to see the onset of SW monsoon (contributing about 80% of the annual mean rainfall), which commences in Kerala in early June and progresses to other parts of the country. Though, the onset of monsoon in Kerala is normally considered as on 1st June, it is highly uncertain.In some years it commences by the first week of May and by last half of June in some other years.This brings up the importance of monsoon prediction. According to Ananthakrishnan and Soman, (1988), May 30th was found to be the onset date for South Kerala and 1st june for North Kerala with a S.D of about 9 days in both cases.

Though several attempts were made, predictions fail due to the large air-sea interactions and the scientists were forced to modify their work and try for better alternative methods and models. Sooraj, (2004) was of the opinion that a consistent dramatic reversal in wind direction over the western Arabian Sea would occur about three weeks prior to the onset of monsoon. Because of the complexity of the monsoon of South Asia it is very difficult to accurately predict the quantity, timing and geographic distribution of the precipitation which determine the water availability for any given year.

*Author for correspondences: Phone: 9496168264, Email: ajitha.tk@kau.in

Numerous efforts have been done by several scientists to improve the existing series of forecasting models. In this study attempts have been made to predict the onset of SW monsoon by using a meteorological as well as statistical approach. Comparing the relative efficiency of the models, a potential model has been identified.

## Materials and Methods

The dates of onset of SW monsoon for 147 years for the period from 1870- 2016 published by IMD were used for the study. Daily data on the meteorological variables viz., rainfall, maximum temperature, minimum temperature, relative humidity, sunshine hours and wind velocity for 33 years for the period from 1984 to 2016 recorded at the meteorological observatories in the College of Horticulture, Vellanikkara (bounded by $10^0$ 31'; $76^0$13'), under Kerala Agricultural University were also used. As a meteorological approach, the pre-monsoon rainfall peak (PMRP) for each year was identified to construct prediction equations for onset of SW monsoon by regressing PMRP on onset dates. In the statistical approach of prediction, the dates of onset were quantified to an index for the ease of computation. The index 518 denotes June 1.It is an arbitrarily assumed value attached to June 1. In Ms Excel dates can be stored as serial values so that they can be subjected to algebraic operations. Actually 518 corresponds to June 1, 1901. To develop suitable models by regressing onset values on its own past values, the weighted least square method with heteroscedasticity correction has been used which will account for the auto correlations present in the univariate series of onset data.

## Results and Discussion

To characterise the internal structure of the onset data the summary statistics were computed. The average date of onset of SW monsoon was found to be June 1st (518) with a S.D of 7.08 ≈ 7 days. The results showed that the onset dates ranged between as early as 11$^{th}$ May (497) in 1918 to as late as June 18$^{th}$ (535) in 1972.

*Table 1.* Frequency distribution of onset dates according to ± 1S.D limit

| Date of Onset | Frequency | Percentage |
|---|---|---|
| Before 25$^{th}$ May | 18 | 12 |
| 25$^{th}$ May – 8$^{th}$ June | 108 | 74 |
| After 8$^{th}$ June | 21 | 14 |
| Total | 147 | 100 |

Out of the 147 data points of onset dates, 74 percentage were in the range of 518 ± 1S.D i.e., between 25$^{th}$ May – 8$^{th}$ June. Twelve percentage (12%) of the cases were before 25$^{th}$ May and 14

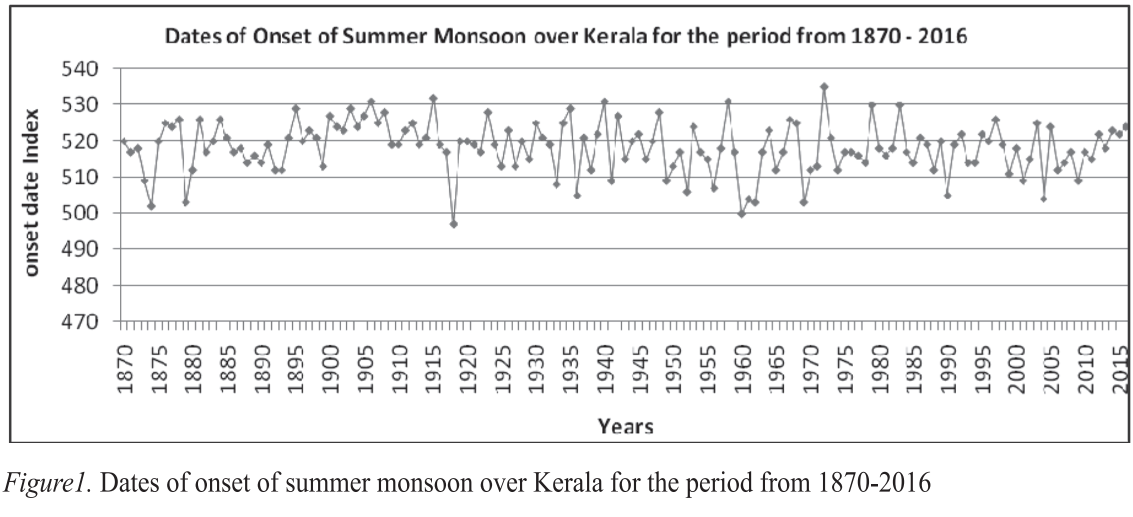


*Figure1.* Dates of onset of summer monsoon over Kerala for the period from 1870-2016

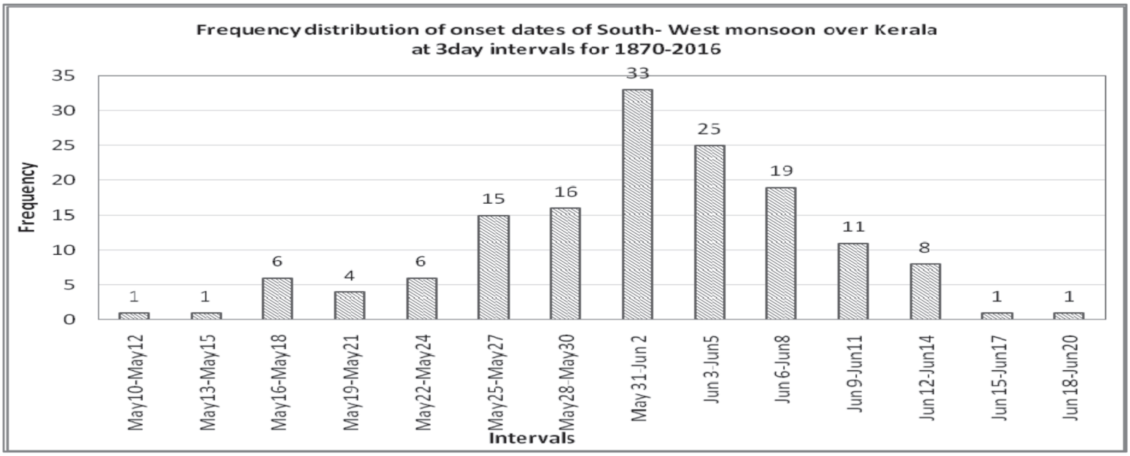


*Figure 2.* Frequency distribution of onset dates of South-West monsoon over Kerala at 3 day intervals for 1870-2016

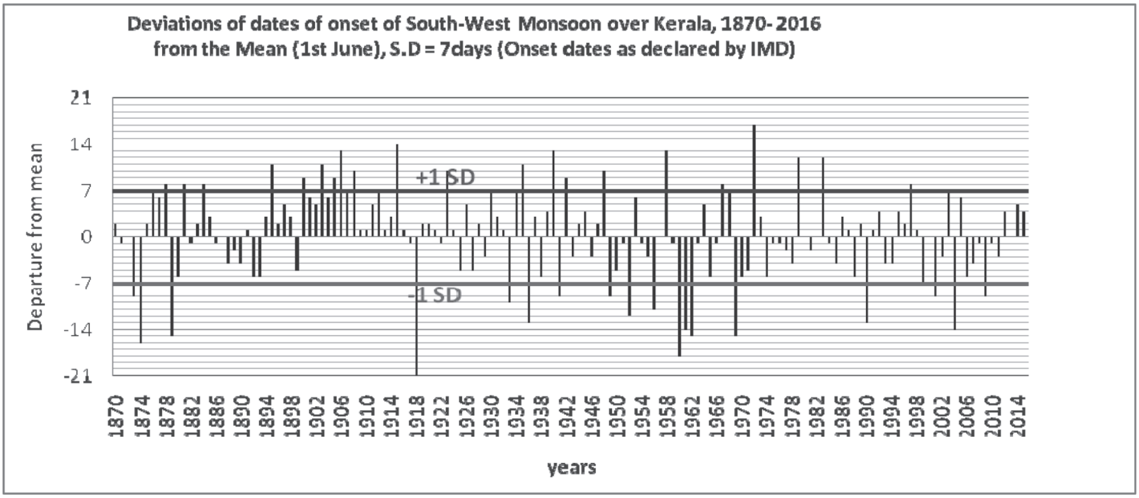


*Figure 3.* Deviations of dates of onset of South-West monsoon over Kerala, 1870-2016 from the mean (1st June), S.D= 7 days (onset dates as declared by IMD)

percentage of them were beyond 8th June. Goswami and Gouda, (2010) were of the opinion that eventhough the standard deviation in date of onset was only 7 days for the past hundred years, almost 50 percentage of the cases showed large (> 1 S.D) deviations . Forecasting of date of monsoon onset was rarely attempted because of the poor skill of most of the General circular models in the long range prediction of daily rainfall.

According to several forecasting studies, pentad averages (five day mean) seemed to be a better representation of the distribution of rainfall. Hence in the meteorological approach of prediction of onset date, the pentad average of the daily climatic variables viz; rainfall, maximum temperature, minimum temperature, relative humidity, sunshine hours and wind velocity were computed from March 2nd to onset date for 30 years from 1984 - 2013. The correlation coefficients of pentad averages for each climatic variable were computed with the pentad average rainfall during the onset period. A significant correlation coefficient of 0.46 was obtained between the quantity of rainfall at 8

*Table 2.* Descriptive statistics of mean pentad rainfall from 1 to 15 pentads before onset at Vellanikkara, Thrissur district

| X pentad before | Mean | S.D | N | Correlation with mean rainfall at onset | Significance |
|---|---|---|---|---|---|
| -15 | 0.68 | 2.76 | 30 | -0.22 | 0.25 |
| -14 | 0.26 | 0.90 | 30 | -0.07 | 0.71 |
| -13 | 0.50 | 1.08 | 30 | -0.23 | 0.22 |
| -12 | 2.29 | 4.00 | 30 | 0.03 | 0.89 |
| -11 | 0.99 | 2.10 | 30 | -0.27 | 0.16 |
| -10 | 1.54 | 2.79 | 30 | -0.07 | 0.72 |
| -9 | 3.17 | 4.46 | 30 | -0.01 | 0.96 |
| -8 | 4.05 | 7.10 | 30 | 0.46** | 0.01 |
| -7 | 3.13 | 5.37 | 30 | -0.13 | 0.50 |
| -6 | 3.5 | 8.87 | 30 | -0.24 | 0.20 |
| -5 | 4.04 | 5.86 | 30 | -0.02 | 0.92 |
| -4 | 2.28 | 4.2 | 30 | -0.12 | 0.54 |
| -3 | 4.74 | 7.07 | 30 | -0.24 | 0.21 |
| -2 | 6.39 | 10.07 | 30 | 0.11 | 0.58 |
| -1 | 3.77 | 3.44 | 30 | 0.37* | 0.05 |
| 0 | 21.28 | 19.36 | 30 | 1 | |

pentads before onset and the quantity of rainfall at the onset pentad.The correlation coefficient was not significant for other climatic variables except in the just previous pentad before onset. Joseph et al. (2006) have found that, eight pentads before the monsoon onset over Kerala, a spatially large area of deep convection would form near the equator, South of Bay of Bengal, which would move to South East Asia marking the onset of South China Sea monsoon for many years. Eight pentads before monsoon onset of Kerala (MOK) a warm pool was located over central Bay of Bengal and an area of active convection would form to its South near the equator in the region of large sea surface temperature gradient.

The composite values over the years 1984-2013 of the rainfall data were computed using the pentad average values in such a way that the MOK would coincide with the zero pentad. The intensity of rainfall was observed at -15 pentads to + 2 pentads with respect to the composite values. Fig. 4 shows the existence of a PMRP around 8 pentads before the onset of monsoon. For every year during the period from April 1 to May 10, the mid day of the pentad with rainfall peak was identified and recognised as PMRP. The occurrence of PMRP was

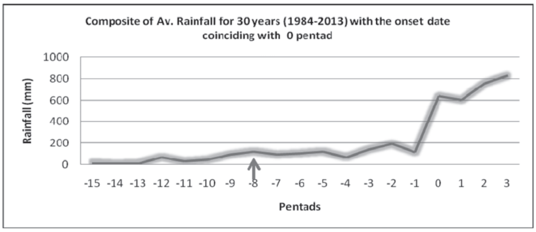


*Figure 4.* Composite of average rainfall for 30 years (1984-2013) with the onset date coinciding with 0 pentad

marked as on day 1, day 2 etc. if it has realised on April 1, April 2 and so on. A scatter diagram was drawn taking PMRP (x days) and the actual onset dates (y days). A correlation coefficient of 0.59 was obtained between the dates of PMRP and onset dates which was highly significant. The regression of y on x was as follows which can be used for the prediction of onset dates for the future.

$$Y = 48.54 + \overset{**}{0.52} X$$ with a S.E = 0.14 for the regression coefficient

The actual and estimated monsoon onset dates obtained using the regression of onset dates on PMRP are shown in Table 3. The regression model was validated by computing the predicted onset dates for 4 independent years, 2014 – 2017 which

*Table 3.* Yearly distribution of PMRP, MOK, Estimated MOK and Residuals for 1984 - 2016

| Year | PMRP | MOK (A) | Estimated MOK (B) | Error (A-B) | Year (A) | PMRP | MOK | Estimated MOK (B) | Error (A-B) |
|---|---|---|---|---|---|---|---|---|---|
| 1984 | 23 | 61 | 61 | 0 | 2001 | 23 | 53 | 61 | -8 |
| 1985 | 13 | 58 | 55 | 3 | 2002 | 23 | 59 | 61 | -2 |
| 1986 | 33 | 65 | 66 | -1 | 2003 | 23 | 69 | 61 | 8 |
| 1987 | 23 | 63 | 61 | 2 | 2004 | 18 | 48 | 58 | -10 |
| 1988 | 18 | 56 | 58 | -2 | 2005 | 23 | 68 | 61 | 7 |
| 1989 | 23 | 64 | 61 | 3 | 2006 | 13 | 56 | 55 | 1 |
| 1990 | 18 | 49 | 58 | -9 | 2007 | 23 | 58 | 61 | -3 |
| 1991 | 23 | 63 | 61 | 2 | 2008 | 28 | 61 | 63 | -2 |
| 1992 | 23 | 66 | 61 | 5 | 2009 | 18 | 53 | 58 | -5 |
| 1993 | 18 | 58 | 58 | 0 | 2010 | 28 | 61 | 63 | -2 |
| 1994 | 13 | 58 | 55 | 3 | 2011 | 28 | 59 | 63 | -4 |
| 1995 | 33 | 66 | 66 | 0 | 2012 | 38 | 66 | 68 | -2 |
| 1996 | 23 | 64 | 61 | 3 | 2013 | 18 | 62 | 58 | 4 |
| 1997 | 33 | 70 | 66 | 4 | 2014* | 28 | 67 | 63 | 4 |
| 1998 | 28 | 63 | 63 | 0 | 2015* | 33 | 66 | 66 | 0 |
| 1999 | 18 | 55 | 58 | -3 | 2016* | 23 | 68 | 61 | 7 |
| 2000 | 18 | 62 | 58 | 4 | 2017* | 14 | 60 | 56 | 4 |

*Validation of the regression model

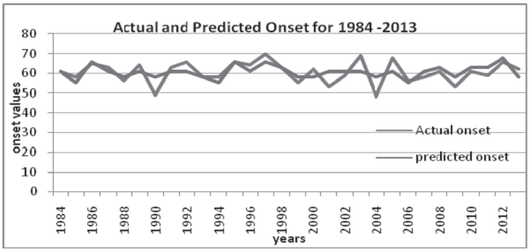


*Figure 5.* Actual and predicted onset for 1984-2013

were not included for the model fitting. In 2015 the error of prediction has turned out to be zero. The actual and predicted onset dates are graphically shown in Fig. 5. It could be observed that the predictability of the model is adequate except for some odd years where there was too early or too late onset events. The insignicant Z value in the Runs test in Table (4) shows that the errors of forecasts were randomly distributed. So this model can be recommended for predicting the arrival of SW monsoon in Kerala, especially in the particular locality of study as the errors were negligibly small for the concerned region.

*Table 4.* Runs test for the residuals

| | |
|---|---|
| Test value | - 2 |
| Cases < Test value | 7 |
| Cases >= Test value | 23 |
| Total cases | 30 |
| Number of Runs | 15 |
| Z | 1.46 |
| Asymp. Sig. (2 tailed) | 0.15 |

Kumar, (2004) has stressed the potential of satellite data to predict the onset of SW monsoon and used the Global precipitation and In Situ gauge data for 23 years from 1979 – 2001 and PMRP was identified around six pentads before the monsoon onset. In their study the correlation coefficient between the dates of PMRP and onset dates was 0.64. Even though the correlation coefficient between dates of PMRP and onset dates of SW monsoon was 0.59 in the present study, the PMRP was realised 8 pentads in advance of the onset as against the findings of the above author. The results of this study reveals that the surface data is also equally capable as satellite data for predicting onset dates as against the existing belief.

In the meteorological approach already discussed, only short term predictions were tried. But a long term prediction is tried in the statistical approach. Time series forecasting is made use of to predict future onset dates based on its own previously observed values. If the current level of the dependent variable is heavily determined by its past levels or if the error terms are correlated across time it will generate autocorrelation in the Time series. In other words the present outcome of the model is greatly affected by the past errors. In such situation the assumption of homoscedasticity needed for BLUE (Best Linear Unbiased Estimator) is violated resulting in biased standard errors. A statistically valid estimation of the effect of the independent variables on the dependent variable cannot be made if the standard errors are biased.

Lagged dependent variables can be included as regressors in the model to make the results reliable by reducing the occurrence of autocorrelation arising from model misspecification. Here the correlation between the lagged dependent variable and the current disturbance are to be considered to estimate the parameters. This is a violation from the critical assumption of the classical linear regression model and due to inconsistency, the ordinary least square estimator become unacceptable. Hence the weighted least square method with heteroscedasticity correction has been made use of to predict the onset of SW monsoon. The dates of onset of SW monsoon for the period from 1870 -2004 have been used to fit the regression model and validation of the model has been done using the data from 2005 – 2016. Predictions for 5 years from 2017 – 2021 have been made together with the confidence interval.

As a first step, the stationarity of the data was tested using the Augmented Dickey fuller test. The results showed that the series was stationary with auto correlated lags . A regression model using 37 lags yielded an adjusted $R^2$ of 86% which was statistically significant at 1% level of significance. The other statistics based on the weighted data were as follows.

The relative goodness of fit of statistical models are measured using several Information criteria. These are measures of the trade off between the uncertainity in the model and the number of parameters in the model. These tools are mainly used to attain much explanatory power with only a few parameters. The R – squared value would always increase when additional regression parameters are added. But the increase in accuracy would decrease the parsimony of the model. Practically, R – squared often increases dramatically for the first few added regression parameters and then levels off as more parameters are added. Therefore instead of fitting a model with 37 lagged variables involving insignificant regression coefficients, a non contiguous model considering the significant lags which is both parsimonious (does not over fit the data with too many parameters) and accurately modelling the data has been developed. To capture the dynamic effects

*Table 5.* Statistics based on the weighted data using 37 lags

| | |
|---|---|
| Sum of squared residuals | 122.32 |
| S.E of regression | 1.43 |
| Log – likelihood | -149.92 |
| Akaike criterion | 375.83 |
| Schwarz criterion | 474.07 |
| Hannan - Quinn | 415.57 |
| Rho | 0.05 |
| Durbin’s h | 0.71 |

*Table 6.* Weighted Least Square Estimates used for prediction of onset of SW monsoon in Kerala

| Estimates | Coefficient | Std. Error | t-ratio | P - value |
|---|---|---|---|---|
| Constant | 452.39 | 105.33 | 4.29 | < 0.0001*** |
| b1 (Y_5) | 0.19 | 0.09 | 2.04 | 0.04** |
| b2 (Y_12) | - 0.12 | 0.09 | -1.25 | 0.21 |
| b3 (Y_19) | 0.23 | 0.10 | 2.29 | 0.02** |
| b4 (Y_22) | - 0.18 | 0.09 | -1.87 | 0.06* |

*Table 7.* Statistics based on the weighted data using 4 lags

| | |
|---|---|
| Sum of squared residuals | 588.62 |
| S.E of regression | 2.33 |
| Log -likelihood | —253.59 |
| Akaike criterion | 517.18 |
| Schwarz criterion | 530.81 |
| Hannan- Quinn | 522.71 |
| Rho | 0.03 |
| Durbin-Watson | 1.90 |

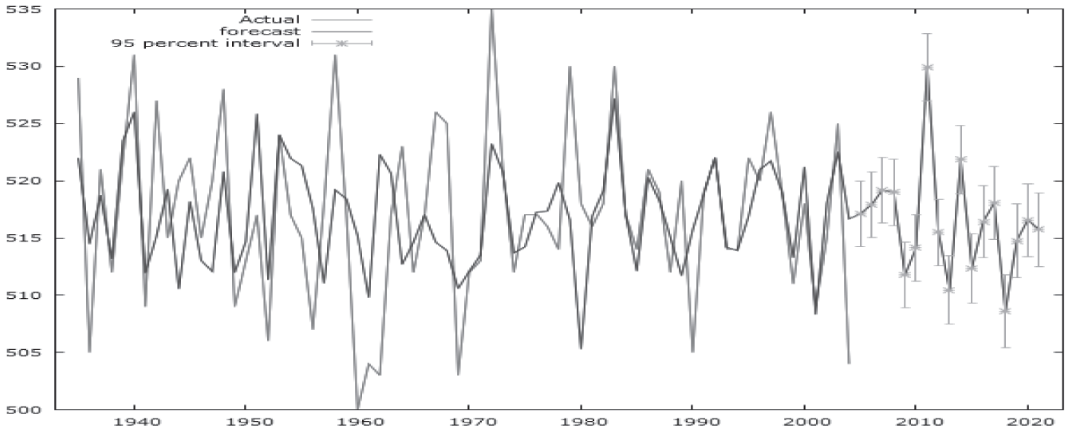


*Figure 6.* Actual and Forecasted dates of onset of SW monsoon in Kerala

*Table 8.* Runs test for residuals

| | |
|---|---|
| Test value | 0.00 |
| Cases < Test value | 47 |
| Cases >= Test value | 66 |
| Total cases | 113 |
| Number of Runs | 57 |
| Z | 0.213 |
| Asymp. Sig. (2 tailed) | 0.831 |

and to get rid of the autocorrelations present in the data, the significant auto correlated lags were first identified and a regression model with heteroscedasticity correction was fitted using the onset dates corresponding to the significant lags viz; 5, 12, 19 and 22 as the regressors. This model could predict the onset dates with minimum error when compared to the above models.

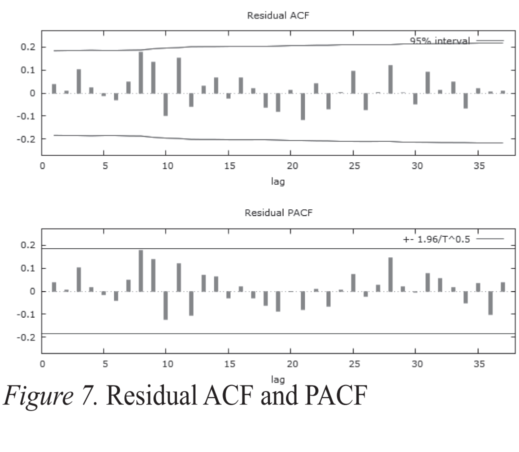


*Figure 7.* Residual ACF and PACF

The adequacy of the parsimonious model using the significant lags has been tested using the standard criterion on residuals. The errors satisfy the normality test (Fig.8) and peak of the curve corresponds to zero error. So it can be concluded that in most of the years the predictions can be done with zero error that is with absolute accuracy as illustrated by the Runs test (Table 8) which shows that the modal value of error is zero. The non significant Z value of the Runs test indicate the acceptance of the null hypothesis that the errors of prediction were randomly distributed. The ACF (Auto correlation function) and PACF (Partial autocorrelation function) plots give a picture of white noise residual terms which shows that

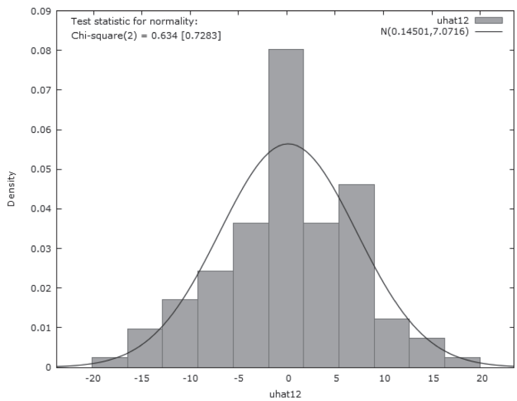


*Figure 8.* Normality Test for Forecast Errors

majority of the information content in the data has been extracted by the model and the residual series is not contaminated with any auto correlated terms. Long range forecasting of the Indian summer monsoon onset and rainfall in terms of antecedent upper air circulation have been attempted by Kung and Sharif (1982).

From Table 9, it can be noticed that in almost all the years, the forecast errors are much lower in the case of predictions using regression model with selected significant lags as regressors when compared to the other models using 37 lags as well as the model using PMRP. The mean absolute error (MAE) used to compare the accuracy of prediction models turned out to be 3.08, 6.25 and 3.75 respectively for the models using 4 lags, 37 lags and using PMRP. The existing methods of monsoon onset forecasts make use of several variables of the current year and predictions are done for the particular year only. The recommended model (using 4 lags with lowest MAE = 3.08) in this study make use of the past values of a single variable viz; onset dates and the predictions can be done for several years in the future.

The statistical analysis of onset dates of SW monsoon for 147 years from 1870 – 2016 and daily climatic variables for 33 years from 1984 – 2016 for the region Vellanikkara (bounded by $10^0$ 31'; $76^0$13'), Thrissur has resulted in the conclusion that the prime variable which can be used for prediction of onset dates is rainfall. The average date of onset was found to be June 1 with a S.D of 7 days. In 74% of the years, the onset was in between 25th May to 8th june. On an average, a pre monsoon rainfall peak was found to occur 8 pentads before onset. The correlation coefficient between dates of PMRP

*Table 9.* Validation of the regression model (out of sample forecasts from 2005 - 2016) to predict the onset of SW monsoon

| Year (1) | Actual dates of onset (2) | Predicted dates of onset using 37 lags (3) | Error in prediction (2) – (3) | Predicted dates of onset using significant lags 5,12,19,22 (4) | Error in prediction (2) – (4) |
|---|---|---|---|---|---|
| 2005 | 524 | 517 | 7 | 517 | 7 |
| 2006 | 512 | 518 | -6 | 517 | -5 |
| 2007 | 514 | 519 | -5 | 516 | -2 |
| 2008 | 517 | 519 | -2 | 519 | -2 |
| 2009 | 509 | 512 | -3 | 511 | -2 |
| 2010 | 517 | 514 | 3 | 519 | -2 |
| 2011 | 515 | 530 | -15 | 519 | -4 |
| 2012 | 522 | 515 | 7 | 519 | 3 |
| 2013 | 518 | 510 | 8 | 518 | 0 |
| 2014 | 523 | 522 | 1 | 517 | 6 |
| 2015 | 522 | 512 | 10 | 519 | 3 |
| 2016 | 524 | 516 | 8 | 523 | 1 |

*Table 10.* Forecasts of the onset of SW monsoon for the years from 2017 – 2021 using the parsimonious model (using lags 5,12,19,22)

| Year | prediction | Corresponding date of onset | S.E | 95% confidence interval |
|---|---|---|---|---|
| 2017 | 518 | 1st June | 2.39 | 513.31 – 522.82 |
| 2018 | 516 | 30th May | 2.40 | 511.63 – 521.14 |
| 2019 | 517 | 31st May | 2.38 | 512.10 – 521.60 |
| 2020 | 516 | 30th May | 2.39 | 511.18 – 520.69 |
| 2021 | 521 | 4th June | 2.40 | 515.77 – 525.27 |

*Table 11*. Forecasts of the onset of SW monsoon for the years from 2017 – 2021 using the model with 37 lags

| Year | Prediction | Corresponding date of onset | S.E | 95% confidence interval |
|---|---|---|---|---|
| 2017 | 518 | 1st June | 1.58 | 514.90— 521.22 |
| 2018 | 509 | 23rd May | 1.58 | 505.44— 511.77 |
| 2019 | 515 | 29th May | 1.60 | 511.56— 517.95 |
| 2020 | 517 | 31st May | 1.60 | 513.37— 519.77 |
| 2021 | 516 | 30th May | 1.61 | 512.52— 518.96 |

and onset dates of SW monsoon was 0.59 which is highly statistically significant. The study shows that the Agrometeorological data recorded at automatic weather stations are also equally capable like satellite data to predict the onset dates using the details of PMRP. As against this meteorological approach, long term predictions were tried in the statistical approach of prediction using weighted least square method with heteroscedasticity correction. Regression models were fitted to predict onset dates using its own significant auto correlated lags. Promising results with high level of accuracy could be realised using the statistical approach of forecasting.

## Acknowledgement

Authors gratefully acknowledge the Department of Agricultural Meteorology, College of Horticulture, Kerala Agricultural University, Vellanikkara for providing the data used for the study.

## References

Ananthakrishnan, R. and Soman, M.K. 1988. The onset of the south west monsoon over Kerala 1901 – 1980. J. Climatol., 8: 283-296.

Goswami, P. and Gouda, K.C. 2010. Evaluation of a dynamical basis for advance forecasting of the date of onset of monsoon rainfall over India. Mon. Wea. Rev., 138: 3120–3141.

Joseph, P.V., Sooraj, K.P. and Rajan, C.K. 2003. Conditions leading to monsoon onset over Kerala and the associated. Hadley cell. Mausam., 54(1): 155–164.

Kumar, M.R.R. 2004. Forecasting of onset of south west monsoon over Kerala coast using satellite data. IEEE Geosci. remote sensing letters., 1(4): 265-267.

Kung, E.C. and Sharif, T.A. 1982. Long range forecasting of the Indian summer monsoon onset and rainfall with upper air parameters and sea – surface temperature. J. Meteorol. Soc. Japan., 60: 672 – 682.

Sooraj, K.P. 2004. A Diagnostic and modelling study on the onset of summer monsoon over south Asia. Ph.D. Thesis Cochin University of Science and Technology. Kochi. 262p