Mysore J. Agric. Sci., 52 (3) : 606-612 (2018)

Bias Correction and ARIMA Model for Paddy Crop Production in Selected Districts of Karnataka

SUNIL GAYAKAWAD, G. B. MALLIKARJUNA, K. N. KRISHNAMURTHY AND A. SATHISH
Department of Agricultural Statistics, Applied Mathematics and Computer Science
College of Agriculture, UAS, GKVK, Bengaluru - 560 065
E-mail : kkmurthy01@yahoo.com

ABSTRACT

Forecasting is an essential tool in agriculture to study the area and production of various crops. Estimates of Paddy Production are taken from crop cutting experiments and through satellite. The data obtained from the crop cutting experiment estimates are always more accurate than that of the satellite estimates. Crop cutting experiments have its own limitations mainly in the coverage of area in a limited period. Remote sensing data collected covers larger area. In view of its large resolution, biasness could be observed in the data collected. To reduce this, different bias corrective methods such as Difference Method (DM) and Modified Difference Method (MDM) were used. Based on the Normalized Mean Square Error (NRMSE) value the best method for each data set is identified. The study resulted that MDM was the best method for bias correction showing the least value of NRMSE and it can be used for smoothening of the modeled data. In the presence of auto correlation, model fitting was done by ARIMA models for the crop production data of all selected districts of Karnataka. ARIMA (0,1,0) was found to be the best fit for production of Paddy crop for all the selected districts.

Keywords: Difference method, Modified difference method

ONE of the important areas in agriculture is crop yield forecasting. Their use includes monitoring of agricultural production changes, planning of agricultural interventions, development projects, development of early warning systems and preparation of macroeconomic accounts. Poor agricultural data can lead to misallocation of scarce resources and policy formulations that fail to resolve critical development problems. The advance estimates of crop production are needed much before the actual harvest of the crops for making various decisions such as pricing, distribution, export and import *etc.* However, the final estimates of crop production which are based on area through complete enumeration and yield rate through Crop Cutting Experiments are made available much after the harvest of the crop. Therefore, there is great need for developing suitable and reliable models using information from different sources like agricultural inputs, meteorological data and remote sensing data for providing the reliable and timely forecast of crop Area/Production. Accurately estimating crop yields is never easy and is even more of a challenge in the context of farming systems that are characterized by small area holder farms that produce a wide range of diverse crops. Challenges that may occur include information on land use, intercropping, non-uniform plots in a wide range of sizes, not all planted area is harvested and significant post-harvest losses.

Crop Cutting Experiments which are more precise for small areas, become invalidate at country level. Currently the agriculture department officials visit the village or tahsil where they inquire about crop acreage and expected yield. Based on these types of sampling the results are projected to acquire the acreage and yield information. This methodology, though prevalent from a long time is neither very accurate nor very scientific. It is having other limitations such as extremely tedious, time-consuming, costly, inconsistent and labor-intensive.

Alternatively, Remote sensing data has been used for forecasting purpose. It does not require close contact between the sensing organs and the external objects. It deals with remote sensing data attained through earth observation satellite. Remote sensing-based methods

have already been proven as an effective alternative for mapping crop area and forecasting crop production. The benefits of remote sensing technology include: (i) spatial coverage over a large geographic area; (ii) availability during all seasons; (iii) relatively low cost, since some optical images are freely available although radar data are usually a bit costly; (iv) efficient analysis; (v) they provide information in a timely manner and (vi) they are capable of delineating detailed spatial distributions of areas under crop cultivation. Problems that limit the current usefulness of remote sensing for developing countries include cloud coverage, the need for expensive ground truthing, the need for specialist knowledge, and the need of expensive image processing software (Reynolds *et al.,* 2000). Under this situation precise estimate will be done only by smoothing (Bias correction) the data generated for minimizing the variation. Smoothing of the data has to be done by using appropriate bias correction to the data before having the proper prediction model. Gallego (2006) indicated that crop area estimation from satellite imagery is typically calculated using the product of the resolution of an image and the area of an agricultural feature delineated with a spectral classifier. It was revealed that, Co-location inaccuracies and considerable overlap between spectral categories can induce further error. Graham *et al.* (2007) and Weiland *et al.* (2010) used delta method for the bias correction.

Material and Methods

The present study was based on the secondary data on Paddy crop production of selected districts of Karnataka *viz.,* Bellary, Davanagere and Raichur. The data over a period of 17 years (1998-2015) was collected from the Directorate of Economics and Statistics (DES), Government of Karnataka and Karnataka State Remote Sensing Application Centre (KSRSAC), Bangaluru. The data obtained from the Directorate of Economics and Statistics (DES) is an observed data which is based on Crop Cutting Experiment. Remote estimates which are obtained from the Karnataka State Remote Sensing Application Center is a modeled data. Here, observed data is normally accurate compared to modeled data. But, because of limitations of area coverage, timely availability and so on, remote estimates of KSRSAC have been considered. Since the data generated by KSRSAC is having long resolution and pixels, it might have not been so accurate compared to Crop Cutting Experiment estimates i.e. bias might have been noticed. In this study two bias methods are used to bring modeled data (satellite estimates) close to observed data (crop cutting experiment estimates). Further, appropriate prediction models were evaluated for the bias corrected data by following the procedures of model fitting.

Bias corrections

Following two methods were applied to bring the modeled (remote estimates) data close to the observed. Each value is converted with the correction methods.

1. Difference method

In this method, averaged yearly difference (Δx) of observed and modeled values of cropped area is taken. The term (Δx) was considered as a correction factor, which was added to the modeled uncorrected value ($x\ model_{uncor}$) to correct it as ($x\ model_{cor}$) so that the values approach the observed ones.

$$Model_{cor} = Model_{uncor} + (\Delta x)$$

Δ(x) - Averaged difference of observed and modeled values of cropped area.

2. Modified difference method

The modified difference method (MDM) is similar to the difference method (DM); however, some statistical parameters were added to improve the correction function. For example, in area correction, μ and σ are added which aimed at shifting and scaling to adjust the μ and σ^2 (Leander and Buishand, 2007).

$$Model_{cor} = (Model_{uncor} + (\Delta x)) \times \left(\frac{\sigma Area_{obs}}{\sigma Area_{\mathrm{mod}}} \right)$$

Δ(x) - Averaged difference of observed and modeled values of a parameter

Validation of bias corrective measures

The correction capability of the correction measures were tested by coefficient of variation (CV%) expressed as Normalized Root Mean Square Error (NRMSE).

$$NRMSE = \frac{\left[\sum_{i=1}^{n} \frac{(P_i - O_i)^2}{n}\right]^{0.5}}{\bar{O}} \times 100$$

Where,

P_i = Predicted value O_i = Observed value

$\bar{O}$ = Mean of observed value n = Number of observations ranging from 1 to n

Model fitting for bias corrected model data

Data collected is a time series data; Durbin Watson test for autocorrelation was performed to know absence or presence of autocorrelation to the bias corrected data. Growth models (linear/ non-linear such as Linear, Quadratic, Cubic, Exponential, MMF, Rational, Sinusoidal and Logisitic models) or AR/MA/ ARIMA models were considered depending on the outcome of Durbin-Watson test.

The best fit models for Paddy Production were assessed based on the R^2 (Coefficient of determination), Adj. R^2 and NRMSE values. The model with the highest R^2, Adj R^2 and the lowest NRMSE value is considered as the best model.

Diagnostic checking **:** Different models obtained for various combinations of AR and MA individually and collectively are tested using the diagnostics checking such as Plot of residual ACF (plotting the ACF of residuals of the fitted model) and Non-significance of auto correlations of residuals via Portmonteau tests (Q-tests based on Chi-square statistics)-Box-Pierce or Ljung-Box tests.

Box-Pierce statistic (a function of autocorrelations of residuals) whose approximate distribution is Chi-square and is computed as follows:

$$Q = n\sum_{j=1}^{k} r_{(j)}^{\ 2}$$

where n is the number of observations in the series, r (j) is the estimated autocorrelation at lag j; k can be any positive integer and is usually around 20. Q follows Chi-square with (k-m-1) degrees of freedom where m-1 is the number of parameters estimated in the model. A modified Q statistic is the Ljung-box statistic which is given by

$$q = n(n+2)\sum_{j=1}^{k} \frac{r_{(j)}^{\ 2}}{(n-j)}$$

Durbin-Watson (DW) is the ratio of the distance between the errors to their overall variance.

$$DW = \frac{\sum_{t=2}^{N} (e_t - e_{t-1})^2}{\sum_{t=1}^{N} (e_t)^2}$$

where, e_t are residuals from an ordinary least squares regression.

The Durbin Watson test reports a test statistic, with a value from 0 to 4, where DW between 1.5-2.5, implies that auto correlation is absent, otherwise auto correlation is present. If auto correlation is present ARIMA models are tried, otherwise linear and nonlinear models are attempted.

Results and Discussion

The Paddy crop production modeled data (Remote sense data) was subjected to bias correction using 2 methods *viz.,* Difference method (DM) and Modified difference method (MDM). To identify suitable methodology to smoothen the model data NRMSE for each (Model uncorrected, Model corrected by DM and MDM methods) was worked out and results are presented in Table 1.

Table 1 showed that calculated NRMSE values for Paddy Production is least in MDM for all the Districts. This indicated that MDM was a better bias correction method for getting smoothening data compared to DM. Kim *et al.* (2016) indicated that, raw satellite-based rainfall estimates require a post processing of bias correction before data can be useful for forecasting and impact studies. To address this issue, several bias correction methods were suggested.

TABLE 1

NRMSE values of Paddy Production (ha) for selected districts

Districts	Model Uncorrected	Model Corrected	
		DM	MDM
Bellary	10.63	10.36	7.76
Davanagere	10.94	10.58	8.40
Raichur	9.51	9.11	8.16

Bias corrected time series data of Paddy Production has been checked for the auto correlation. Results of Autocorrelation test made with the Durbin-Watson test and are presented in Table 2.

TABLE 2

Durbin-Watson values of Paddy Production

Districts	DW
Bellary	1.43
Davanagere	1.21
Raichur	1.04

* Absence of auto correlation (1.5 < DW > 2.5)

From Table 2, it was observed that Durbin-Watson value is less than 1.5 for paddy production for Bellary, Davanagere and Raichur. This indicated presence of autocorrelation and this leads to fitting of autoregressive models for the Paddy Production. The model fit has been carried out using ARIMA for above districts. The ARIMA models have been fitted for 17 years (1999-2015) model corrected data.

ARIMA model identification involves the determination of the appropriate order of AR and MA polynomials i.e. values of p and q. The graphical representation of Paddy production (tonnes) of Bellary, Davanagere and Raichur districts dipicted in Fig. 1 to 3 clearly indicates that the data series are non-stationary.

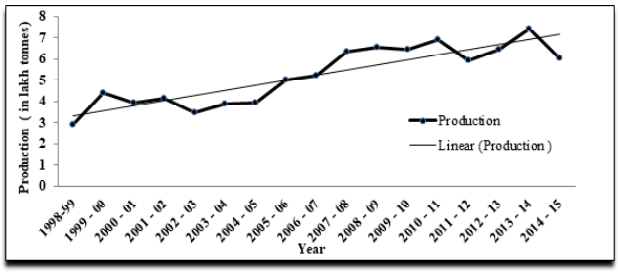

Fig. 1 : Annual Production (lakh tonnes) of Paddy for Bellary district

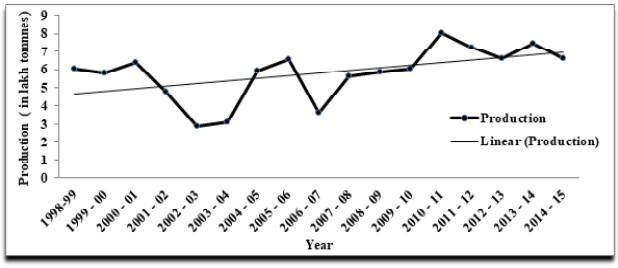

Fig. 2 : Annual Production (lakh tonnes) of Paddy for Davangere district

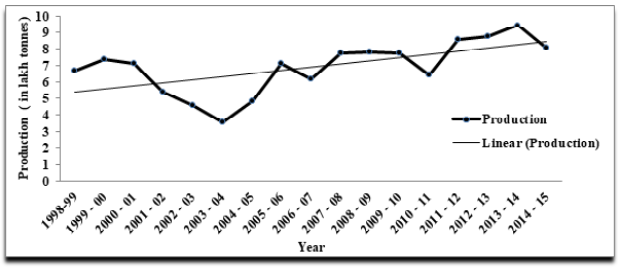

Fig. 3 : Annual Production (lakh tonnes) of Paddy for Raichur district

The plotting ACF (Fig. 4 to 6) indicates that the ACF's decline gradually implying non stationary respectively for Bellary, Davanagere and Raichur districts.

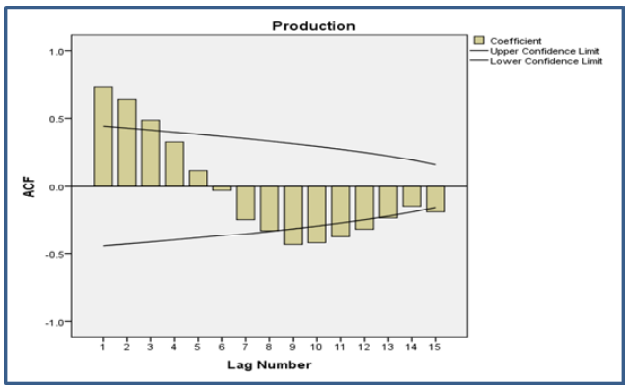

Fig. 4 : Autocorrelations: Bellary district

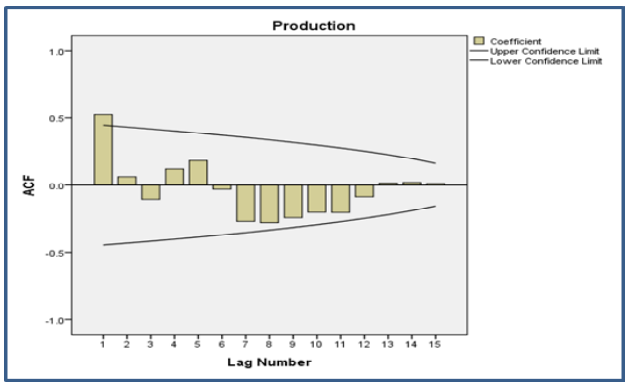

Fig. 5 : Autocorrelations: Davangere district

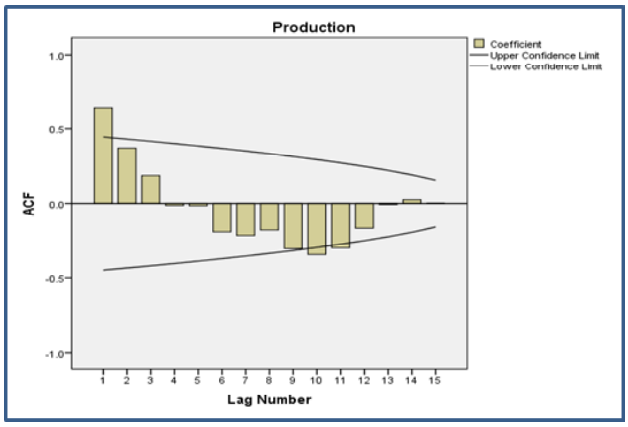

Fig. 6 : Autocorrelations: Raichur district

However, the PACF (Fig. 7 to 9) showed the presence of one significant spike, indicating that the series may have autoregressive component of order one respectively for Bellary, Davanagere and Raichur districts.

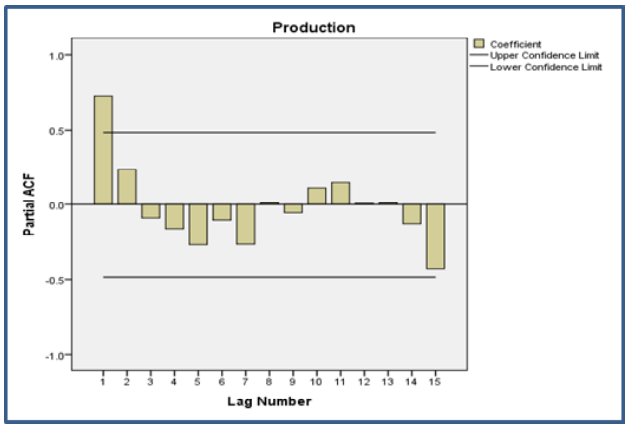

Fig. 7 : Partial Autocorrelations: Bellary district

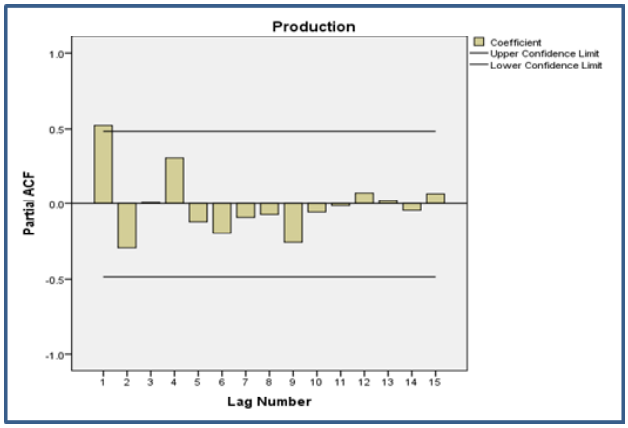

Fig. 8 : Partial Autocorrelations: Davangere district

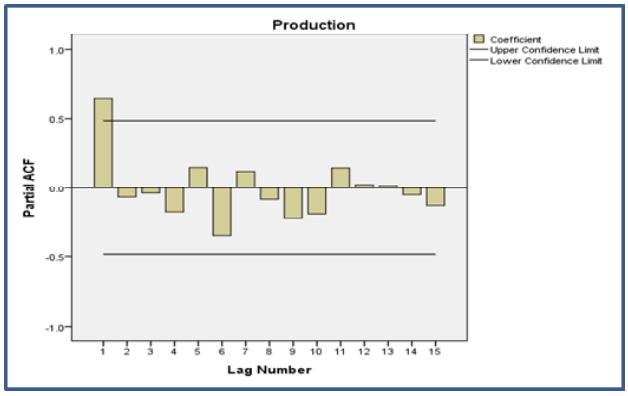

Fig. 9 : Partial Autocorrelations: Raichur district

The non-stationary data series of Bellary, Davanagere and Raichur were transformed into stationary series by the first differencing of the original data series. The plotting Differentiated ACF (Fig. 10 to 12) and Differentiated PACF (Fig. 13 to 15) indicated that differencing of order one *i.e.,* d=1 was enough for getting an approximate stationary series in all the districts.

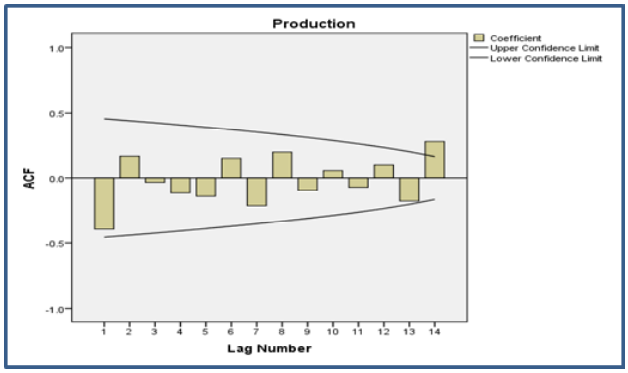

Fig. 10 : Autocorrelations after Difference (1): Bellary district

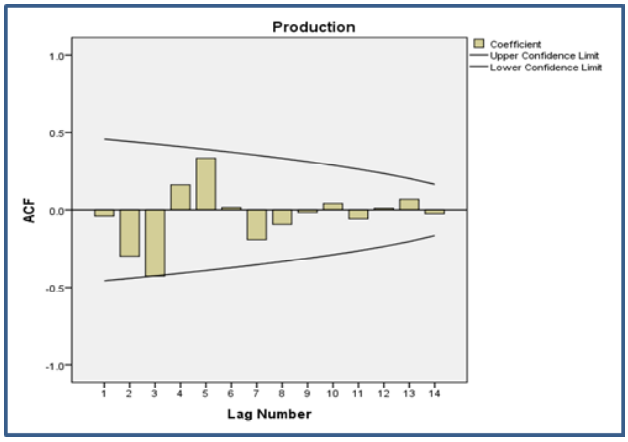

Fig. 11 : Autocorrelations after Difference (1): Davangere district

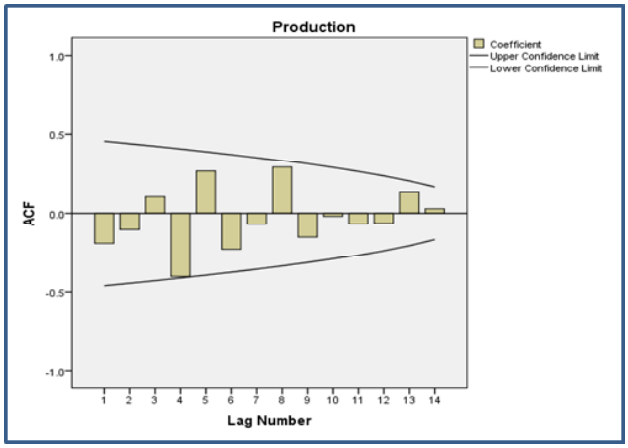

Fig. 12 : Autocorrelations after Difference (1): Raichur district

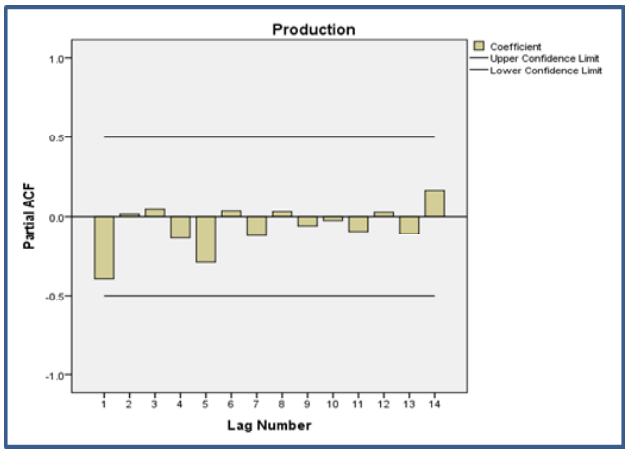

Fig. 13 : Partial Autocorrelations after Difference (1): Bellary district

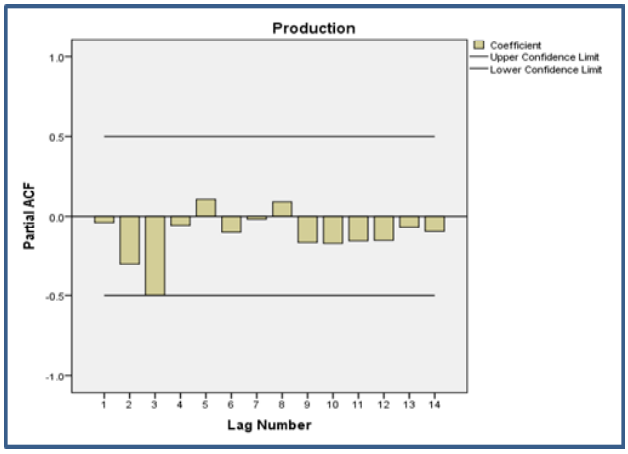

Fig.14 : Partial Autocorrelations after Difference (1): Davangere district

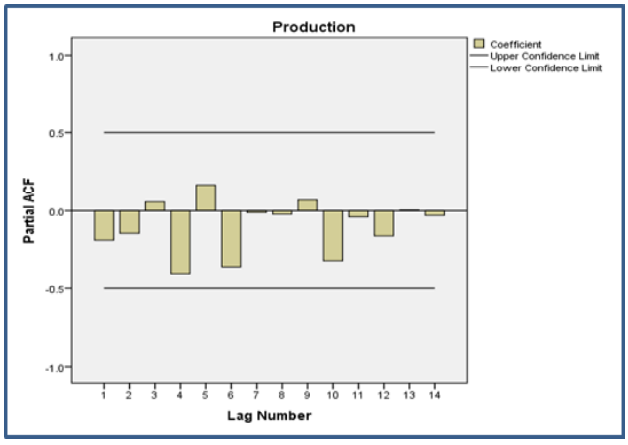

Fig. 15 : Partial Autocorrelations after Difference (1): Raichur district

After experimenting with different lags of moving average and autoregressive process, ARIMA (0,1,0), ARIMA (0,1,0) and ARIMA (0,1,0) were found to be the best fitted models for Bellary, Davanagere and Raichur districts respectively. ARIMA models were also used by Debnath *et al*. (2013) in cotton crop, Gupta *et al*. (2009) in Jute crop, Hamjah (2014) in paddy, Manoj and Madhu (2014) in sugarcane, Meena *et al*. (2014) in oil prices and Prabakaran *et al*. (2014) in pulse producton.

The study resulted that MDM was best method for bias correction showing the least value of NRMSE for area and production of selected districts of Karnataka such as Bellary, Davanagere and Raichur. Hence, MDM can be used for bias correction to the modeled data. In the presence of autocorrelation model fitting was done by ARIMA models for all selected districts for production. ARIMA (0,1,0) was found to be the best fit for production of Paddy crop for all selected districts of Karnataka.

REFERENCES

DEBNATH, M. K., BERA, K. AND MISHRA, P. 2013, Forecasting Area, Production and Yield of Cotton in India using ARIMA Model. *J. of Space Sci.& Tech.,* 2 (1) : 16 - 20.

GALLEGO, F. J., 2006, Review of the Main remote Sensing Methods for crop Area Estimates Agriculture unit", Compilation of ISPRS WG VIII/10 Workshop 2006, remote Sensing Support to crop Yield Forecast and

Area Estimates, Stresa, Italy, Agriculture Unit, IPSC, JRC.

GRAHAM, L. P., ANDRÉASSON, J. AND CARLSSON, B., 2007, Assessing climate change impacts on hydrology from an ensemble of regional climate models, model scales and linking methods - A case study on the Lule River basin.*Clim. Change*, **81** : 293 - 307.

GUPTA, SAHU, D. AND BENERJEE, P. K.. 2009, Forecasting jute production in major contributing countries in the world. *J. of Natural Fibres,* **6** (2) : 127 - 137.

HAMJAH, M. A., 2014, Rice production forecasting in Bangladesh - An application of Box-Jenkins ARIMA model. *Mathematical Theory and Modelling.* **4** (4) : 1 - 11.

KIM, K.B., BRAY, M. AND HAN, D., 2016, An improved bias correction scheme based on comparative precipitation characteristics. *Hydrological Processes*, **29** (9) : 2258 - 2266.

LEANDER, R. AND BUISHAND, T. A., 2007, Resampling of regional climate model output for the simulation of extreme river flows. *J. of Hydrology*, **332**(3) : 487- 496.

MANOJ, K. AND MADHU, A., 2014, An Application of Time Series ARIMA Forecasting Model For Predicting Sugarcane Production In India. *Studies in Business and Eco. J.,* **9** (1) : 81 - 94.

MEENA, D. C., SINGH, O. P. AND SINGH, R., 2014, Forecasting mustard seed and oil prices in India using ARIMA model. *Ann. of Agri-Bio Research,* **19** (1) : 183 - 189.

PRABAKARAN, K., NADHIYA, P., BHARATHI, S. AND ISAIVANI, M., 2014., Forecasting of pulses area and production in India - An ARIMA Approach. *Indian Streams Research J.,* **4** (3) : 1 - 8.

REYNOLDS, C. A., YITAYEW, M., SLACK, D. C., HUTCHISON, C. F., HUELE, A. AND PETERSEN, M. S., 2000, Estimating crop yields and production by integrating FAO crop specific water balance model with real - Time satellite Data and Ground-based Auxiliary Data. *Int. J. of Remote Sensing,* **21** : 3487 - 3508.

WEILAND, S. F. C., BEEK, V. L. P. H., KWADIJK, J. C. J. AND BIERKENS, M. F. P., 2010, The ability of a GCM - forced hydrological model to reproduce global discharge variability, *Hydrol. Earth Sys. Sci.*, **14** (8) : 1595 - 1621.

(*Received :* Janauary, 2018 *Accepted :* April, 2018)