Pol. J. Environ. Stud. Vol. 26, No. 2 (2017), 635-642
DOI: *10.15244/pjoes/65155*

Original Research

Applying Time Series and a Non-Parametric Approach to Predict Pattern, Variability, and Number of Rainy Days Per Month

Alamgir Khalil[1], Subhan Ullah[2], Sajjad Ahmad Khan[3], Sadaf Manzoor[4], Asma Gul[5], Muhammad Shafiq[6]*

[1]Department of Statistics University of Peshawar
[2]M.Phil Scholar Department of Statistics Allama Iqbal Open University Islamabad
[3]Department of Statistics Abdul Wali Khan University Mardan
[4]Department of Statistics Islamia College University Peshawar
[5]Department of Statistics Shaheed Benazir Bhutto Women University Peshawar
[6]Department of Economics, Kohat University of Science and Technology, Kohat

Received: 11 February 2016
Accepted: 8 September 2016

Abstract

In the past 100 years, the annual global temperature has increased by almost 0.5ºC and is expected to increase further with time. This increase in temperature negatively affects the management of water resources globally as well as locally. Rain is an important phenomenon for agriculture, particularly in hilly areas where there is no feasible irrigation system. The present study is concerned with the analysis and modeling of the rain pattern, its variability, and prediction of monthly number of rainy days for the Abbottabad District, which is considered to be one of the greenest and most beautiful areas of Khyber Pakhtunkhwa, Pakistan, by incorporating both parametric and nonparametric techniques. Non-parametric statistical techniques are used for movement detection and significance testing; in this context, statistical tests were incorporated for inspection of homogeneity of rainy days among successive periods. A time series data for the period 1971-2013 was analyzed. Box Jenkins methodology and time series decomposition were applied for fitting the selected model, which was assessed for forecasting the monthly number of rainy days for 2015-2020. In this study several time series parametric and non-parametric approaches were applied to model rainfall data. The results showed that SARIMA (1, 0, 1) (0, 1, 1) was a better choice in predicting the monthly number of rainy days. Further analysis of the data suggests that January, March, May, July, and December have a considerable declining tendency in the number of rainy days.

Keywords: decomposition, Mann Kendall test, SARIMA, Sen slope, time series

*e-mail: mshafiq_stat@yahoo.com

Introduction

Pakistan is a developing country whose economy depends mainly on agriculture, a sector that is more susceptible to climate change than others, as 70% of the country's population depends on this sector. In Pakistan, agriculture makes a significant contribution to the total gross domestic product (GDP), about 21.46%. In consequence, agricultural planners are always keen in boosting its contribution to the country's development [1]. Because crops mostly depend on rainfall, its analysis gains relevance. The only costless indicator accountable for enhancing production of crops is precipitation [2]. Pakistan has been severely affected by floods in the recent past, especially by the flood of 2010 that was caused by historical rains that hit Khyber Pakhtunkhwa and Punjab [3].

Decreasing the amount of rainfall may cause drought. There could be severe impacts of drought on human health such as respiratory problems, mental health, illnesses, food, and water security problems and infectious diseases [4].

The aim of this study was to analyze the monthly number of rainy days for Abbottabad, Khyber Pakhtunkhwa City, Pakistan, located at coordinates 34°11′0″N, 73°15′0″E, and elevation of 1,300 m, with 30,000 inhabitants. This city is important for Pakistan due to its pleasing typical weather and its high-standard learning institutions and military establishments (Pakistan Military Academy) located in the eastern side of the valley. It is a popular hill station that attracts thousands of tourists each year [5]. Its annual rainfall average is 1,193.8 mm, while in monsoon (July and August) 635 mm of precipitation falls [1].

A study [6] reported that hourly rainfall data follows the autoregressive moving average (ARMA) process. Hourly rainfall data was modelled for two stations in the USA and several stations of Italy, which revealed that an event-based estimation procedure presents a better forecast.

Several researchers have investigated annual precipitation trends for various regions of Italy: [7] found a significant negative precipitation trend for Basilicata region while [8] and [9] reported a significant negative precipitation trend for the Sicily region.

Study [10] examined rainfall for the Dodoma region of Tanzania. The amount of rainfall in 1981-2010 was analyzed and fitted a linear regression to test the significance of time trend. The Kruskal-Wallis and Mann-Kendall tau test statistics were also employed for trend detection and testing the homogeneity of mean rainfall. The author concluded that mean annual rainfall through the period 1981-2010 displayed a declining non-significant drift. In study [11] the researchers used 40 years of monthly rainfall data for six stations of western India. They applied a Modified Mann-Kendall test and Thiel Sen's slope estimator, and they developed a Box-Jenkins autoregressive integrated moving average (ARIMA) model for forecasting. The analysis suggested an overall positive trend.

Material and Methods

Time series data for the monthly number of rainy days for Abbottabad in 1971-2013, taken from Regional Meteorology Department Peshawar, Pakistan, were selected for analyzing and forecasting the precipitation trend. For this purpose, a day is considered rainy if the amount of rainfall is 2.5 mm or more [12]. The programs used were Minitab 17, Gretl 1.9.8, and MS Excel 2007. Both parametric and non-parametric statistical techniques were incorporated in the analysis. A brief introduction of these techniques is given in the next section.

The Box-Jenkins methodology consists of four stages: identification, estimation, diagnostic check, and forecasting. The first plot of ACF and PACF is constructed. This not only tells about the stationarity of the series but it also identifies the parameters of the models: autoregressive (AR) terms, moving average (MA) terms, and order of the differencing (*d*) for the non-seasonal model, which is written as ARIMA (*p, d, q*; *p* = order of autoregressive term and *q* = order of moving average term). When the data exhibit seasonality and the purpose is to utilize this behavior, then the SARIMA model is used. The multiplicative seasonal arima is known as SARIMA (*P, D, Q) * (p, d, q*; *p* = seasonal autoregressive, *q* = seasonal moving average component, and *d* = seasonal difference). In the backshift notation the SARIMA model without intercept is given as follows:

$$\phi_{AR}(B)\Phi_{SAR}(B)(1-B)^{d}(1-B^{s})^{D}X_{t} = \\ = \theta_{MA}(B)_{SMA}(B^{s})\epsilon_{t} \quad (1)$$

$$\phi_{AR} = (1 - \emptyset_1 B - \emptyset_2 B^2 - \cdots - \emptyset_p B^p) \quad (2)$$

$$\Phi_{SAR} = (1 - \Phi_1 B^S - \Phi_2 B^{2S} - \cdots - \Phi_P B^{PS}) \quad (3)$$

$$\theta_{MA}(B) = \left(1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q\right) \quad (4)$$

$$_{SMA}(B^s) = (1 + {}_1B^S + {}_2B^{2S} + \cdots + {}_Q B^{QS}) \quad (5)$$

…where ϕ_{AR}, Φ_{SAR}, θ_{MA}, Θ_{SMA} and represent non-seasonal autoregressive operator, seasonal autoregressive operator, non-seasonal moving average operator, and seasonal moving average operator respectively; "s" is the number of periods per season. For seasonal model, the significance of seasonal lags are considered; for this case, if monthly data is used then the significant behavior of lag12, lag24, lag36, and so on are used for modeling. If the data is collected quarterly, then lag4, lag8, lag12, and so on are used for model fitting.

For non-seasonal, a pure AR model is fitted if autocorrelation decays geometrically, while a pure MA model is fitted if partial autocorrelation decays

geometrically. If both autocorrelation and partial autocorrelation decay exponentially, then the ARMA model is fitted. The parameters of autoregressive (both AR and Seasonal AR) and moving average (MA Seasonal MA) components are estimated generally by applying the least square technique and maximum likelihood estimation.

Decomposition

Decomposition is concerned with disintegrating time series observations into trend, seasonal, and random components:

$$Y_t = T_t + S_t + E_t$$

…where Y_t is the value of time series variable at time t, and T_t, S_t, and E_t are the trend component, seasonal effect, and random error, respectively, at time t. This technique fits the trend line to the seasonal adjusted data; further seasonal indices are computed to de-trended data, which is the seasonal effect at that time.

Seasonal index = average of season – overall average

Finally, it computes estimation (prediction) by adding (multiplying) appropriate seasonal indices to the trend-computed values (depending on whether the model is additive or multiplicative).

Thiel Sen's Estimator

Thiel Sen's estimation is a non-parametric technique used for finding the slope of the time series; in order to calculate this estimator for slope of the trend equation, all possible slopes are computed using the formula:

$$m_i = \frac{x_j - x_k}{j - k}$$

For all $j > k$ and i =1, 2, 3..., N (N = number of observations).

Median of all slopes is the slope of Thiel Sen's estimator.

Mann-Kendall Trend Test

The Mann-Kendall (MK) test is a non-parametric test widely used for trend detection (not necessarily linear) of time series data, when the assumptions of homoscedasticity and normality are not satisfied, and data show missing values and outliers. The test statistic to be used is S, calculated as follows:

$$S = \sum_{k}^{N-1} \sum_{j=k+1}^{N} sign(x_j - x_k)$$

…where x_j and x_k are the values of the series at time j and k, for all j greater than k. The total number of observations is N. The sign function is as follows:

$$sign(\delta) = sign(x_j - x_k)$$

$$= \begin{cases} 1 & if \quad \delta > 0 \\ 0 & if \quad \delta = 0 \\ -1 & if \quad \delta < 0 \end{cases}$$

When the size of the series is large with N greater than 10, then the distribution of S is approximately normal with mean equal to zero and variance given by:

$$var(S) = \frac{N(N-1)(2N+5) - \sum_{k=1}^{n} t_k(t_k-1)(2t_k+5)}{18}$$

Here t_k represents the number of ties of kth observation. Then the Z-test is as follows:

$$Z = \begin{cases} \frac{S-1}{\sqrt{var(S)}} & if\ S > 0 \\ 0 & if\ S = 0 \\ \frac{S+1}{\sqrt{var(S)}} & if\ S < 0 \end{cases}$$

This is a two-tailed test. The null hypothesis is rejected if $|Z| \geq Z_{\frac{\propto}{2}}$, where α is the size of critical region [13].

Kruskal-Wallis Test

The Kruskal-Wallis test is a non-parametric alternative to the one-way analysis of variance. It is used to test whether K sample came from the same distribution, as samples share the same median. The test statistic to be used is:

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1)$$

Under the null hypothesis of the same median, statistic H is a chi square distribution with (k-1) degrees of freedom; in this test, n_j is the number of observations in collection j, while N is the number of total observations in all the collections.

R_j = Sum of ranks of the n_j observations of the collection j (= 1, 2, 3, 4…)

In this paper, the data period (1971-2013) were divided into five time periods (1971-79, 1980-88, 1989-97, 1998-2006, and 2007-13). The null hypothesis to be tested was

Table 1. Patterns of ACF and PACF.

Type of model	Trend of ACF	Trend of PACF
AR(P)	Decays exponentially	Cuts up after P seasonal lags
MA(Q)	Cuts up after Q seasonal lags	Decays exponentially
ARMA(P, Q)	Decays exponentially	Decays exponentially

that the monthly numbers of rainy days follow the same distribution in the five time periods (t_1, t_2, t_3, t_4, t_5) [11].

Results and Discussion

Descriptive statistics computed for the monthly number of rainy days for 1971-2012 are shown in Table 2. It is evident from the correlogram in Fig.1 that data shows strong seasonality, hence a seasonal difference was assumed to make it stationary. The correlogram in Fig. 2 was constructed after accepting seasonal difference. Table 3 shows the result of the augmented Dicky Fuller test that was applied before and after taking seasonal difference. The result suggests that the series is stationary after seasonal differencing.

Identification of the Model

In Fig. 2 it is noticeable that autocorrelation (ACF) dies down, which suggests that $q = 0$; similarly, partial autocorrelation (PACF) also dies down, suggesting $p = 0$. On the other hand, the autocorrelation at lag 12 (seasonal lag) was significant but non-significant at seasonal lag24. ACF at seasonal lags is dying down, while partial autocorrelations at lag 12, lag 24, and lag 36 were significant. This suggests $Q = 1$ and $P = 0$. So the tentative model is *SARIMA* $(0,0,0)(0,11)_{12}$.

Besides this model, other models in the neighborhood were also fitted. Table 4 shows the best models on

Table 2. Descriptive statistics.

Mean	Median	Minimum	Maximum
9.58929	9.00000	0.000000	27.0000
Std. Dev.	C.V.	Skewness	Kurtosis
5.22381	0.544755	0.389920	-0.171064

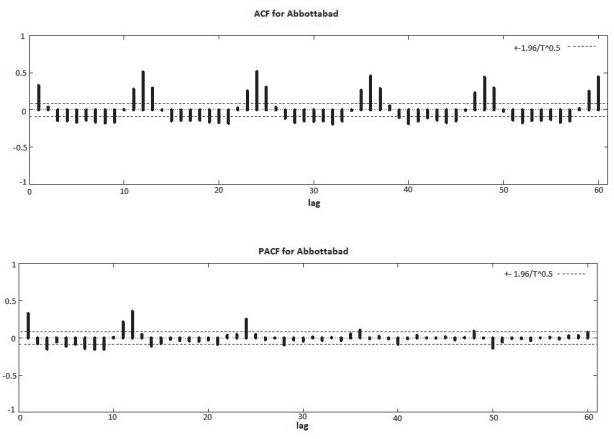

Fig. 1. Correlogram at level.

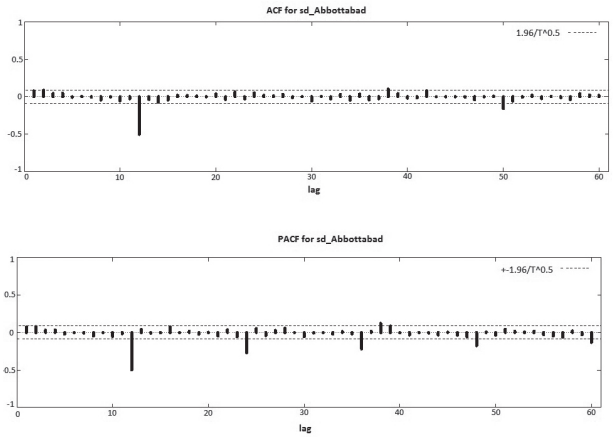

Fig. 2. Correlogram after seasonal difference.

Table 3. Unit root test.

Test at level		Test after seasonal difference	
Test statistic	-0.881377	Test statistic	-13.7863
p-value	0.3344	p-value	
Truncated lag	10	Truncated lag	10
Remarks	There is a unit root	Remarks	Series is stationary

Table 4. Models with minimum AIC value.

S. no	Model	AIC
1	*SARIMA* $(1,0,1)(0,1,1)_{12}$	2,744.235
2	*SARIMA* $(1,0,1)(1,1,1)_{12}$	2,745.301
3	*SARIMA* $(1,0,1)(0,1,2)_{12}$	2,745.434
4	*SARIMA* $(1,0,1)(2,1,1)_{12}$	2,744.956

Table 5. *SARIMA* (1, 0, 1) (0, 1, 1).

Model	Parameter	Estimate	z-value	p-value
SARIMA $(1,0,1)$ $(0,1,1)_{12}$	AR1	0.845982	6.7063	<0.00001
	MA1	-0.777589	-5.3055	<0.00001
	SMA1	-0.942366	-27.8174	<0.00001

Table 6. Modified Box-Pierce (Ljung-Box) test.

Lag	12	24	36	48	60
Chi-Square	5.0444	13.5699	19.1297	24.0106	37.4240
P-Value	0.956	0.675	0.990	0.999	0.990

the basis of minimum AIC value. It is evident from Table 4 that the value of AIC was the smallest for *SARIMA* (1, 0, 1) (0, 1, 1).

Numerical Diagnostics

Results of the Ljung Box test were non-significant at all lags, as they did not show a serial correlation among the errors (Table 6). Table 7 contains the result of the Jarque-Bera test of normality for residuals showing normality.

Decomposition

The trend equation fitted to the series after seasonal adjustment is:

$$Y_t = 10.438 - (0.00336) * t$$

The negative slope indicates that rainy days decrease by 0.00336 for unit change in time when seasonal effect is eliminated. In Table 8, the seasonal indices refer to changes occurring in data for the selected period (month), which are negative for January, May, and September-December. The trend suggests that mean rainy days decreases in these months on average, as compared with overall data. The maximum decrease is observed for November (Seasonal Index = -6.858).

The seasonal indices are positive for the periods of February-April and June-August, and the highest index is for the month of July (Seasonal Index = 7.684), which is the first month of the monsoon rainy season in Abbottabad.

Forecast Accuracy Comparison of SARIMA and Decomposition

Table 9 was constructed for comparing the forecasting accuracy of the fitted models using MFE (mean forecast error), MAFE (mean absolute forecast error), SSFE (sum of square of forecast error), and MSFE (mean square forecast error). From Table 9 it can be inferred that

Table 7. Jarque-Bera test for normality.

Test statistic	0.435651
p-value	0.804266
Conclusion	Data is normally distributed

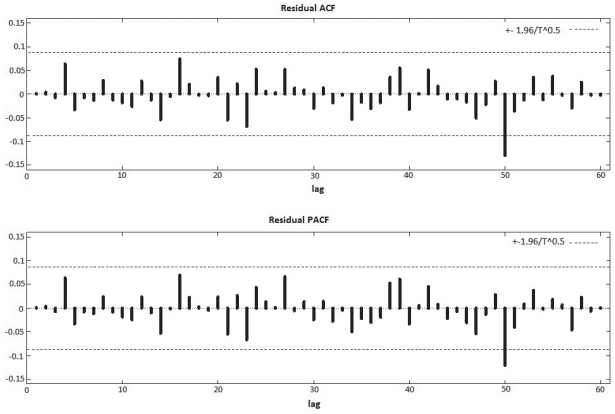

Fig. 3. Correlogram of residual.

Table 8. Seasonal indices for rainy days of Abbottabad.

Period	January	February	March	April	May	June
Seasonal index	-1.191	1.017	3.726	1.809	-0.233	0.726
Period	July	August	September	October	November	December
Seasonal index	7.684	2.976	-1.066	-4.024	-6.858	-4.566

Table 9. Comparison of forecast errors.

Model	MFE	MAFE	SSFE	MSFE
SARIMA $(1,0,1)(0,1,1)_{12}$	-0.04858	2.223182	129.1306	10.76088
Decomposition	0.694914	2.582516	161.0717	13.42265

Table 10. Forecasted monthly number of rainy days for the period 2015-20.

Month	2015	2016	2017	2018	2019	2020
January	7.42138	7.41536	7.41456	7.41445	7.41443	7.41443
February	10.1627	10.1576	10.1569	10.1568	10.1568	10.1568
March	11.9253	11.9210	11.9204	11.9203	11.9203	11.9203
April	11.3333	11.3296	11.3291	11.3291	11.3291	11.3291
May	8.56628	8.56320	8.56279	8.56273	8.56272	8.56272
June	10.1951	10.1925	10.1921	10.1921	10.1921	10.1921
July	16.1853	16.1831	16.1828	16.1828	16.1828	16.1828
August	13.1412	13.1393	13.1391	13.1390	13.1390	13.1390
September	8.98336	8.98179	8.98157	8.98155	8.98154	8.98154
October	5.09732	5.09599	5.09581	5.09579	5.09578	5.09578
November	3.37855	3.37742	3.37727	3.37725	3.37724	3.37724
December	5.06999	5.06903	5.06890	5.06888	5.06888	5.06888

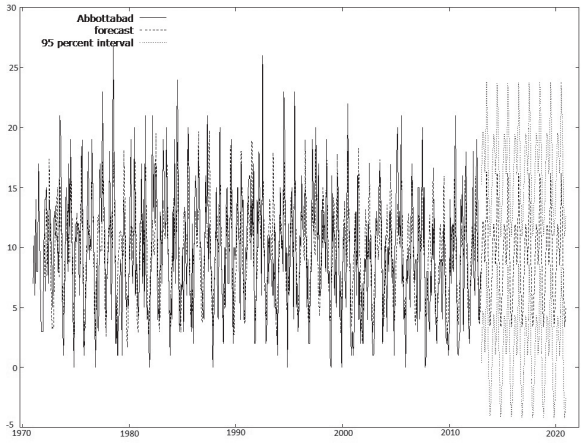

Fig. 4. Time series graph for 1971-2020.

the error measurements for the SARIMA model are a minimum. Hence, the SARIMA model provides a better fit in the current study.

Table 12. Periods, sample size, and median for Kruskal-Wallis test.

Periods	Sample size	Median	Average rank
1971-79	108	10	265.7
1980-88	108	9	263.0
1989-97	108	10	275.6
1998-2006	108	9	237.7
2007-13	84	9	248.2

Table 11. Sen slope and Mann-Kendall tests.

Month	Sen slope	Mann-Kendall statistic	p-value	Remarks
January	-0.0667	-2.002*	0.0226	Significant decreasing trend
February	-0.0357	-0.885	0.1880	Non-significant increasing trend
March	-0.1200	-1.777*	0.0378	Significant decreasing trend
April	0	-0.494	0.3110	Non-significant decreasing trend
May	-0.0870	-1.795*	0.0364	Significant decreasing trend
June	0.0250	0.802	0.2110	Non-significant increasing trend
July	-0.1000	-1.787*	0.0370	Significant decreasing trend
August	0	-0.3260	0.3720	Non-significant decreasing trend
September	0	-0.2320	0.4080	Non-significant decreasing trend
October	0	-0.4550	0.3250	Non significant decreasing trend
November	0	0.2640	0.3960	Non-significant increasing trend
December	-0.0909	-1.7050*	0.0441	Significant decreasing trend

Table 13. Kruskal-Wallis test.

Test without adjusting for ties		Test adjusting for ties	
Test statistic (H)	4.27	Test statistic (H)	4.28
Degrees of freedom	4	Degrees of freedom	4
P-value	0.371	p-value	0.369

Forecast with *SARIMA* (1, 0, 1) (0, 1, 1)

Table 10 shows the forecasted monthly number of rainy days for 2015-20 by applying *SARIMA* (1, 0, 1) (0, 1, 1). Fig. 4 shows the complete graph for 1971-2020, where observed series consist of the period 1971-2012 and the forecast period 2013-20. The shaded area refers to the forecasted prediction interval.

Non-parametric Analysis

Table 11 shows that the Mann-Kendall test for negative trend was significant only for January, March, May, July, and December – months when the monthly number of rainy days declined. Kruskal-Walls test was applied on the data for 1971-2013 (Table 12). The Kruskal-Wallis test results were applied with and without adjusting matched observations (Table 13). Based on the outcome of the Kruskal-Wallis tests, we concluded that the monthly number of rainy days follows almost the same pattern.

Conclusion

In Pakistan's Abbottabad District, agricultural production is highly vulnerable to rainfall variability. In consequence, farmers in this area are eventually affected by changes in the rainfall pattern. This study evaluated the pattern, variability, and trend of the number of rainy days based on past data to help farmers with information about possible climate change in the future.

In this paper, data on the monthly number of rainy days were analyzed by applying time series analysis and a non-parametric approach. Several SARIMA models were fitted and the best models were chosen. It was concluded that SARIMA was a better choice in predicting the monthly number of rainy days as compared to decomposition. Forecasts for the period 2015-20 were also obtained and presented, and are expected to be very useful for future planning and policy making.

Analysis of the data suggests that January, March, May, July, and December have a considerable declining tendency in the number of rainy days. Additionally, it was assessed that the distribution of monthly number of rainy days is identical among different time periods.

Based on the findings of this study, we recommend planning at all levels for risks involved in climate change and to help farmers face the declining effects of the number of rainy days. Such information is required to be conveyed and propagated on time in order to increase its practical significance to the general public and to farmers. Investing in new irrigation schemes is recommended, particularly for those months where there is a declining trend in the amount of rainfall.

Abbreviations

ACF - Autocorrelation Function
PACF - Partial Autocorrelation
AR - Autoregressive
SAR - Seasonal Autoregressive
MA - Moving Averages
SMA - Seasonal Moving Average
ARIMA - Autoregressive Integrated Moving Average
SARIMA - Seasonal autoregressive integrated moving average
MK – Mann-Kendall
AIC - Akaike Information Criterion
GDP- Gross Domestic Product
MFE - Mean forecast error
MAFE - Mean absolute forecast error
SSFE - Sum of square of forecast error
MSFE - Mean square forecast error

References

1. PBS, GOVT. OF PAKISTAN *Agriculture Statistics*. Retrieved from PBS Web site: http://www.pbs.gov.pk/content/agriculture-statistics. (accessed October 3, 2014). **2014**.
2. HUSSAIN Z., MAHMOOD Z., HAYAT Y. Modelling the daily rainfall amonts of north-west Pakistan for agriculture planning. Sarhad Journal of Agriculture, **27** (2), 313-321, **2011**. http://www.aup.edu.pk/SJA-search.php
3. ASHRAF S., IFTIKHAR M., SHAHBAZ B., KHAN G.A., LUQMAN M. Impacts of flood on livelihoods and food security of rural communities: a case study of southern Punjab, a Pakistan. Pakistan Journal of Agricultural Science, **50** (4), 751-758, **2013**.
4. YUSA A., BERRY P., J CHENG J., OGDEN N., BONSAL B., STEWART R., WALDICK R. Climate change, drought and human health in Canada. International journal of environmental research and public health, **12** (7), 8359, **2015**.
5. KAUSAR R., MIRZA S.N., SABOOR A., SALEEM A., KHALID B. Role of ecotourism in promoting and sustaining conservation of nature: A case study of Murree forest recreational resort. Pakistan Journal of Agricultural Science, **50** (3), 463, **2013**.

6. BURLANDO P., ROSSO R., LUIS G.C., JOSE D.S. Forecasting of short-term rainfall using ARMA models. Journal of Hydrology, **144** (1-4), 193, **1993**.
7. PICCARRETA M., CAPOLONGO D., BOENZI F. Trend analysis of precipitation and drought in Basilicata from 1923 to 2000 within a Southern Italy context. International Journal of Climatology, **24**, 907, **2004**.
8. CANCELLIERE A., ROSSI G. Droughts in Sicily and comparison of identified droughts in Mediterranean regions. In Tools for drought mitigation in Mediterranean regions, Rossi G., Cancelliere A., Pereira LS., Oweis T., Shatanawi M., Zairi A. (eds), 103, **2003**.
9. CANNAROZZO M., NOTO LV., VIOLA F. Spatial distribution of rainfall trends in Sicily (1921-2000) . Physics and Chemistry of the Earth, **31**, 1201, **2006**.
10. KASSILE T. Trend analysis of monthly rainfall data in central zone. Journal of Mathematics and Statistics, **9** (1), 1, **2013**. DOI:10.3844/jmssp.2013.1.11.
11. NAYARANAN P., BASISTHA A., SARKAR S., SACHDEVA K. Trend analysis and ARIMA modelling of pre-monsoon rainfall data for western India. Comptes Rendus Geoscience, **345** (1), 22, **2013**. DOI: 10.1016/j.crte.2012.12.001.
12. SHUKLA P. Climate Change and India: Vulnerability Assessment and Adaptation. Universities Press, Hyderabad, **2003**
13. KAHYA E., KALAYCI S. Trend analysis of streamflow in Turkey. Journal of Hydrology, **289**, 128, **2004**.