Int. j. adv. multidisc. res. stud. 2022; 2(4):883-887

Received: 08-07-2022
Accepted: 18-08-2022

International Journal of Advanced Multidisciplinary Research and Studies

ISSN: 2583-049X

Analyzing the Impacts of Climatological Variables on Rainfall using ARIMA and Multi Linear Regressions models: A case study of Algiers

[1] Hüseyin Gökçekuş, [2] Youssef Kassem, [3] Ansumana A Bility

[1, 2] Department of Civil Engineering, Civil and Environmental Engineering Faculty, Near East University, 99138 Nicosia, Cyprus
[2] Department of Mechanical Engineering, Engineering Faculty, Near East University, 99138 Nicosia, Cyprus
[1, 2] Energy, Environment, and Water Research Center, Near East University, 99138 Nicosia, Cyprus
[1, 2] Engineering Faculty, Kyrenia University, 99138 Kyrenia, Cyprus
[3] Department of Environmental Engineering, Civil and Environmental Engineering Faculty, Near East University, 99138 Nicosia, Cyprus

Corresponding Author: **Youssef Kassem**

Abstract

As humanity races for a solution to its making, largely, the crisis is degenerating and becoming vicious- solving one problem creates another. For instance, the call for an energy transition from oil to hydro-power partially tackles the emission issue but offsets marine ecosystem stability. The most significant damage of the humanly induced climate crisis has been observed in the water resource sector. Drier regions are becoming desserts while flood takes over wetter places. The hydrological cycle is most affected as the water crisis surges globally. Precipitation patterns are affected and vary from region to region. Understanding the underlying variables that influence the irregularity is required to tackle this crisis. The current report summarizes several methods used to predict precipitation for the North African city of Algiers, Algeria. The ARIMA and the Multiple Linear Regression models were used to analyze the relationship between rainfall and other parameters such as temperature, runoff, vapor pressure, windspeed, evapotranspiration, soil moisture, etc.). The probability plot was used to determine the best distribution function for precipitation. The results suggest that solar radiation, soil moisture, runoff, and maximum temperature affect precipitation more, recording an R-square value of 35.16% when using the MLR compared to other explanatory variables. For the ARIMA model, The P-value is less than the significance level ($p \leq 0.05$). This suggests the coefficient for the autoregressive term is statistically significant; therefore, the term should be maintained in the model. Using the Modified Box-Pierce (Ljung-Box) Chi-Square statistic, it was found that the p-values are all less than the significance level of 0.05 and significant correlations exist for the autocorrelation function of the residuals. Thus, we can conclude that the model meets the assumption that the residuals are dependent. For the probability plot, the 3-Parameter gamma produced the most negligible AD value (11.008). This makes the 3-parameter gamma the best fit for the distribution function.

Keywords: Machine learning models, climate change, water resources, ARIMA, MLR, Algiers

1. Introduction

1.1 Background

Climate change is arguably the most complicated issue facing humanity and produces the most uncertain solution-based approach than at any moment in human history. As humanity races for a solution to its making, largely, the crisis is degenerating and becoming vicious- solving one problem creates another. For instance, the call for an energy transition from oil to hydro-power partially tackles the emission issue but offsets marine ecosystem stability. The most significant damage of the humanly induced climate crisis has been observed in the water resource sector. The hydrological cycle is under siege as climate pattern fluctuates from region to region. Global temperature surge is accompanied by rapid evaporation- something that offsets the balance in the hydrological cycle, thus, leading to declining in available water for drier regions and cases of flood in wetter areas. Rapid urbanization, the construction of complex infrastructure, and changes in precipitation patterns caused by anthropogenic climate change make cities more susceptible to flooding. Algiers, Algeria's capital, is considered one of the water-stressed settlements in the world. The climate is semi-arid, and population growth is unprecedented, leading to over-exploitation of groundwater resources.

The major water resources in Algiers are groundwater, surface water, and desalination. The current review assesses the impact of other climatological variables on rainfall using machine learning models to understand which variables have the most influence on precipitation in the city. Considering the uncertainties associated with predicting climatic conditions, multiple models have been developed and applied to simulate the impact of changing climates on rainfall and other constituents in the natural hydrological cycle. These models are generally termed General Circulation Models, GCMs. The selection of an appropriate model depends on factors such as parameter, duration, and magnitude of change taking place.

1.2 Rainfall Prediction Models (Models)

There are several models available for predicting rainfall. Varying factors determine the choice model, as stated above. The models are categorized into two: dynamic and empirical models. The dynamic approaches focus primarily on physics-based models, such as numerical weather prediction models. In contrast, empirical methods, such as regression, fuzzy logic, artificial neural network (ANN), and ARIMA, assess past precipitation data and their relationship with other meteorological variables (Swain *et al*., 2018) [6]. Two empirical models popular for predicting rainfall are the Autoregressive Integrated Moving Average (ARIMA) and the Artificial Neural Networks, ANNs.

1.2.1 The ARIMA Model

The ARIMA model assumes that all-time series are linear and obey a particular distribution, such as the normal distribution. Box and Jenkins (1970) were the ones who initially created this method of forecasting, which is currently used extensively in the field of hydrology (Somvanshi, *et al*., 2006) [5]. An ARIMA model is made up of three operators: an autoregressive (AR) function applied to historical values of a parameter (rainfall), a moving average (MA) function applied to completely random values, and an integration (I) part to reduce the difference between them, which is referred to as the differencing operator (Swain *et al*., 2018; Valipour *et al*., 2013; Kaushik & Singh, 2008) [6, 7, 4].

1.2.2 Artificial Neural Networks (ANNS)

Due to its non-linear characteristics and accuracy, Artificial Neural Networks are becoming increasingly valuable for predicting rainfall. Unlike the ARIMA, the ANNs are non-presumptive. The ANN approach has several advantages over conventional phenomenological or semi-empirical models because it requires a set of known input data without making any assumptions (Chattopadhyay, 2007 [1]; Gardner & Dorling, 1998; Nagendra & Khare, 2006).

1.2.3 Multiple Linear Regression

Multiple Linear Regression, MLR is a model that explains the relationship between input and output variables.

1.3 Aim of the study (Aim of study)

The study aimed to analyze the relationship between rainfall and other variables such as maximum temperature based on available data to inform predictions. The overall objective was to inform policy decisions on effective water management, given the significant need for water in an era that has seen a rapid decline in the quantity and quality of global freshwater due to climate change.

2. Materials and methods (Location, Climate, and Map)

2.1 Study area

The rainfall and temperature data were analyzed for the city of Algiers, Algeria, Northern Africa. Algiers is the capital of Algeria, with an estimated population of approximately 2.5 million people. The city is planned mainly on the hills, overlooking the sea. The city uses a 4km grid cell center at 37°C North and Longitude 3°C East. Algiers's climate is the Mediterranean according to the Koppen Climate Classification. Due to its situation along the Mediterranean, the city's annual temperature is moderate, with yearly rainfall averaging 600mm. The city receives peak rainfall between October and April. In July, the city's midday temperatures reach 83 degrees Fahrenheit (28 degrees Celsius), dropping to around 70 degrees Fahrenheit (21 degrees Celsius) at night, while daily temperatures in January range between 59- and 49 degrees Fahrenheit (15 and 9 degrees Celsius). Between October and March, the city receives four-fifths of its annual precipitation (760 mm), whereas July and August are often dry. Below (Fig 1) is an imported map of Algiers city. As can be observed, the city is situated along the Mediterranean.

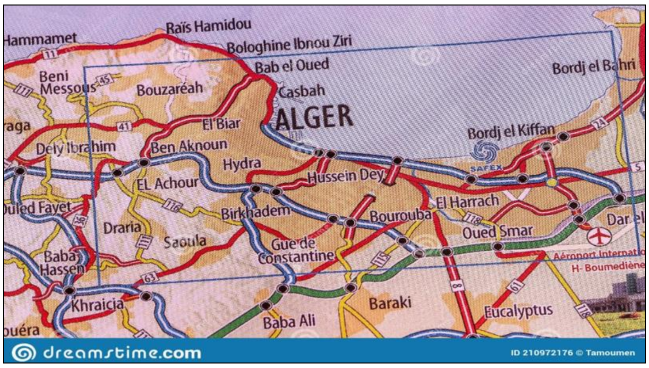

Fig 1: Map of Algiers (https://thumbs.dreamstime.com/z/map-algiers-map-capital-city-algeria-algiers-bigest-city-maghreb-region-210972176.jpg)

2.2 Methods

2.2.1 Data collection

All Data used for the analysis, including but not limited to precipitation, wind speed, temperature, humidity, soil cover, runoff, solar radiation, vapor pressure, etc., were downloaded using the climate database at https://climatetoolbox.org/tool/data-download. The compiled data is a summary of climatic data from 1958 to 2020. The unit for measurement employed was the metric unit.

2.2.2 Why satellite data

In today's climate research, satellite data presents researchers with readily available data in short and real-time than normal meteorological stations would do. The study uses satellite data because it is 1) cost-effective and 2) because it can ensure the collection and analysis of massive datasets in a short period.

2.2.3 Data analysis

To analyze and predict the outcome, this research used the ARIMA model in conjunction with the Multiple Linear

Regression using the software Minitab, version 17.

1) ARIMA is a statistical analysis model that uses time-series data to either better comprehend the data set or predict future trends. Three parameters are used for analysis in the ARIMA model: the number of lag observations denoted by p, the number of times these raw observations are differenced indicated by d, and the order of moving average denoted by q. To interpret an ARIMA model, three considered: 1) determine whether each term in the model is significant using the p-value against the significance level. The null hypothesis holds if the p-value is less than or equal to the significance level. 2) determine how well the model fits the data using the Mean Square Error (MS) value. The smaller the value, the better the fit. 3) determine whether the model agrees with the assumption of the analysis Ljung-Box chi-square statistics and the residuals' autocorrelation function.

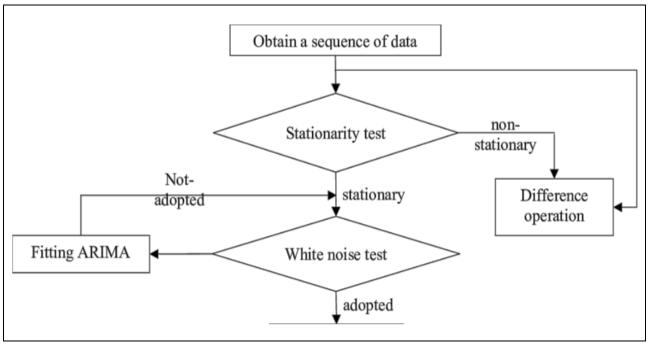

Fig 2: ARIMA Flowchat (https://www.researchgate.net/figure/Flow-chart-of-ARIMA-model_fig1_336549097

2) Multiple linear regression is a statistical technique that takes several independent variables (Explanatory) and manipulates them to derive a dependent variable (Response). The function gives the formula for the MLR:

$$yi=\beta0+\beta1xi1+\beta2xi2+\ldots.+\beta pxip+\epsilon$$

where for $i=n$ observations:
yi=dependent variable
xi=explanatory variables
$\beta0$=y-intercept (constant term)
βp=slope coefficients for each explanatory variable
ϵ=the model's error term (also known as the residuals)

We deployed the R-Square values to test the model accuracy, generated using excel. The R-Square values are used to determine the accuracy of models for predicting dependent variables. The values range from 0 to 1. The closer the r-square value is to 1, the more accurate the model explains the relationship between the explanatory and response variables. The closer the matter is to 0, the more inaccurate the model determines the connection from a regression trendline. R- square values are given as percentages. The plotting on the left indicates a low R-Square compared to the one on the right.

3) The Distribution function depicted by the probability plot was also used to determine the best distribution function to analyze the goodness of fit. In a probability plot, each data point is displayed in relation to the percentage of values in the sample that is either greater than or the same. The most important value in the probability plot is the Anderson-Darling goodness of fit, AD value. AD measures the deviations between the fitted line (based on the specified distribution) and the nonparametric step function (based on the data points). Smaller AD values generally imply the data following the distribution more closely.

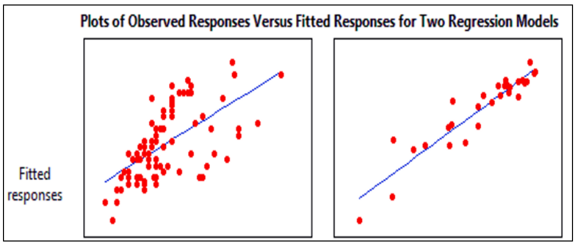

Fig 3: Typical R-Square determination graph https://blog.minitab.com/hubfs/Imported_Blog_Media/fittedxobserved.gif

3. Results and discussion

3.1 General Characteristics of data

The general characteristic of the data portrays slightly right-skewness, as depicted by the histogram (Fig 3). The kurtosis and skewness values, as shown in table 1; 0.74 & 1.04, respectively, suggest that the distribution is almost normal with a positive tail. The mean and standard deviation (59.7 & 54.8) indicate moderate dispersion of the data. The Q2 value (93.3) suggests that a significant amount of the data lie within the upper quartile.

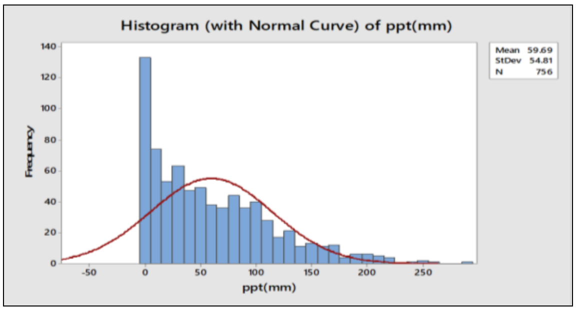

Fig 4: Histogram

Table 1: Descriptive Statistic

Mean	59.7
Standard Deviation	54.8
Variance	3000.2
Covariance	91.8
Minimum	0.0
Q1	12.3
Median	46.6
Q2	93.3
Maximum	286.9
IQR	81
Skewness	1.04
Kurtosis	0.74

3.2 Probability Plot

The table below (Table 2) displays results from several distributions generated from the use of probability plots used to measure the accuracy of precipitation. The AD values used to measure the goodness of fit were considered for the different distributions. The 3-Parameter gamma produced the most negligible AD value (11.008). This makes the 3-parameter gamma the best fit for the distribution function.

Table 2: distributions generated from the use of probability plots

Distribution	AD value	Meaning
Normal	20.071	
Lognormal	N/A	It does not work with negative values
3-parameter lognormal	17.055	
Gamma	N/A	It does not work with negative values
3-parameter gamma	11.008	**Best fit**
Exponential	N/A	It does not work with negative values
2-parameter exponential	19.485	
Smallest extreme value	40.042	
Weibull	N/A	It does not work with negative values
3-parameter Weibull	13.147	
Largest extreme value	12.662	
Logistic	16.775	
loglogistic	N/A	It does not work with negative values
3-parameter loglogistic	26.05	

In addition to the Anderson-Darling value, the P-value also offers good ex for the goodness of fit for the distribution. Below is the probability plot (Fig 3) displaying the 3-parameter gamma.

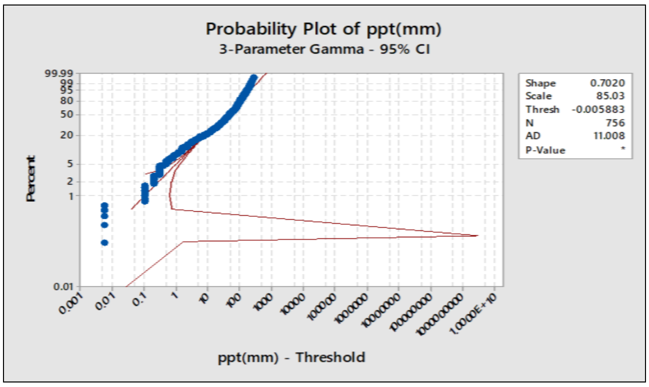

Fig 5: Fig 3: Probability plot for 3-parameter gamma (best fit)

3.3 Multiple Linear Regression

The multiple linear regression was carried out in Minitab using precipitation as the response variable and the rest of the parameters (Windspeed, temperature, runoff, soil cover, vapor pressure, radiation, evapotranspiration, etc.) as the explanatory variables. The figure below (Fig 6) shows the original and predicted values for precipitation using variables: ctual evapotranspiration, climate water deficit, palmer drought severity index, and gross reference evapotranspiration. The series shows a slightly upward or positive trend with an R-Square value of 0.2625 or 26.25%. The graph demonstrates that the independent variables can be used to explain something about the response variable. The positive trend depicts that an increase in any of the values in the explanatory variables leads to an increase in the dependent variable.

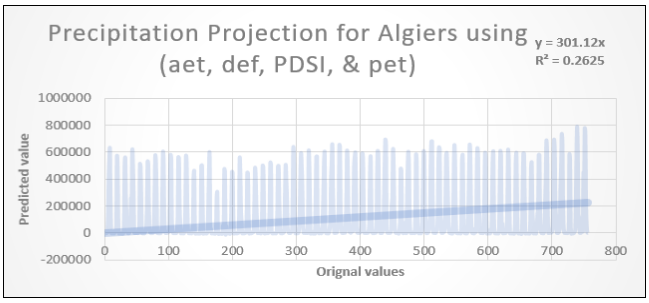

Fig 6: Precipitation Projection for Algiers using (aet, def, PDSI, & pet)

The series graph below shows the original and predicted precipitation values for Algiers using other parameters such as runoff, soil moisture, solar radiation, and maximum temperature. The trend of the series indicates a downwardness or negative pattern. The R-Square value, as seen, is 35.16%. There is a relationship though slightly negligible. As the x values increase, the y values decrease. The connection is disproportional.

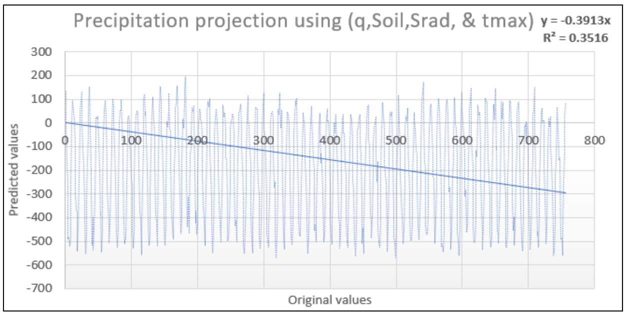

Fig 7: Precipitation projection using (q, Soil, Srad, & tmax)

Below is a time-series graph showing the city's original and predicted rainfall values based on variables such as minimum temperature, vapor pressure, wind speed, and vapor pressure deficit. The trend of the graph is slightly negative. The R-square is found to be 0.2822 or 28.22%. The relationship is negative. Thus, an increase in the independent variables leads to a decrease in the dependent variable. The explanatory variables do not explain much about the dependent variable.

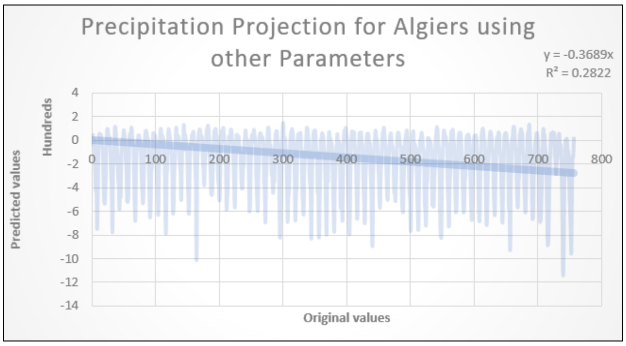

Fig 8: Precipitation Projection for Algiers using other Parameters

3.4 ARIMA Model

In addition to the MLR, the Autoregressive Integrated Moving Average, ARIMA, was also used to determine the goodness of fit and establish a relationship between the explanatory variables and the response variable. The table below (Table 3) summarizes the final estimates of the parameters. The P-value is less than the significance level ($p \leq 0.05$). This suggests the coefficient for the autoregressive term is statistically significant; therefore, the term should be maintained in the model.

Table 3: Final Estimates of Parameters

Type	Coef	SE Coef	T	P
AR 1	-0.4905	0.0326	-15.02	0.000
SAR 12	-1.0769	0.0317	-33.95	0.000
SAR 24	-0.5421	0.0322	-16.83	0.000
Constant	0.112	2.882	0.04	0.969

Table 4 displays the Modified Box-Pierce (Ljung-Box) Chi-Square statistic for the ARIMA model. The p-values are all less than the significance level of 0.05, and significant correlations exist for the autocorrelation function of the residuals. Thus, we can conclude that the model meets the assumption that the residuals are dependent.

Table 4: Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	143.3	227.9	293.1	330.6
DF	8	20	32	44
P-Value	0.000	0.000	0.000	0.000

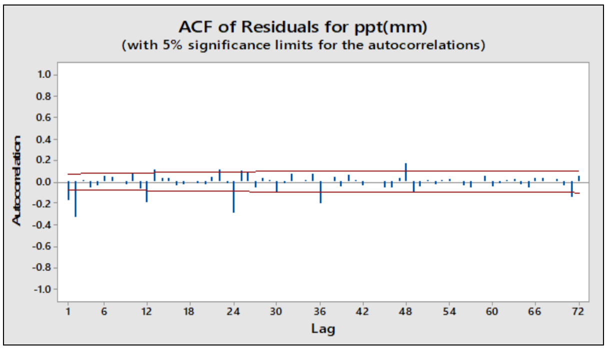

Fig 9: ACF of residuals for ppt(mm) (with 5 % significance limits for the autocorrelations)

4. Conclusion

Climate change is ragging and affects all human and natural entities. Drier regions are becoming desserts while flood takes over wetter places. The hydrological cycle is most affected as the water crisis surges globally. Precipitation patterns are affected and vary from region to region. Understanding the underlying variables that influence the irregularity is required to tackle this crisis. The current report summarizes several methods used to predict precipitation for the North African city of Algiers, Algeria. The ARIMA and the Multiple Linear Regression models were used to analyze the relationship between rainfall and other parameters such as temperature, runoff, vapor pressure, windspeed, evapotranspiration, soil moisture, etc.). Results indicate that solar radiation, soil moisture, runoff, and maximum temperature affect precipitation more, recording an R-square value of 35.16% when using the MLR compared to other explanatory variables. Though the series trend shows a negative pattern, these variables explain more about the precipitation than other variables. For the ARIMA model, using the Modified Box-Pierce (Ljung-Box) Chi-Square statistic, it was observed that p-values are all less than the significance level of 0.05 and significant correlations exist for the autocorrelation function of the residuals. Thus, we can conclude that the model meets the assumption that the residuals are dependent.

It can be concluded that climate change has and will continue to have negative implications on rainfall from area to area. The most significant climatic parameters affecting rainfall in the study area are temperature, runoff, evapotranspiration, soil moisture, and radiation. The 3-parameter distribution was found to be the best fit distribution for analyzing precipitation. The best model for predicting rainfall would be the MLR because it recorded the highest R-square value.

5. References

1. Chattopadhyay, S. Feed forward Artificial Neural Network model to predict the average summer-monsoon rainfall in India. Acta Geophys. 2007; 55:369-382. Doi: https://doi.org/10.2478/s11600-007-0020-8
2. Das R, Mishra J, Mishra S, Pattnaik PK. Design of mathematical model for the prediction of rainfall. Journal of Interdisciplinary Mathematics. 2022; 25(3):587-613.
3. Elouissi, Abdelkader, *et al*. Climate change impact on rainfall Spatio-temporal variability (Macta watershed case, Algeria). Arabian Journal of Geosciences. 2017; 10(22):1-14.
4. Kaushik, Inderjeet, Sabita Madhvi Singh. Seasonal ARIMA model for forecasting of monthly rainfall and temperature. Journal of Environmental Research and Development. 2008; 3(2):506-514.
5. Somvanshi VK, *et al*. Modeling and predicting rainfall using artificial neural network and ARIMA techniques. J. Ind. Geophys. Union. 2006; 10(2):141-151.
6. Swain S, Nandi S, Patel P. Development of an ARIMA model for monthly rainfall forecasting over Khordha district, Odisha, India. Recent Findings in Intelligent Computing Techniques. Springer, Singapore, 2018, 325-331.
7. Valipour M, Banihabib ME, Behbahani SMR. Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of the Dez dam reservoir. Journal of Hydrology. 2013; 476:433-441.
8. Willems, Patrick, *et al*. Climate change impact assessment on urban rainfall extremes and urban drainage: Methods and shortcomings. Atmospheric Research. 2012; 103:106-118.
9. Tarmizi, Aainaa Hatin Ahmad. Climate change and its impact on rainfall. International Journal of Integrated Engineering. 2019; 11(1).