Application of two statistical methods for rainfall network development in northeast of Algeria

[1]M. TOURKI and [2]K. KHANCHOUL
[1] Laboratoire Sols et Développement Durable, Institut des Sciences et Technologies, Centre Universitaire de Mila, B.P. 26, 43000 Mila, Algérie. Email: tourki_mahmoud@yahoo.fr
[2] Laboratoire Sols et Développement Durable, Département de Géologie, Université Badji Mokhtar-Annaba, B.P.12, 23000 Annaba, Algeria. Email: kam.khanchoul@gmail.com

ملخص : يمثل المطر عاملا أساسيا في مفهوم المناخ، حيث أن الدراسات و البحوث التي أنشأت في ميدان تتبع هطول الامطار تواجه دائما عراقيل كثيرة وصعوبات بسبب عدم وجود قياسات و بيانات متصلة و كذلك نقص في كثافة شبكة الرصد المتخصصة في ترقب الأمطار المتهاطلة.

و نظرا لهذه المشكلة، تقدم هذه الدراسة منهجيات مستعملة لوضع شبكة دقيقة لرصد التهاطلات السنوية المتوسطة. يتعلق الأمر بالتنبؤ بالتهاطلات باستعمال الانحدار الخطي المتعدد و شبكات نورون الاصطناعية و هذا اعتبارا من 40 محطة رصد موزعة في الشمال الشرقي للجزائر.

تم استعمال الأساليب الإحصائية، لنمذجة العلاقة بين المتغير التابع (الأمطار السنوية المتوسطة المقاسة) و المتغير المستقيل (ارتفاع المحطة، مسافة المحطة بالنسبة للبحر، الإحداثيات الجغرافية للمحطة).

تبين النتائج التي تم الحصول عليها في هذه الدراسة، أن نموذج شبكات نورون الاصطناعية (MCP) قادر على منح تنبؤات جيدة فيما يخص كميات الأمطار السنوية المتوسطة مقارنة مع نموذج الانحدار الخطي المتعدد. في هذه الحالة، التقدير المقدم من طرف شبكات نورون الاصطناعية هو أقرب إلى القيم الفعلية من التهاطلات المرتبطة بطريقة الانحدار الخطي المتعدد المتغيرات مع خطأ متوسط و ضعيف بنسبة 0.12 و كذلك عامل الكفاءة و عامل ارتباط مرتفع بنسبة 0.90 و 0.95 على التوالي.

الكلمات الأساسية : المتغيرات الفيزيائية، التنبؤ، تساقط الأمطار، النمذجة، الانحدار الغير خطي، شبكة نورون الاصطناعية.

Résumé : La pluie est un paramètre important du climat et les études sur les précipitations sont généralement entravées en raison du manque de données continues ou de la faible densité du réseau pluviométrique, en particulier dans les pays en développement. Compte tenu de ce problème, l'étude présente des méthodologies utilisée pour l'optimisation d'un réseau d'observation des précipitations moyennes annuelles. Il s'agit de la prédiction des précipitations en utilisant la régression linéaire multiple et les réseaux de neurones artificiels, à partir de 40 stations pluviométriques réparties dans le nord-est de l'Algérie. Les deux méthodes statistiques sont utilisées pour modéliser les relations entre une variable dépendante (données pluviométriques) et les variables explicatives (altitude, distance de la mer, les coordonnées latitude et longitude). Les résultats obtenus dans cette étude indiquent que le modèle du réseau de neurone Multi-Couches Perceptron (MCP) est capable de fournir une meilleure représentation des estimations des précipitations en comparaison avec le modèle de régression linéaire multiple. Dans ce cas, l'estimation présentée par les réseaux de neurones artificiels est plus proche des valeurs réelles que les précipitations liées à la régression multivariée avec une erreur quadratique moyenne faible de 0,12 et un facteur d'efficacité et un coefficient de corrélation élevés de 0,90 et 0,95 respectivement.

Mots clés : variables physiques, prévision, précipitation, modélisation, régression non linéaire multiple, réseau de neurone artificiel.

Abstract : Rainfall is an important climatic parameter and the studies on rainfall are commonly hampered due to lack of continuous data or low density of rainfall network, especially in the developing countries. In view of this problem, this study presents the methodologies used for the optimization of a rainfall observation network by predicting annual averaged rainfall. This is the prediction of rainfall using multiple nonlinear regression and artificial neural networks from 40 rainfall stations in the north-eastern Algeria. The two statistical methods are used to model relationships between a dependent variable (rainfall data) and explanatory variables (altitude, distance from the sea, latitude and longitude coordinates). The results obtained in this study indicate that Multi-Layer Perceptron Neural Network (MLP) model is able to provide more or less a better representation of rainfall prediction in comparison with the multiple nonlinear regression model. In this case, forecasting result exhibited by the proposed artificial neural network networks is the closest to actual rainfall values among the multivariate regression taken the low root mean square error of 0.129 and high efficiency factor and coefficient of correlation of respectively 0.88 and 0.94.

Keywords : physical variables, prediction, rainfall, modeling, multiple nonlinear regression, artificial neural network.

1. Introduction

Precipitation plays a significant role in agriculture and it is a major area in climatic studies (Ayoade, 1983). The availability of precipitation data is vital for hydrologic analysis such as design of water resources systems. In fact, studying about precipitation can lead to the identification of its characteristics, the analysis of temporal and spatial variability, and resolving the problems such as floods, droughts, mass wastings, etc.

Often Algerian hydrologists encounter the problem of missing data where the consistency and continuity of rainfall data are very important in statistical analyses such as time series analysis. Both consistency and continuity may be disturbed due to change in observational procedure and incomplete records (missing observations) which may vary in length from one or two days to several months or years (De Silva, 2007). Nevertheless, filling of the gaps provided by inconsistent rainfall data is crucial, and different methods and approaches are available to accomplish task.

It is also true that more rainfall gauges are needed when dealing with rainfall characteristics, soil erosion, and hydrological models. The input of the later models is given by rain gauge measurements so that the accuracy of the output depends essentially on the rain gauge network density configuration (Maheepala et al., 2001). In situ monitoring of rainfall is accurate and reliable, but is usually limited in its usefulness at the regional and global scale because of the high temporal and spatial variability of rainfall. The areal rainfall estimated by rain gauges exhibits a great deal of uncertainty where the rain gauge network is sparse (Collischonn et al., 2008). The development of methods to interpolate climatic data from sparse networks of stations has been a focus of research for much of this century (Thiessen, 1911; Shepard, 1968; Hughes, 1982; Hutchinson and Bischof, 1983; Phillips et al., 1992; Daly, 1994).

In the northeast of Algeria, the existing rainfall gauges in the drainage basins are not adequate to characterize the spatial variation of rainfall, especially in the upper and central parts of the basins. Consequently, studying precipitation is important, not only because it can help up better understanding the precipitation pattern such as its distribution within a watershed, but also it can be applied in forecasting and flood frequency estimation.

However, a mathematical model is not so easy to build because the precipitation is related to enormous factors, which lead to the difficulty to use a specific model to describe the precipitation.

More recently, empirical models such as artificial neural networks (ANNs) are becoming increasingly popular. Their application has gained enormous interest in the hydrology and water resources research for application to a number of hydrological prediction problems (Govindaraju and Rao, 2000).
The main objective of the current study is to identify a better method for the estimation of rainfall observations in ungauged stations by increasing rain gauge network. The specific objectives of the research are to develop and introduce empirical models for missing data estimation at a regional scale, compare and evaluate the estimates obtained from each used method, and to study the suitability of the important factors such as topography, distances from the sea, aerial coverage of each gauge, etc., which can prove to have significant influences on rainfall estimates.

2. Study area

Eastern Algeria is divided into hydrological units called drainage basins (Figure 1). The later are largely nuanced due to their large geographical area, and given the loose of hydroclimatological network, need to rely on data from a higher number of observation stations.

The eastern Algeria is the wettest region of the country and has, therefore, the largest share of water surface. With a mean annual runoff that can exceed 200-300 mm on the Tellian basins, which are opposed to the Western basins, where the dominant semi-arid regions allow lower flow in majority less than 50 mm per year (Mebarki, 2005). The northeastern selected basins have an exoreic flow type (Coastal basins of Constantine, Rhumel-Kebir, Seybouse, and Medjerda). Thus, the mountainous character of the Tell and abundant rainfall provide to the streams an outlet to the Mediterranean Sea. Moreover, the dominant influence of the climatic factor, especially rainfall, rhythms the availability of surface water resources, their scarcity and their recurrent oversupply.

The eastern Algeria is distinguished by a variety of the physical-geographical context, where from south to north and over 250 km as the crow flies, we move from a barren terrain, highly flat and sub-arid to arid climate (e.g. Cherf and Mellegue drainage basins) to humid mountains with beautiful forests of cork oak and by overlooking on several hundred meters, the Mediterranean Sea. The organization of the eastern orographic part of Algeria is affected by a high latitudinal gradient that concerns both topographic elevations and bioclimatic zones (Côte, 1996). This set is geologically completed by a variety of structural units.

Areas of high rainfall (above 900 mm) are located in the northern mountainous along the coast in the northeast. Rainfalls of 600 to 800 mm sweep the rest of the Tellian Atlas, with areas that can get much rainfall in El Kala. In the High Plains, rainfall decreases towards the southern part (350 mm) with relatively less than 300 mm.

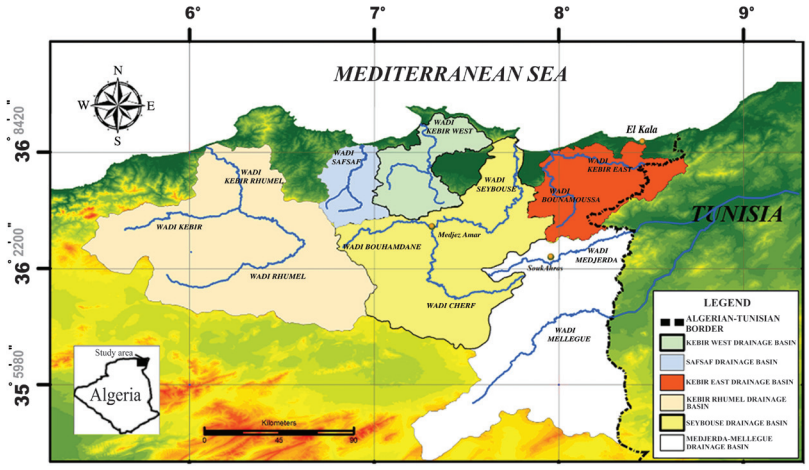


Fig. 1 Location map of the study area

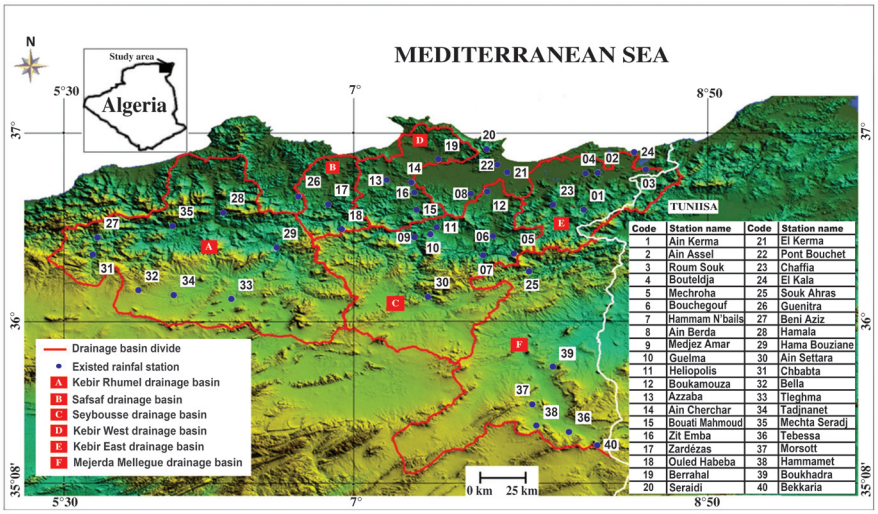
The Seybouse basin (6450 km²) covers a distance of 160 km, with an orientation significantly Southwest Northeast. The union of Wadi Cherf that starts in the High Plains, and Wadi Bouhamdane forms the Wadi Seybouse at Medjez Amar village. Wadi Rhumel that belongs to the Rhumel Kebir basin (8811 km²) starts from an elevation of 1160 m in the southern edge of the Tell. The basin is drained mainly by two main rivers, namely Wadi Rhumel and Wadi Kebir (Figure 1). The Coastal Basins of Constantine are composed mainly of Bounamoussa and Kebir East drainage basins with an area of 3203 km², and the Kebir West, Safsaf (5524 km²). The drainage basins of Medjerda at Souk Ahras gauge station (217 km²) and Mellegue (4575 km²) belong to the great Medjerda basin whose network system starts in Algeria and continues its way in Tunisia to the Mediterranean Sea.

The nature and distribution of the vegetation are generally controlled by the physical and climatic aspect. The major divisions of forest species are determined by climate, especially by the amount of annual rainfall, which in turn depends on relief and soil conditions. Despite its small extension in terms of discontinuity surfaces, the forest canopy extends from the Mediterranean mountainous forests to the sparse forests (Oak) relayed by a dominance of cultures and rangeland of the High Plains (Cherf and Mellegue basins).

3. Materials and methods

3.1 Density of Rain Gauges

For the majority of the catchments, network density of point-measuring rain-gauges is a fundamental tool that provides, with more accuracy, an estimate of rainfall at a point and predicts weather conditions, floods, and droughts. The areal rainfall estimated by the 40 selected rain gauges exhibits a great deal of uncertainty where the rain gauge network is sparse (Figure 2). Based on this limited number of stations, it is difficult to determine how much rain actually falls across the whole north-eastern part of Algeria. Any prediction of rainfall is limited by the difficulties of measuring precipitation in high land areas where the frequency of occurrence of thunderstorms is higher. For these selected stations a regionalization approach developed by statistical analysis is used to estimate point rainfall.



Code	Station name	Code	Station name
1	Ain Kerma	21	El Kerma
2	Ain Assel	22	Pont Bouchet
3	Roum Souk	23	Chaffia
4	Bouteldja	24	El Kala
5	Mechroha	25	Souk Ahras
6	Bouchegouf	26	Guenitra
7	Hammam N'bails	27	Beni Aziz
8	Ain Berda	28	Hamala
9	Medjez Amar	29	Hama Bouziane
10	Guelma	30	Ain Settara
11	Heliopolis	31	Chbabta
12	Boukamouza	32	Bella
13	Azzaba	33	Tleghma
14	Ain Cherchar	34	Tadjnanet
15	Bouati Mahmoud	35	Mechta Seradj
16	Zit Emba	36	Tebessa
17	Zardézas	37	Morsott
18	Ouled Habeba	38	Hammamet
19	Berrahal	39	Boukhadra
20	Seraidi	40	Bekkaria

Fig. 2 Location map of the rainfall gauge stations

3.2 Input data

Annual climate data for the study north-eastern regions of the country are extracted from the National Agency of Hydraulic Resources and National Office of Meteorology. For all the existed rainfall stations, the annual rainfall means are available for a period varying from 28 to 37 complete years, from 1970 to 2010; most of them vary between 28 and 34 years. Most of the stations are found on plains and over coasts, the number of the stations decreases toward the south and over mountainous regions (over 600m).

The prime and available factors dependent upon rainfall are distance from the sea, elevation of the station, and station coordinates such as latitude and longitude (Table 1). A consideration is taken for the former variable because an air stream moving inland becomes progressively drier as moisture is lost in the form of rain and consequently there is a diminution in rainfall unless this is off-set by an increase in elevation. Moreover, this factor is largely associated with the prevailing rain-bearing winds where the directions and trajectories of winds have an important effect on their rain-bearing properties.

Table 1. Data of the topographic parameters at the study rainfall stations

Station codes	Rainfall stations	Rainfall (mm)	Elevation (m)	Distance from the sea (km)	Longitude (UTM in m)	Latitude (UTM in m)
1	Ain Kerma	676.45	235	34	427738	4049894
2	Ain Assel	813.90	35	14	434378	4071936
3	Roum Souk	706.61	150	11	456663	4073346
4	Bouteldja	743.36	20	15	428981	4071269
5	Mechroha	1082.36	748	54	395160	4024340
6	Bouchegouf	545.66	120	45	385320	4034731
7	Hammam N'bails	694.60	478	57	380781	4023484
8	Ain Berda	631.45	85	26	375660	4060140
9	Medjez Amar	600.28	250	54	348794	4034501
10	Guelma	523.81	301	55	356629	4036542
11	Heliopolis	598.57	280	51	359637	4040544
12	Boukamouza	652.45	40	20	383000	4061440
13	Azzaba	626.08	93	18	336620	4068603
14	Ain Cherchar	768.81	34	24	348586	4066927
15	Bouati Mahmoud	669.01	156	39	350375	4051230
16	Zit Emba	544.18	50	26	348630	4064717
17	Zardézas	637.83	180	30	309445	4054897
18	Ouled Habeba	816.00	886	44	315438	4039853
19	Berrahal	676.24	33	18	360820	4080541
20	Seraidi	801.70	840	4	383123	4085980
21	El Kerma	575.10	14	6	388190	4074940
22	Pont Bouchet	629.07	6	5	387890	4077020
23	Chaffia	826.00	170	28.15	413534	4052897
24	El Kala	862.30	13	13.00	450973	4083547
25	Souk Ahras	556.36	580	63.43	401968	4014264
26	Guenitra	660.29	169	31.87	295956	4060010
27	Beni Aziz	663.80	770	40.16	201939	4038562
28	Hamala	809.21	660	55.67	260746	4051356
29	Hama Bouziane	518.47	460	65.26	284692	4029854
30	Ain Settara	289.06	741	89.00	354601	3999410
31	Chbabta	382.38	710	49.20	199480	4028618
32	Bella	360.24	990	73.04	219872	4006677
33	Tleghma	335.79	750	107.04	263439	4000056
34	Tadjnanet	336.14	845	81.67	236401	4003391
35	Mechta Seradj	460.29	350	40.88	237110	4044041
36	Tebessa	313.79	850	164.90	419820	3918624
37	Morsott	266.70	740	143.00	402902	3935553
38	Hammamet	325.26	880	155.00	404580	3922963
39	Boukhadra	299.00	885	125.31	412684	3957421
40	Bekkaria	216.04	950	172.56	432680	3911339

Broadly speaking there is a tendency for rainfall to increase with increasing elevation but the relationship is so broad that little pattern is discernable and a wide range of rainfall occurs at the same elevation (Hounam, 1958). Hence it is generally necessary to include elevation of the rain station as an explanatory variable in the interpolation method. The relationship between rainfall and distance to the sea is also evident as mentioned in some studies mainly during the rainy monthly periods (Hounam, 1958; Hayward and Clarke, 1996; Kieffer Weisse and Bois, 2000). This distance from each rain station to the sea was measured using Google Earth ruler in kilometers.

The global climatology of the vertical gradient of rainfall rate can characterize precipitation-system structure more concretely. According to Hirose and Nakamura (2004) the vertical gradient of rainfall rate at low levels in the coastal upwelling region obstruct the large moist inflow and generate copious orographic downward increase of rain. In general, the vertical gradient for individual systems decreased as rain area increased or as storms developed vertically. Zahar and Laborde (2007) have stated in there study that it exists an elevation gradient that varies only with distance to the sea.

Concerning the rain station coordinates, the longitudes and latitudes in degrees have been converted to Universal Transverse Mercator (UTM) in meters to be used in the statistical analysis and represented in maps using GIS program.

Nevertheless, other approaches have been adopted in different studies; a background of this literature can be cited here. For predicting rainfall, correlation and regression was carried out by Omogbai (2010) to study the rainfall pattern of Northern Nigeria and its relationship with sea surface temperature. The data set (available in the form of monthly and seasonal rainfall totals) of Systat comprising of rainfall data and sea surface temperature were used for this work. In another study Kumar and al. (2007) have used relation between regional rainfall over Orissa and the large scale climate indices like El-Niño Southern Oscillation (ENSO), Equatorial Indian Ocean Oscillation (EQUINOO) and a local climate index of Ocean-Land Temperature as predictor variables to predict the monthly as well as seasonal rainfall using artificial neural networks (ANNs) methodology.

Since global climatic factors like, distance from sea, latitude the Himalayan Mountain, distribution of land and water, surface pressure and wind, upper air circulation and western cyclones are affecting Indian sub-divisional rainfall. The researchers have correlated these independent parameters to predict rainfall (Mahajan and Mazumdar, 2013). Moreover, the Outgoing long wave radiation, global temperature and sunspot numbers have been used as firm predictors of rainfall in other techniques like artificial neural network and Multiple linear regression for Tamil Nadu rainfall prediction. Similarly, An ArcView GIS-based spatial interpolation module has been presented in the island of Crete. The response variable is "Precipitation" and the predictor variables are elevation, longitude, and latitude. The developed module using multiple linear regression to predict precipitation at ungauged locations has performed satisfactorily.

3.3 Statistical techniques

3.3.1 Multiple regression

The multiple regression is to learn more about the relationship between several explanatory or predictor variables and a dependent or criterion variable. In multivariate linear regression, the model specification is defined as a statistic method that is used to model a linear relationship between a dependent variable and one or more explanatory variables (Aksornsingchai and Srinilta, 2011). The Multivariate regression model can be expressed as a linear function shown in equation (1).

$$y = \beta_1 x_1 + \beta_2 x_2 + ...+ \beta_n x_n, \ldots \quad (1)$$

Where y is the value of a dependent, xi is the value of the ith explanatory variable, and βi is an adjustable error coefficient of the ith explanatory variable. Nonlinear regression aims to describe the relationship between a response variable and one or more explanatory variables in a non-linear fashion. When the model function is not linear in the parameters, the sum of squares must be minimized by an iterative procedure.

The regression analysis to derive rainfall equations using key precipitation production parameter is adopted by introducing correlation matrix and multiple regression analyses. The performance is followed using a combination of the four explanatory variables. Throughout this study, the following symbols are used: P= mean annual rainfall, H = elevation, D = distance from the sea, Lat = latitude, Long = longitude. It is strongly advised to view early a scatterplot of data to determine which model to use for each dependent variable and its corresponding explanatory variable. If the variables appear to be related linearly, a linear regression model can be used but in the case that the variables are not linearly related, data transformation might help and/or a mathematical function (eg. nonlinear function) that fits the data to that type of model can be adopted.

The result of the multiple regression analysis is tested for significance within a 95% confidence interval using the ANOVA (Analysis of Variance). This technique is used whenever an alternative procedure is needed for testing hypotheses concerning means when there are several samples. Estimates of the amount of variation are obtained separately and compared using an F-test and conclusions are drawn using the value of F. To test the significance of the observed F, we must state the rejection of the hypothesis of independence when $F > F^*$ (critical value taken from the table of the F distribution). In this case, F^* is equal to 2.64 at degrees of freedom of $\upsilon_1 = c-1, \upsilon_2 = N-c$ and a significance level of 5%, knowing that c and N are respectively the number of used variables and data.

It is possible that the observed correlation between two variables (X and Y) may be in part because of a third or more variables that are related to both of these variables. When this or these confounding variables are also observed, we may be interested in estimating the correlation between X an Y after eliminating the effect of their correlation with the control variables. Thus, the principle of this method is

to measure the correlation between a dependent variable and one explanatory variable when all other variables involved are kept constant ; that is, when the effects of other variables are removed. Indeed, using statistical techniques for the partial correlation of order three, the Tanagra software, working under Excel, initially calculates the correlations of all the variables 2 by 2 from data. Then it updates closer and closer this correlation matrix by introducing the first control variable z_1, then the second z_2, and the third z_3, until it gets the desired third order partial correlation and t-statistic for the significance testing. In this study, the critical t value is taken from the t-Student table (t critical = 2.03) by considering the significance level of 0.05 (alpha) and degrees of freedom of n – k – 2 = 40 – 3 – 2 = 35.

3.3.2 Artificial neural networks

In addition to the multiple regression analysis application, an artificial neural network (ANN) training algorithm, Multi-Layer Perceptron (MLP), is used in the present study. The basic Multi-Layer Perceptron (MLP) model employed in this study possesses a three layer learning network consisting of three distinctive layers, the input layer, where the data are introduced to the ANN, the hidden layer, where data are processed, and the output layer, where the results of ANN are produced (Figure 3).

The MLP is a layered feed-forward network, which means that the units each performed a biased weighted sum of their inputs and pass this activation level through a transfer function to produce their output, and the units are arranged in a layered feed-forward topology (Figure 3). Then, the network's weights and thresholds must be set so as to minimize the prediction error made by the network. This is the role of the training algorithms and the best-known example of a neural network training algorithm is back propagation.

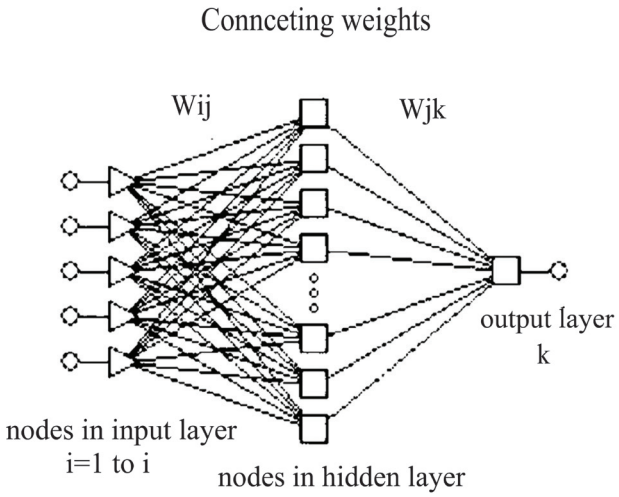


Fig. 3 An exemplary three-layer feedward ANN structure

Back propagation involves two phases: a feed forward phase in which the external input information is propagated forward to a hidden layer node usually through a sigmoid activation function, and a backward phase in which modifications to the connection strengths are made based on the differences between the computed and observed information signals at the output units. The difference or error of the later information signals is minimised by adjusting the weights and biases through some training algorithm, where the error (E) calculated at the output is propagated back to hidden layer and finally to input layer by updating the weights of interconnection. The error (E) is defined as:

$$E = \tfrac{1}{2} \Sigma k \, [d(k) - O(k)]^2 \qquad (2)$$

where d(k) is the observed output at the kth node of the output layer and O(k) is the estimated output at the k^{th} node of the output layer. The same response procedure is repeated for each hidden node (Kuo et al., 2007). The number of nodes in a hidden layer provides the best training results is the initial process of the training procedure.

The input combinations that are tested to estimate annual rainfall values are covering the topographical factors and the target layer are consisting of the unique mean annual rainfall data. Using neural network program, the data sets of 40 patterns are divided into three sets for the purpose of training (60%, 70%, 80%), verification (20%, 15%, and 10%), and testing (20%, 15%, and 10%) to reach the best generalization. The training data set is used to train the neural network by minimizing the error of this data set during the training at different iterations. The cross verification data are used to find the network performance by monitoring the training and guard against overtraining. Then, the test set is used for checking the overall performance of the trained network.

The networks that have been created are inserted into the network set in performance order, so that the last one inserted is the best discovered. The ANNs program has generated these results and statistics to indicate the performance of the best network obtained. During the training process, we have created a large number of networks, and we have retained more than just the best of these.

3.3.3 Performance evaluation criteria

Besides the statistical parameters that are automatically performed using STATISTICA such as coefficient of correlation and Mean Squared Error (MSE) to estimate the accuracy of the proposed methodology, a variety of verification criteria, which could be used for the evaluation and inter-comparison of different models, are proposed by World Meteorological Organisation (WMO) (1975). A suitable one for the present study is chosen,

Root-Mean-Squared Error (RMSE). This can be defined as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Pi - \hat{P}i)^2}{N}} \qquad (3)$$

where $\hat{P}i$ is observed annual rainfall value; is calculated annual rainfall value; and N is number of elements. RMSE gives a quantitative indication for the model error; it measures deviation of the forecasted and/or simulated value from the actual observed value. The ideal value for RMSE is 0.

Also, a percentage forecast error named percentage root mean squared error (PRMSE) in ANN models with three different proportions of ratio is illustrated as follows:

$$PRMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\frac{(Pi - \hat{P}i)^2}{Pi^2}} \times 100 \qquad (4)$$

Other possible mathematical associations are tried with the independent variables. The model efficiency factor EF of observed and predicted values are estimated for different predictions on validation datasets. The best model is selected based on the EF value approaching one. The Nash-Sutcliffe efficiency factor is estimated for all the validation sets using the equation:

$$EF = 1 - \frac{\sum_{i=1}^{n}(Pi - \hat{P}i)^2}{\sum_{i=1}^{n}(Pi - \overline{P})^2} \qquad (5)$$

where $\overline{P}$ is the mean of observed values.

4. Results and discussion

Preliminary univariate analyses of the relationship between annual rainfall, altitude and distance from the sea with the 40 rainfall gauges have shown that the relationship between rainfall and elevation is negatively moderate (r = -0.475) and it increases to -0.739 with rainfall and distance (Table 2a). After screening the data graphically (e.g. by a scatterplot) in order to determine how the explanatory and dependent variables are related (linearly, exponentially, etc.), we have discovered that the data resemble an exponential function between the dependent variable and each of the explanatory ones.

Applying multiple nonlinear regression for the rainfall data set using XLSTAT software and choosing the best function among the preprogrammed functions, the results show an evident relationship between the rainfall and the topographical parameters (r = 0.84). The regression parameters related to this relation are represented by the following equation:

P = 719974+0.55**H*+4.86**D*+1.75E-03**Lat*-0.36**Long*-3.11E-04**H*^2-5.09E-02**D*^2-8.18E-10**Lat*^2+4.5E-08**Long*^2 (6)

By making logarithmic transformation of the five used variables, the correlation matrix (table 2b) has shown a slight increase of the association between altitude, distance to the sea, and longitude. The purpose of the logarithmic transformation is to remove that systematic change in spread, achieving approximate «homoscedasticity». Thus, one reason researchers (although not the only reason) utilize data transformations is improving the normality of variables.

The aforementioned procedure of nonlinear multiple regression for the logarithmic transformed data was used throughout the study variables, and the investigation has revealed that the relationship is quite significant, with a coefficient of correlation equals to 0.92. When another nonlinear multivariate regression is used by taking 70% of the data for training and 30% for the model validation, the coefficient of correlation becomes equal to 0.94. The equation related to this analysis can be written as follows:

P = 88051.39-0.47**H*+1.75**D*+28.53**Lat*-11612.53**Long*+6.50E-02**H*^2-0.31**D*^2-1.09**Lat*^2+382.09**Long*^2 (7)

The analyses have shown that relationship between annual rainfall and altitude is quite different for the two regional climatic groups of gauges (humid-subhumid and semi-arid regions). In drier twelve gauges, located in south-western and south-eastern semi arid climates, there is evidence, but not consistently, that rainfall is negatively related to altitude whilst in the eastern wetter regions, annual rainfall generally has increased positively with altitude. Hence, in the semi-arid regions, higher elevations do not always receive more rainfall; some locations with low elevation receive more rainfall where the rainfall is chiefly due to a shelter effect of the elevation influence and continentality. The relationship reflects low precipitations at higher elevations, particularly on the leeward plateaus, depending on exposure of the station to the prevailing wind, orientation and slope of the hills and mountains. For instance, the positive correlation is not such good fit in the humid and subhumid areas because much of the variance remained unexplained in those areas.

The interaction between rainfall and distance from the sea for the study gauges is less evident with the logarithm of both values (Table 2b). This relationship becomes less marked and more spatially variable in humid-subhumid regions. Decreases in rainfall with distance are greatest in the drier areas, and therefore are dissimilar in magnitude in the two groups of gauges.

Table 2. Correlation matrix related to rainfall, altitude, distance, and geographic coordinates

a- Result of non transformed data

Parameters	*P*	*H*	*D*	*Lat*	*Long*
P	1				
H	-0,475	1			
D	-0,739	0,730	1		
Lat	0,208	-0,315	0,024	1	
Long	0,737	-0,734	-0,988	-0,051	1

b- Result of natural logarithm transformed data

Parameters	*P*	*H*	*D*	*Lat*	*Long*
P	1				
H	-0,480	1			
D	-0,682	0,746	1		
Lat	0,163	-0,376	-0,210	1	
Long	0,820	-0,662	-0,860	-0,036	1

From the results, it may be noted that the value of Root-Mean-Squared Error (RMSE = 0.176) and PRMSE (2.40%) are considered not too high compared to the values of the non transformed data with PRMSE and EF equal respectively to 18.63 and 0.73. In fact, the former multivariate model according to its EF value of 0.88 has provided a more satisfactory performance in estimating rainfall (Table 3).

The scatter plots of observed (measured) rainfalls and predicted rainfalls of all the gauges are given in Figure 4. As it can be seen from the figure, the observed rainfalls are not very close to the simulated ones; more residuals are observed outside the 95% confidence band. So it proved that the multiple regression model has not increased in accuracy, which suggests that its contribution could be important at a drainage basin level, but the impact of the physical parameters is insignificant at a regional level, due to regional fluctuations in average rainfall.

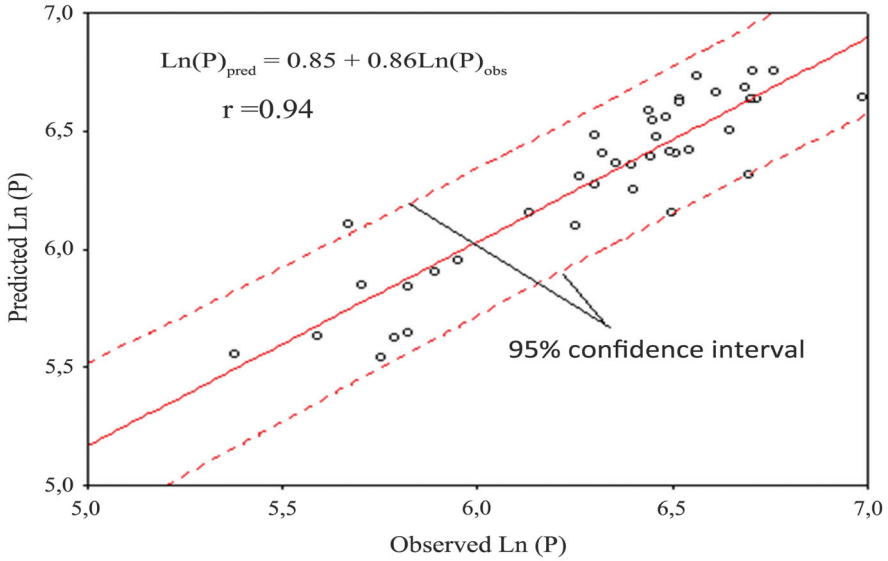


Fig. 4 Predicted against observed rainfalls using multiple nonlinear regression

Therefore, it is not feasible to use the model to predict rainfall under such conditions. Nevertheless, we can conclude from the used ANOVA (analysis of variance) that $F_{4,35}$, having the value of 29.16, is too high to have been merely the result of sampling error (Table 3). The sample is significant evidence for detecting a relationship between the five variables.

Table 3. Statistical parameters in multiple nonlinear regression model

Parameters	Values
RMSE	0.176
PRMSE (%)	2.40
Coefficient of correlation (r)	0.94
EF	0.88
ANOVA	
$df_1 = C-1$	4
$df_2 = N-C$	35
Fcalculated	29.16
Fcritical at $\alpha = 5\%$	2.64

df_1 and df_2 : degrees of freedom ; Fcritical at α= 5%: value taken from F-distribution table at significance level (α) of 5%.

In this study we have used multiple regression to predict a single precipitation variable from four explanatory variables. However, we may suspect the relationship between two variables to be influenced by other variables, and to provide explanation to this correlational relationship we have used partial correlation. By considering the univariate coefficients of correlation shown in table 2b, we note that most of these coefficients are higher than the partial coefficients of correlation, except for the precipitation versus latitude (Table 4). It appears before testing that the data are consistent with lack of connection between the explanatory variables and those of dependent ones once we have retrieved the information provided by the control variables. The t-statistic for the significance testing in $r_{12.345}$ is higher than the critical t values of 2.03, which means that there is a significant correlation between precipitation and altitude, even with removing the three variables (D, Lat and Long), even though there is opposite interaction between the simple correlation and partial correlation. In addition, the p-value is lower than 0.05 for detecting a rejection of an interaction between the explanatory and dependent variables.

In $r_{15.234}$, if we control for the three variables H, D, latitude, we see there's still a connection between precipitation and both longitude. It is the same thing for the partial correlation $r_{14.235}$ when holding constant the variables altitude, distance to the sea, longitude, where latitude does not loose from its connection with precipitation. Since t-statistic is greater than critical t at significance level of 0.05, and p-value is less than alpha, we can't reject the hypothesis that both longitude and latitude might influence the rainfall distribution. In contrary, when applying partial correlation for the substantial significant interaction between precipitation and distance to the sea, we notice that the partial correlation is greatly smaller than the simple correlation. Based on the testing significance (t < critical t) and p-values which is greater than alpha (Table 4), it is clear to suggest that after eliminating the contribution of the control variables, we find the moderate relation between precipitation and distance to the sea vanishes. Moreover, it appears that the direction of the relation changes as well, suggesting that after removing the contribution of the three variables (altitude, latitude and longitude), areas receiving more rainfall in fact are inaccurately diagnosed with distance to the sea.

Table 4. Partial coefficients of correlation and significance tests

Partial correlation designation	Target	Input	Control variables	r	t-test	p-value
$r_{12.345}$	*P*	*H*	*D, Lat, Long*	0.361	2.291	0.028
$r_{13.245}$	*P*	*D*	*H, Lat, Long*	0.20	1.213	0.233
$r_{14.235}$	*P*	*Lat*	*D, H, Long*	0.528	3.678	0.0008
$r_{15.234}$	*P*	*Long*	*H, D, Lat*	0.752	6.750	0.0000

From a series of ANN exercises, the training ratio of 60:20:20 and 70:15:15 are found the best division for the models. Table 5 summarizes the networks performance during the training, validation, and testing stages for the three chosen ratios. In this table, the retained models possess low error standard deviation and high coefficients of correlation.

Table 5. Performance of MLP network stages and forecast errors in ANN models

Ratio	Statistical parameters	Training	Validation	Testing	Nodes	RMSE	PRMSE (%)	r	EF
60:20:20	Data ratio Error S.D. S.D.ratio Correlation	0.363 0.120 0.329 0.944	0.403 0.111 0.276 0.963	0.449 0.151 0.336 0.942	10	0.129	2.074	0.943	0.88
70:15:15	Data ratio Error S.D. S.D.ratio Correlation	0.373 0.128 0.343 0.939	0.363 0.064 0.176 0.985	0.422 0.179 0.425 0.906	7	0.133	2.20	0.938	0.88
80:10:10	Data ratio Error S.D. S.D.ratio Correlation	0.375 0.142 0.378 0.927	0.495 0.061 0.123 0.992	0.404 0.181 0.449 0.970	8	0.137	2.21	0.933	0.87

Standard deviation ratio (S.D. Ratio) is the division of the error S.D. by the data S.D.

In addition, the configuration with 4 input nodes, 7 and 10 hidden nodes and unique output for the previous ratios have provided the best performance during the training stage of 820 iterations (Figure 5), with the highest coefficient of correlation (0.94) and the lowest RMSE and PRMSE (Table 5). The comparison of the Nash-Sutcliffe efficiency factor in assessing the performance of a model between the ratios has given a better factor for the proposed models, and which are equal to 0.88.

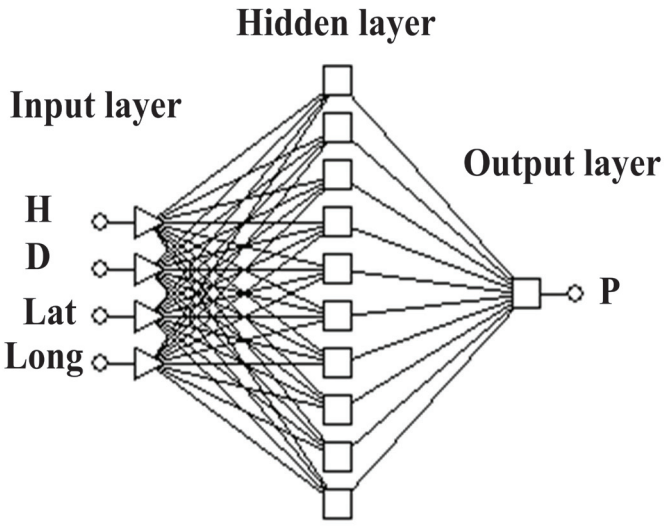


Fig. 5 Neural network illustration using MLP technique

The logarithmic presentation is also given in figure 6 for the 60:20:20 ratio in order to see clearly the model performances for rainfalls of different heights. It can be seen from the fit line equations (assume that the equation is y = ax + b) that the a and b coefficients for the MLP model are respectively closer to the 1 and 0 than the nonlinear multiple regression. This confirms the RMSE and PRMSE statistics evaluated in table 3. During the testing, the MLP model has produced the closest values to the observed mean annual rainfalls by its highest coefficient of correlation (0.94).

The significantly less underestimations of the high logarithmic rainfall values (> 6.50 mm, mostly) for the MLP model are obviously seen from the scatterplots. Compared to MNLR model (multiple nonlinear regression) shown in figure 4, there is less overestimation of low values (< 6.00 mm) for the MLP model. It can be considered that the MLP model has given good prediction in both the training and testing phases where all data lie in line of almost perfect agreement.

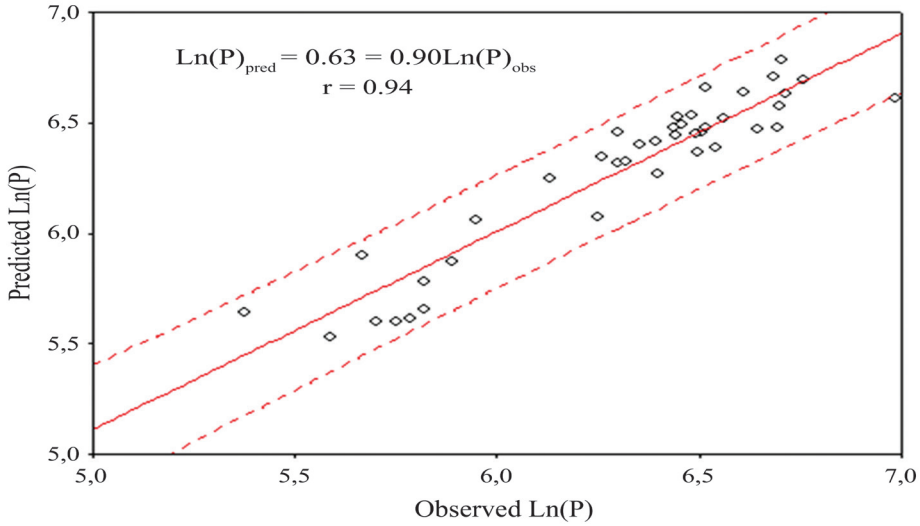


Fig. 6 Comparison of predicted and observed rainfalls using MLP model

When the MLP model is compared with the MNLR method in respect of calculated RMSE, PRMSE. It can be seen from the previous discussion that both MLP at ratio 60:20:20 outperforms the MNLR model in terms of errors. The rainfall estimates of the two models are represented in figure 7 in the form of curves. It is obviously seen from the curves that MLP is closer to the corresponding observed rainfall values. Moreover, the MLP model seems to have fairly better performance than the MNLR from the error viewpoint.

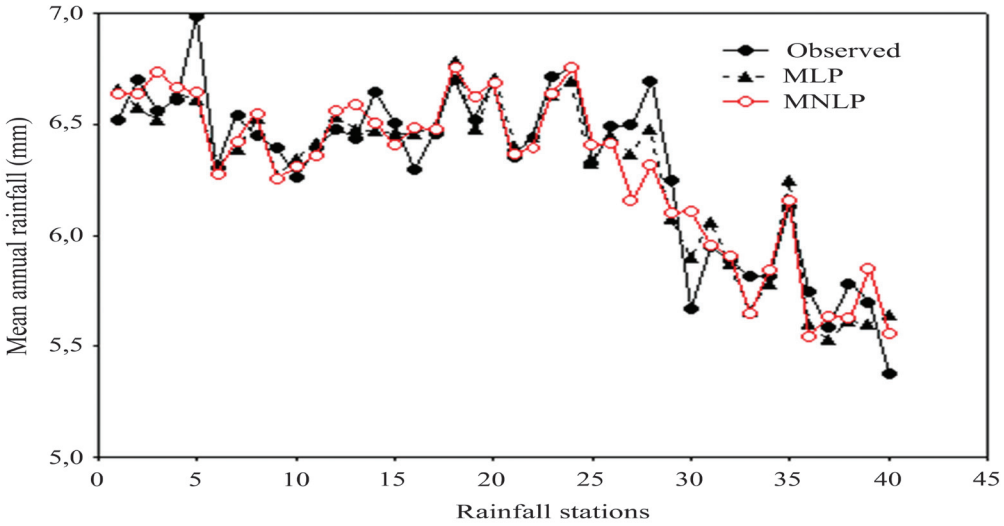


Fig. 7 Mean annual rainfall estimated by MLR and MLP models for the 40 rainfall stations

Further inspection of the ANN model is performed by adding mean annual rainfall data of 13 rain gauge stations distributed over the study area. The resulting ANN-based estimation model, obtained using the 13 stations in analysis, has revealed that the differences between the observed and predicted rainfall values are generally relatively small. The statistical calculations have given a stronger correlation (r = 0.93) between predicted and actual rainfall values and low error (RMSE = 0.164).

The superiority of artificial neural networks over a conventional method in the reviewed prediction study can be attributed to their capability to capture the nonlinear dynamics and generalize the structure of the whole data set (Celikoglu and Cigizoglu, 2007). In fact, the relationships are not linear that any hydrologist, who has at least one neuron, knows that the neural networks have the advantage of not having to worry about this non-linearity. Obviously, using the artificial neural networks such as MLP for modelling rainfall estimation is more reliable than the classical method in the weir studied herein.

5. Rain gauges increase

In the northeastern area, the existing rainfall stations are not sufficient to characterize the spatial variation of rainfall because it often varies spatially with drainage basin topography. For example areas in higher elevations generally receive more rainfall than areas in lower elevations within the drainage basin. To fulfil that criteria twenty six dummy stations are manually distributed over two selected drainage basins, Seybouse and Kebir Rhumel, to reach a rather uniform coverage (Figure 8).

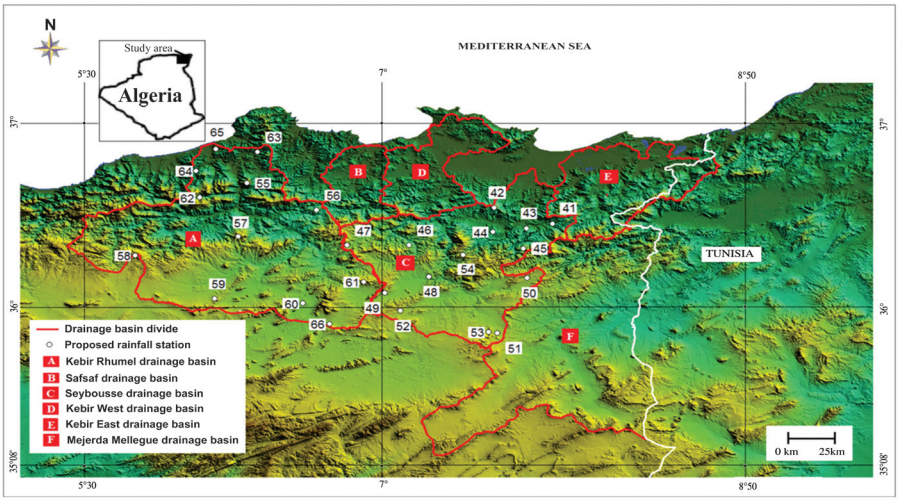


Fig. 8 Map of the proposed rainfall stations (symbolized by white squares with station codes)

The estimation of the mean annual rainfall in the proposed rainfall gauge stations is done using the MLP model and MNLR, and the results are presented in Table 6. From this table, we have taken a closer look at the five variables for the selected stations by analysing the data using a multiple regression application.

Table 6. Suggested rainfall stations with ANN predicted rainfall values.

Station codes	Elevation (m)	Distance to the sea (km)	Longitude (UTM in m)	Latitude (UTM in m)	Predicted rainfall (MLP)	Predicted Rainfall (MNLR)
41	533	39.38	392930	4039550	860.05	855.90
42	816	50.96	371790	4044690	894.39	853.25
43	351	57.81	386170	4032070	594.05	567.84
44	440	65.39	371010	4030240	555.76	539.58
45	978	70.57	384700	4019900	636.95	677.05
46	704	63.53	332810	4022830	558.63	601.81
47	854	67.60	304770	4023540	536.30	575.23
48	825	92.96	341600	4003630	373.47	440.79
49	891	91.16	321300	3994330	347.29	440.06
50	868	88.00	386040	4002330	424.66	507.18
51	785	126.00	371980	3968890	266.05	314.16
52	867	103.77	328250	3983278	304.67	381.72
53	908	124.66	367985	3969849	280.86	336.05
54	651	87.93	357315	4016669	422.43	446.66
55	1120	18.21	260500	4062180	1071.06	1012.49
56	710	46.66	291610	4045170	776.96	716.93
57	1094	69.82	255800	4029980	531.65	520.75
58	1236	62.00	208650	4020090	454.43	437.29
59	954	94.27	243860	3992720	311.75	323.08
60	797	104.90	283780	3988740	289.10	320.25
61	1086	80.93	311870	4001040	417.83	530.83
62	647	30.36	238860	4054160	864.73	693.48
63	113	25.14	265870	4080930	587.78	600.76
64	186	13.45	237890	4070080	650.81	558.75
65	93	3.28	247060	4083100	417.76	246.86
66	869	116.96	295710	3976050	282.42	303.59

The results have been perfect where the experiment has given a high coefficient of correlation. By using logarithm transformations of the explanatory and dependent variables, the relationship has been improved by an increase of the coefficient of correlation to 0.94. Again, these results show the good performance of the MLP neural network and in a way the multiple nonlinear regression in the assessment of rainfalls.

6. Conclusion

This study is prompted by the need for rainfall information, especially for hydrometeorological purposes, over unsettled areas where rain gauge density is quite inadequate. Four factors recognized as affecting the distribution of mean annual rainfall in northeast of Algeria are investigated quantitatively at 40 rain gauges, using multiple nonlinear regression and artificial neural network analysis. The factors are the elevation of the rain gauging station, its distance to the sea, its latitude and its longitude.

For the prediction of mean annual rainfall in suggested gauges, two different models are used, MNLR and MLP. In order to increase our confidence in the use of one of the models, it has been necessary to compare their outputs with root mean squared error, efficiency factor and coefficient of correlation statistics. However, to date, there have been no studies of comparison of the MNLR outputs with those of ANNs methods in Algeria.

From the results of this study, the ANN configuration established shows somehow the highest statistical performance in the rainfall prediction when the four physical parameters are used as associated input variables in the network, while the MNLR has shown also its significance performance in the rainfall estimation. Neverheless, the performance evaluation of the neural network model (eg. MLP) remains statistically more efficient compared to the classical regression regardless of the input sets. As a fact, the ANN model configuration can be suggested as a potential tool for modeling the mean annual rainfall of the northeast of Algeria at ratio 60:20:20 or 70:15:15 when rainfall data are not available.

A similar analysis to be carried out on rainfall prediction using other factors (slope, exposure to rain bearing winds, wind direction and convergence) which are not included in the network inputs could explain more this relationship between rainfall and physical principles. Achieving this could certainly improve the reliability of the whole methodology, by diminishing the variability of the starting dataset.

Bibliographical references

Aksornsingchai P. and Srinilta C. 2011. Statistical downscaling for rainfall and temperature prediction in Thailand. Proceedings of the International MultiConference of Engineers and Computer Scientists, 1: 356-361.

Ayoade J.O. 1983. Introduction to climatology for the Tropics. John Wiley and Sons, New York.

Celikoglu H.B. and Cigizoglu H.K. 2007. Public transportation trip flow modeling with generalized regression neural networks. Advances in Engineering Software **38**: 71-79.

Collischonn B., Collischonn W., Tucci C.E.M. 2008. Daily hydrological modeling in the Amazon basin using TRMM rainfall estimates. Journal of Hydrology **360 (1-4)**: 207-216.

Côte, M. 1996. L'Algérie, espace et société. Masson - Armand Colin, Paris.

Daly C. 1994. A statistical-topographic model for mapping climatological precipitation over mountainous terrain. Journal of Applied Meteorology **33**: 140-158.

De Silva R.P., Dayawansa1 N.D.K., Ratnasiri M.D. 2007. A comparison of methods used in estimating missing rainfall data. Journal of Agricultural Science **3(2)**: 101-108.

Govindaraju R.S. and Rao A.R. 2000. Neural Networks in Hydrology. Kluwer Academic Publishers, Netherlands.

Hayward D. and Clarke R.T. 1996. Relationship between rainfall, altitude and distance from the sea in the Freetown Peninsula, Sierra Leone. Hydro logical Sciences Journal **41(3)**: 377-383.

Hirose M.and Nakamura K. 2004. Spatial and temporal variation of vertical profiles of rainfall rate observed by TRMM precipitation radar. Journal of Climate **17**: 3378–3397.

Hounam C.E. 1958. Estimation of average annual rainfall over the port Phillip region of Victoria. Australian Meteorological Magazine **21**: 1-30.

Hughes D.A. 1982. The relationship between mean annual rainfall and physiographic variables applied to a coastal region of southern Africa. South African Geographical Journal **64**: 41-50.

Humbert J., Mahr, N., Siefert N. 1997. Quantification spatiale des précipitations du bassin Rhin- Meuse. Secteur oriental, Période 1971-1990 (Precipitation spatial mapping of the Rhin-Meuse basin. Eastern zone, period 1971-1990). Final Report, Agence de l'Eau Rhin Meuse, CEREG, Strasbourg.

Hutchinson M.F. and Bischof R.J. 1983. A new method for estimating mean seasonal and annual rainfall for the Hunter Valley, New South Wales. Australian Meteorological Magazine **31**: 179-184.

Kieffer Weisse A. and Bois P.H. 2000. Topographic Effects on Statistical Characteristics of Heavy Rainfall and Mapping in the French Alps. Journal of Applied Meteorology **40**: 720-740.

Kumar D., Reddy M. J. and Maity R. 2007. Regional rainfall forecasting using large scale climate teleconnections and artificial intelligence techniques. Journal of Intelligent Systems **16** (4), 307-322.

Kuo J.T., Hsieh M.H., Lung W.S., She N. 2007. Using artificial neural network for reservoir entrophication prediction. Ecological Modelling **200**: 171-177.

Llasat M. and Puiggerver M. 1992. Pluies extrêmes en Catalogne, influence orographique et caracteristiques synoptiques. Hydrologie Continentale **7**: 99-115.

Mahajan S., Mazumdar H. 2013 .Rainfall prediction using neural net based frequency analysis approach. International Journal of Computer Applications **84**(9): 7-11.

Maheepala U.K., Taolkyi, A.K., Perera B.J. 2001. Hydrological data monitoring for urban stormwater drainage systems. Journal of Hydrology **245**: 32-47.

Mebarki A. 2005. Hydrologie des bassins de l'est algérien : ressources en eau, aménagement et environnement. Thèse de Doctorat d'Etat, Université Mentouri de Constantine, Constantine, Algérie.

Naoum S.and Tsanis I.K. 2003. Estimating rainfall at ungauged locations using topographical and geographical features by means of multiple linear regression. 8th International Conference on Environmental Science and Technology, Lemnos Island, Greece, 604-607.

Phillips D.L., Dolph J., Marks D. 1992. A comparison of geostatistical procedures for spatial analysis of precipitation in mountainous terrain. Agricultural and Forest Meteorology **58**: 119-141.

Omogbai, B.E. 2010. Prediction of Northern Nigeria rainfall using sea surface temperature. Journal of Human Ecology, **32**(2): 127-133.

Selvaraj R. S. and Aditya R. 2011. Statistical method of predicting the northeast rainfall of Tamil Nadu. Universal Journal of Environmental Research and Technology, **1**(4): 557-559.

Shepard D.L. 1968. A two dimensional interpolation function for irregularly spaced data. Proceedings of 23rd National Conference, Association for Computing Machinery, ACM, Washington, 517- 524.

Thiessen A.H. 1911. Precipitation averages for large areas. Mon. Weather Review **39**: 1082-1084.

World Meteorological Organisation, 1975. Intercomparison of conceptual models used in operational hydrological forecasting, W.M.O., Technical series. Water Resources Research **27**(9): 2415-2450.

Zahar Y. et Laborde J.P. 2007. Modélisation statistique et synthèse cartographique des précipitations journalières extrêmes de Tunisie. Revue des sciences de l'eau **20** (4): 409-424.