Interpolation the Missing Data of Air Temperature by Using Artificial Intelligence for Selected Iraqi Weather Stations

Dr. Osama T. AL-Taai
Department of Atmospheric Sciences, College of Science
AL-Mustansiriyah University
Baghdad-Iraq

Abstract

In this work, a branch of artificial intelligence was employed to predict the monthly mean of daily air temperature for different target station using the available data of neighboring stations, which were used as the reference stations. The daily air temperature data, collected by the Iraqi Meteorological Office (IMO) for 14 stations, which cover different Iraqi provinces, were used. The long term air temperature data covers the period between 1993 and 2008. These data were classified in parts according to the correlation coefficients relating them. The reference stations data, as on input layer of the neural network and the hidden layers and neurons were defined; the monthly mean of air temperature for the target station was utilized as an output layer of the neural network. Multi-Layer Perceptron's learning algorithm was applied in present work. The hidden layer and output layer of the network included Sigmoid as an activation function. Finally the interpolated data by (ANN) model were compared with measured data shows very good agreement with Correlation Coefficients (r) ranges between 0.9980 and 0.9767 also Root Mean Square Error (RMSE) ranges from 0.629 °C to 2.221 °C, Mean Percentage Error (MPE) ranges between 0.264 °C and 3.64 °C and Mean Absolute Error (MAE) ranges between 0.367 °C and 1.62 °C for Mosul and Najaf stations respectively.

Keyword:- perception learning , correlation coefficients , MPE, MAE.

استكمال البيانات المفقودة لدرجة حرارة الهواء باستخدام الذكاء الاصطناعي لمحطات طقس عراقية مختارة

د. أسامة طارق الطائي

قسم علوم الجو - كلية العلوم - الجامعة المستنصرية - بغداد - العراق.

الخلاصة

في هذا العمل استخدم فرع الذكاء الاصطناعي للتنبؤ بالمعدل الشهري لدرجة حرارة الهواء اليومية لمحطة مستهدفة مختلفة باستخدام البيانات المتاحة من المحطات المجاورة والتي كانت تستخدم كمحطات مرجعية. إن بيانات درجة حرارة الهواء اليومية المستخدمة تم الحصول عليها من دائرة الأنواء الجوية العراقية (IMO) ولأربعة عشر محطة والتي تغطي مختلف المحافظات العراقية. وقد أخذت مدة طويلة من بيانات درجة حرارة الهواء تغطي الفترة للسنوات بين 1993 و 2008. ولقد تم تصنيف هذه البيانات إلى أجزاء وفقا لمعاملات الارتباط المتعلقة بها. وحددت مراكز البيانات المرجعية على طبقة المدخلات للشبكة العصبية والطبقات المخفية والخلايا العصبية المعرفة، واستخدم المعدل الشهري لدرجة حرارة الهواء للمحطة المستهدفة باعتبارها طبقة المخرجات من الشبكة العصبية وتم تطبيق العمل على خوارزمية. الطبقة المخفية وطبقة الإخراج تضمنت خوارزمية من نوع Sigmoid كدالة تنشيط وأخيرا تمت مقارنة البيانات المقاسة مع بيانات نموذج

(ANN) حيث دلت على توافق جيد جدا لمعاملات الارتباط (r) والتي تراوحت بين 0.9980 و0.9767 وكذلك الجذر التربيعي لمعدل الخطأ (RMSE) والتي تراوحت بين 0.629 °C و 2.221 °C ومعدل الخطأ النسبي (MPE) والتي تراوحت بين 0.264 °C و 3.64 °C ومعدل الخطأ المطلق (MAE) والتي تراوحت بين 0.367 °C و 1.62 °C لمحطتي الموصل والنجف على التوالي.

Introduction

The climate is one of the most important systems governing the energy efficiency of building, especially the air temperature; it's not possible to design on energy efficient building without having the air temperature. Bulut et al. [1] reported that the design of heating, ventilating and air conditioning, solar energy systems and the calculation of heating and cooling depend on solar radiation, air temperature, relative humidity, wind speed and direction. In many environmental data sets, there are missing data (also called missing observations or missing values). This can occur for a number of reasons, such as instrument failure, (human) observer failure, weather conditions, etc. in the file with the data, missing data should not just be left blank. If one did so, it would be impossible for anyone looking at the data file to know whether the data are truly missing, or the data file was damaged during transmission over the internet, or someone made a typo.

Traditional interpolation methods to estimate weather data include inverse distance interpolations and regression models. Inverse-distance is a weighting interpolation method. The number of neighbors necessary in the weighting function is important in terms of reducing computation time while maintaining a smooth surface. Dodson and Marks [2] have suggested that with inverse-squared-distance interpolation using eight nearest neighbors is reasonable. Robeson (1993, 1995) [3, 4] investigated three methods of spatially interpolating temperature anomaly data. He found that the inverse-distance method gave about the same results as triangulated surface patches. In order to consider the elevation effects on climate, Gradient plus Inverse-Distance-Squared (GIDS) interpolation technique was derived (Nalder and Wein, price et al.) [5] From the inverse-distance-square method. Price et al. [6] used gradient plus was suggested that this method is attractively simple and appears to give results adequate for modeling long term forest ecosystem responses to climate in relatively flat terrain.

Regression has been used successfully in weather data estimation (Ollinger et al.) [7] Bolstad et al. [8] used regression to predict air temperature and compared their approach with local lapse models or Kriging methods. They stated that the regression approach provided a more accurate estimate of station temperature. Christine et al., (1998) [9] used a regression technique to predict the monthly precipitation, monthly averaged maximum and minimum temperature, and monthly averaged sunshine hours and compared the regression approach with a modified inverse-distance-square interpolation. They reported that the prediction accuracy did not differ between these two methods. Artificial neural networks (ANNs) provide a potential alternative to estimating weather data. (ANNs) are computer models that mimic the structure and functioning of the human brain [10]. (ANNs) can determine the relationships among the independent variables to predict or estimate dependent variables. Back propagation (ANNs) is known for their ability to generalize well for a wide variety of problems and is well suited for prediction applications. Unlike statistical methods, (ANN) models do not make dependency assumptions among input variables and solve multivariate problem with nonlinear relationship among input variables. This technique has been used in a wide range of applications, such as classification, pattern recognition, automatic control and function approximation [11, 12]. Han and Felker (1997) [13] implemented an (ANN) to estimate daily soil water evaporation from average relative air humidity, air temperature, wind speed, and soil water content in a cactus field study. They found that the (ANN) achieved a good agreement between predicted and measured values.

They concluded that the (ANN) technique appeared to be an improvement over the multi-linear regression technique for estimating soil evaporation. Cook and Wolfe (1991) [14] developed a neural network to predict average air temperatures for a single location. In their study, the monthly average of daily maximum temperatures for three months in advance was predicted. Bruton et al., (2000) [15] developed (ANN) models for estimating daily pan evaporation. The results were compared with those of multi-linear regression and Priestly-Taylor model and they found that the (ANN) model provided the highest accuracy.

The aim of this research was to develop (ANN) models to estimate the missing data of monthly mean air temperature for fourteen meteorological stations in Iraq using the monthly mean air temperature of number of weather stations which required as input.

The air temperature is the one of most important meteorological parameters; therefore it's recorded regularly in weather station worldwide. Guan et al., (2007) [16] showed that the weather forecasting is essential for the resource for various subject, such as urban planning, agriculture production. The importance of the temperature parameter return to use it as an input to a variety of spatially distributed hydrological and the ecological models, which is use the air temperature to derive processes such as evapotranspiration, snow melt, soil decomposition and plant productivity. Several methods employed by many researcher in order to estimate the missed data of a certain station using neighbor stations, missing data present a problem in many fields, including meteorology. The data can be missing at random, in recurring patterns, or in layer sections, incomplete data set can be lead to misleading conclusions as demonstrated by [17].

Multi-Layer Perceptron’s Architecture Methodology

Multi-Layer Perceptron’s (MLP) is used on the present work using by commercial software Forecaster-XL-Version 1.6 [18]. The biological nervous system, through much of the biological detail is neglected. (MLP) as shown in figure below is massively parallel system composed of many processing elements connected by links of variable weights, the network consists of layers of parallel processing elements, called neurons, with each layer being fully connected to processing layer by inter connection strengths, or weights. Four layer networks consisting of layers i, j, with inter connection weights W_{ij}.

Initial estimated weight values are progressively corrected during a training process that compare predicted out puts to known outputs, and back propagate any output to determine the appropriate weight adjustments necessary to minimize the errors [19].

$$y_i = f\left(\sum_{j=1}^{n} w_{ij}\, x_j + b_i\right) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.. \quad (1)$$

where: (y_i) is the output signals, (f) activation function (squashing), (w_{ij}) synaptic weights, (x_j) input signals and (b_i) is the bias, (i) number of neurons of layer, (j) number of elements in input vector. Figure 1. Show the Multi-Layer Perceptron’s architecture and figure 2. Illustrates the (ANN) architecture used in the interpolation process.

Data Set and Study Area

In present work, daily air temperature collected by Iraqi Meteorological Office (IMO) at 14 meteorological stations such as (Sinjar, Mosul, Makhmur, Tikrit, Khanaqin, Al-Qa'im, Baghdad, Ali Al-Garbi, Al-Hayy, Najaf, Diwaniyah, Samawah, Nasiriyah and Basrah) wear used the long term air temperature data cover the period from 1993 to 2008. The monthly mean of daily air temperature categorized in two categories, the first category was the data which covers the year between 1993 and 2008 was considered for training and neural network building. The second category covered the years 2007 – 2008 was utilized for testing procedure and considered as missed data set. The mean air temperature of all stations is presented in table 1. As seen in the table, the mean air temperature varies mainly from 21.16 ^{o}C to 26.55 ^{o}C for Al-Qa'im and Basrah provinces respectively, also figure 3. Shows the map of Iraq which presents the locations of each station.

Result and Discussion

Models have most significant points in the selection of the mentioned station is a good relation with height correlation coefficient between target (output) and other station (input). In the first experiment the mean air temperature data of Baghdad province is used a target station (suggested as a station which contain the missed data on its data set), and (Sinjar, Mosul, Makhmur, Tikrit, Khanaqin, Al-Qa'im, Baghdad, Ali Al-Garbi, Al-Hayy, Najaf, Diwaniyah, Samawah, Nasiriyah and Basrah) station have been employed as an input of the (MLP) neural model. By using the same method, the other thirteen experiments were performed to interpolation the supposed missing data for each station for the period (2007 to 2008) in each layer, every neuron is connected to a neuron of advancement layer having different weights, neurons then produce an output signal by passing the summed signal through an activation function [18]. Each experiment in this work evaluated using some statistical indicators such a Root Mean Square Error (RMSE) which reflect the overall accuracy of the shape of the interpolated data and defined by:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(T_{oi}-T_{ei})^2}{n}} \qquad \text{..} \quad (2)$$

Where: (T_{oi}) is the observed value and (T_{ei}) the calculated value, (n) is the number of observations. The closer the calculated values to the observed, represents by a smaller (RMSE). The second statistical indicator employed in this study was Mean Percentage Error (MPE) and defined as follow:

$$\mathrm{MPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{T_{ei}-T_{oi}}{T_{oi}} \qquad \text{..} \quad (3)$$

The last statistical indicator employed in this study was Mean Absolute Error (MAE) and defined as follow:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\, T_{oi} - T_{ei}| \qquad \text{..} \quad (4)$$

The (MLP) neural models have a two-step procedure, the first called training (which include the creating of the artificial neuron weights), the second step called testing procedure (include the initial evaluation of the designed (MLP) neural model). Fourteen training and testing were performed for each meteorological station spritely, the evaluation results are given in table 2. The correlation coefficient (r) between the actual data and interpolated ranged between 0.9767 and 0.9980 for Najaf and Mosul station respectively but correlation coefficient are not enough indicator for evaluation any model, therefore the (RMSE) have been calculated for each model, Which is ranged from maximum value 2.22 °C and minimum value 0.629 °C also for Najaf and Mosul station. Also the Mean Absolute Error (MAE) appears with maximum value 1.62 °C in Najaf and minimum value 0.367 °C for Mosul station respectively.

In present work, three years of mean air temperature from 2007-2008 were supposed as a missed data set in stations (Sinjar, Mosul, Makhmur, Tikrit, Khanaqin, Al-Qa'im, Baghdad, Ali Al-Garbi, Al-Hayy, Najaf, Diwaniyah, Samawah, Nasiriyah and Basrah) were taken for testing and evaluation the accuracy of the (MLP) neural models, which illustrated in figures (4 – 5).

Conclusion

In this research, an (MLP) algorithm was used to develop fourteen (ANN) models to estimate the missing air temperature data in Iraq. Observation of monthly mean air temperature for the period 1993-2008 were used for training the (ANN) models and the observation for the period 2007-2008 were used for the evaluation process. Statistical analysis, including the (r, RMSE, MPE and MAE) suggested that the (MLP) neural model is suitable to be used for interpolating missing data of monthly mean air temperature for a network consisting many meteorological stations.

References

[1]Bulut, H., Büyükalaca, O., and Yilmaz, T. 2003. New outdoor heating design data for Turkey. Energy 28:1133–1150.

[2]Dodson, R. and Marks, D., 1997. Daily air temperature interpolated at high spatial resolution over a large mountainous region. Clim. Res. 8, 1-20.

[3]Robeson, S.M., 1993. Spatial interpolation, network bias, and terrestrial air temperature variability. Publ Climatol 46(1), 1-51.

[4]Robeson SM (1995) Resampling of network-induced variability in estimates of terrestrial air temperature change. Clim Change 29:213-220.

[5]Nalder, I.A., Wein, R.W., 1998. Spatial interpolation of climatic normal: test of a new methods in the Canadian boreal forest. Agric. & For. Met. 9, 211-225.

[6]Price, D.T., McKenney, D.W., Nalder, I. A., Hutchinson, M. F., Kesteven, J. L., 2000. A comparison of two statistical methods for spatial interpolation of Canadian monthly mean climate data. Agric. & For. Met. 101, 81-94.

[7]Ollinger, S.V., Aber, J.D., Lovett, G. M., Millham, S.E., Lathrop, E.G., Ellis, J.M., 1993. A Spatial model of atmospheric deposition for the northeastern U.S. Ecol. Appl. 3, 459-472.

[8]Bolstad, P. V., Swift, L., Collins, F., Regniere, J., 1998. Measured and predicted air temperatures at basin to regional scales in the southern Appalachian mountains. Agric. & For. Met. 91, 161-176.

[9]Christine L. Goodale, Jhon D. Aber, Scott V. Ollinger, 1998. Mapping monthly precipitation, temperature, and solar radiation for Ireland with polynomial regression and a digital elevation model. Clim. Res. 10, 35-49.

[10]Ward System Group, Frederic, MD, 1993. Manual of Neuron Shell 2. Frederic, MD: Ward Systems Group.

[11]McAvoy, T.J., Wang, N.S., Naidu, S., Bhar, N., Gunter, J. and Simmons, M., 1989. Interpreting biosensor data via back propagation. IEEE International Neural Network Conference in 1989, 8, 227-233.

[12]Leonard, J.A., Kramer, M.A. & Ungar, L.H., 1992. Using radial basis functions to approximate function and its error bonds, IEEE Trans. on Neural Networks 3, 624-627.

[13] Han, H. and Felker, P., 1997. Estimation of daily soil water evaporation using an artificial neural network. Journal of Arid Environments 37, 251-260.

[14]Cook D.F. and Wolfe, M.L., 1991. A back-propagation neural network to predict average air temperatures, AI Applications 5(1), 40-46.

[15]Bruton, J.M., McClendon, R.W., Hoogenboom, G., 2000. Estimating daily Pan Evaporation with Artificial neural networks. Trans. ASAE 43(2), 491-496.

[16]Guan, L., Yang, J., and Bell, J. M. 2007. Cross-correlation between weather variables in Australia. Building & Environ. 42:1054–1070.

[17]Kidson, J.W., and K.E. Trenberth, 1988: Effects of missing data on estimates of monthly mean general circulation statistics, J. Climate, 1, 1261–1275.

[18]http://www.alyuda.com/neural-network-software.htm.

[19]Haykin, S. 1994. Neural Networks, a Comprehensive Foundation. New Jersey: Prentice-Hall, Inc.

[20]Climatic Atlas of Iraq (1941-1970), Iraqi Meteorological Organization, Baghdad, Iraq, 1999.

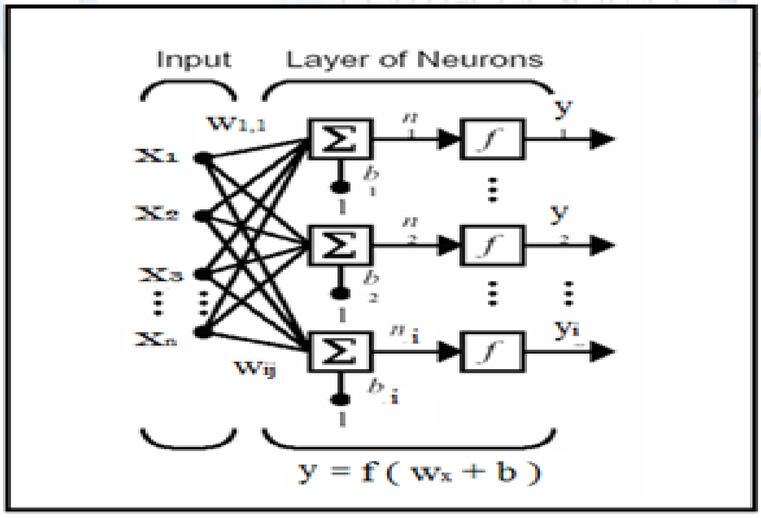

Figure 1. Multi-Layer Perceptron’s Architecture.

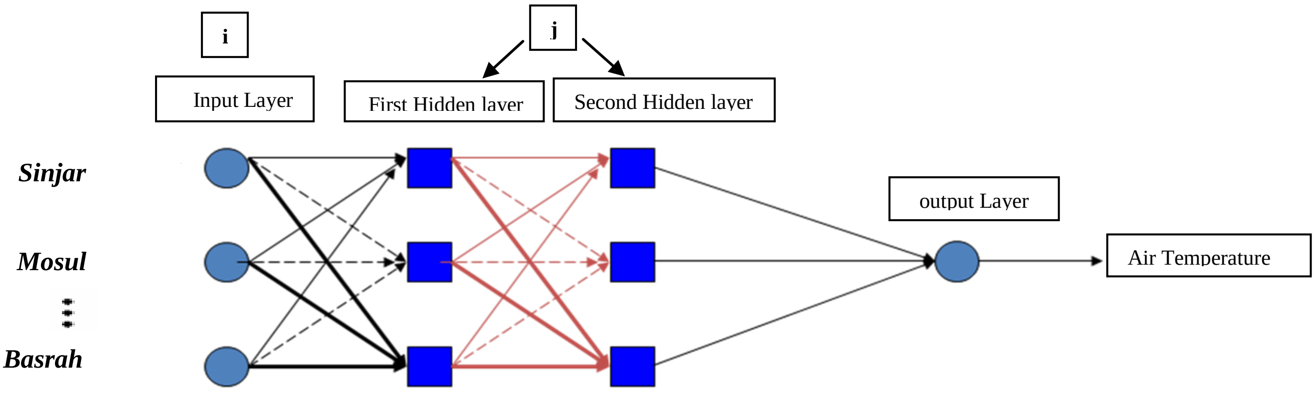

Figure 2. (ANN) architecture used in the interpolation process.

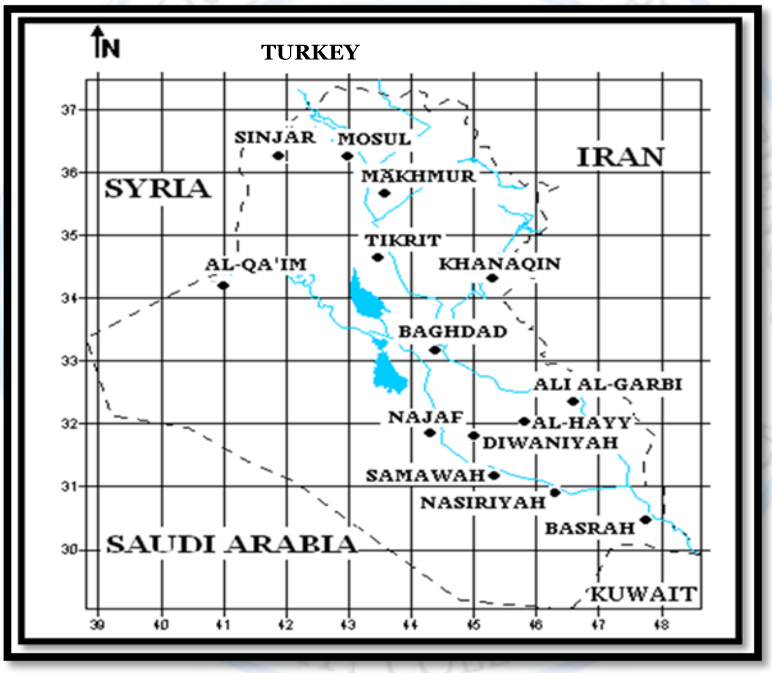

Figure 3. The selected Meteorological station shown in map of Iraq.

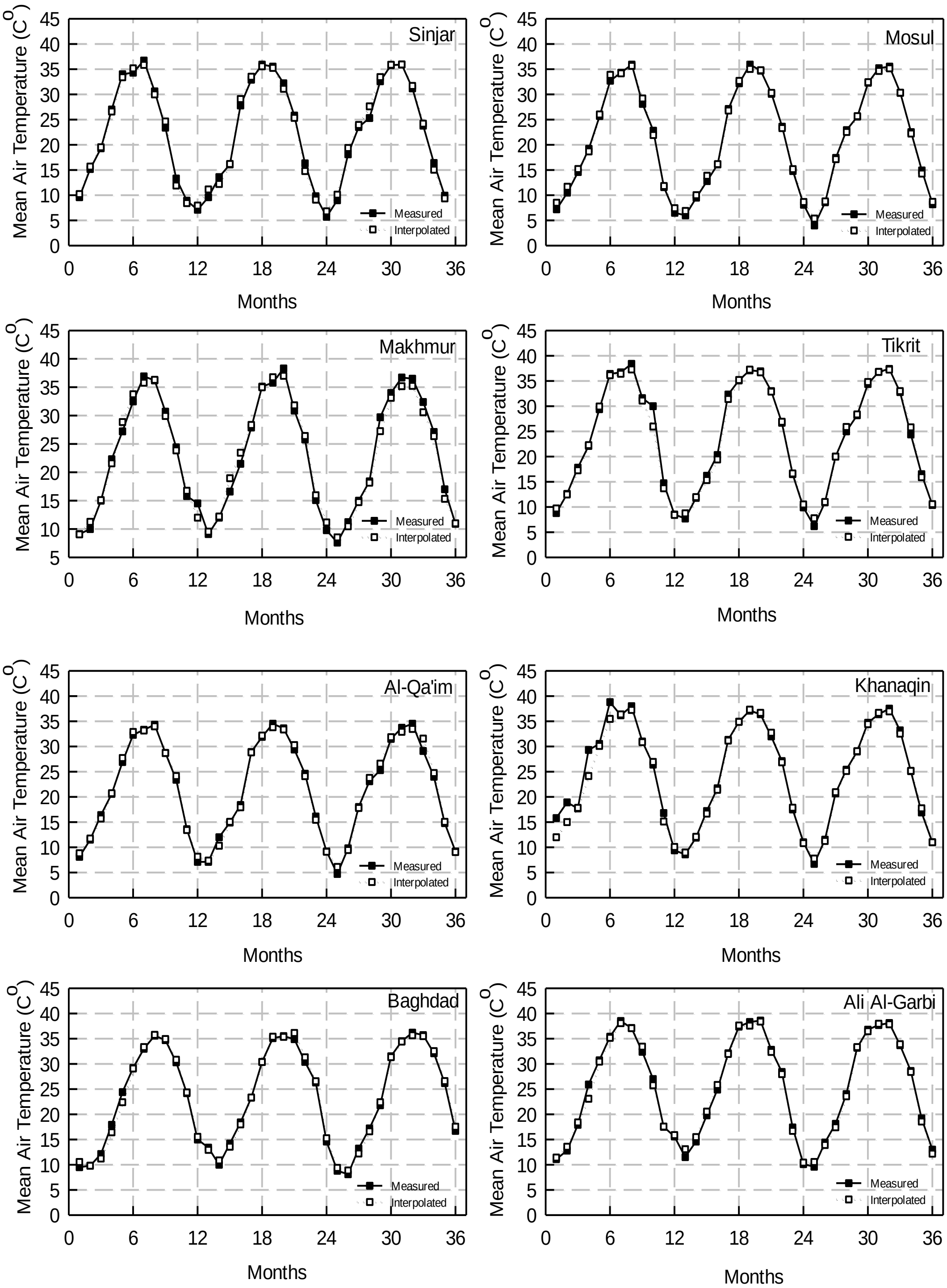

Figure 4. Represents the measured and interpolated monthly mean air temperature for stations Sinjar, Mosul, Makhmur, Tikrit, Khanaqin, Al-Qa'im, Baghdad and Ali Al-Garbi.

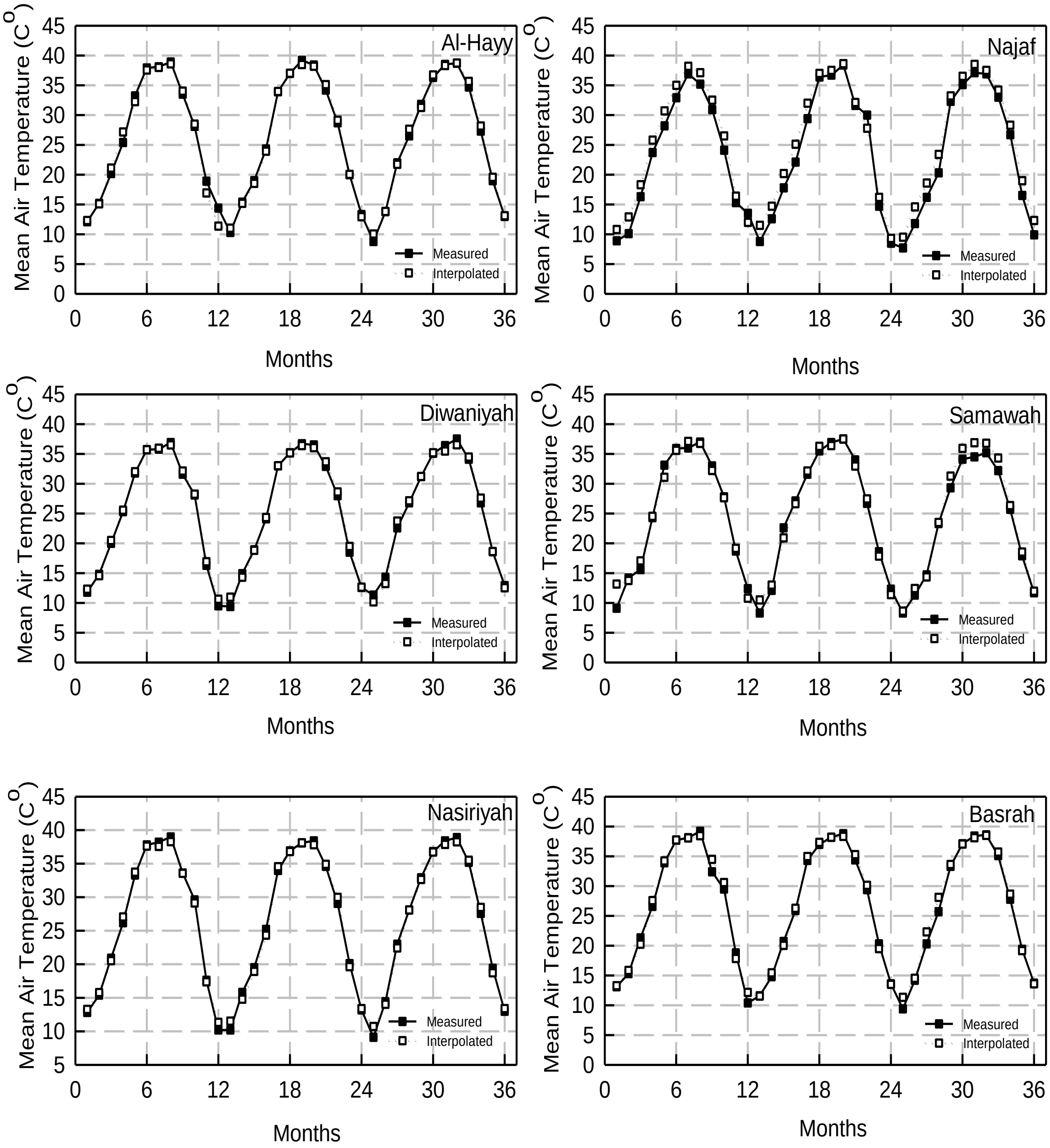

Figure 5. Represents the measured and interpolated monthly mean air temperature for Stations Al-Hayy, Najaf, Diwaniyah, Samawah, Nasiriyah and Basrah.

Tables 1. The geographical coordinates and mean, maximum and minimum air temperature of the Meteorological station [20].

Station	Latitude (°N)	Longitude (°E)	Altitude (meters)	Mean Air Temperature (°C)	Minimum (°C)	Maximum (°C)
Sinjar	36.32	41.38	465	21.45	5.40	38.60
Mosul	36.32	43.15	223	20.64	4.00	37.40
Makhmur	35.46	43.35	270	22.77	6.80	39.40
Tikrit	34.35	44.18	107	23.18	6.20	39.20
Khanaqin	34.30	45.30	202	23.49	6.70	39.60
Al-Qa'im	34.08	14.21	177.5	21.16	4.70	35.80
Baghdad	33.23	44.23	34	23.29	6.70	37.60
Ali Al-Garbi	32.45	46.68	9	25.24	8.70	40.20
Al-Hayy	32.17	46.05	17	25.71	8.80	40.00
Najaf	31.98	44.32	32	25.00	7.80	41.20
Diwaniyah	31.98	44.98	20	24.92	9.40	38.80
Samawah	31.32	45.27	11.4	24.82	8.30	39.60
Nasiriyah	31.08	46.23	7.6	26.05	9.10	40.20
Basrah	30.57	47.78	2.4	26.55	9.40	40.20

Table 2. The statically indicators of the interpolated Meteorological stations.

Station	r	RMSE	MPE	MAE
Sinjar	0.9948	1.020	1.318	0.940
Mosul	0.9980	0.629	0.264	0.367
Makhmur	0.9882	1.554	3.198	1.147
Tikrit	0.9975	0.711	0.400	0.468
Khanaqin	0.9933	1.611	0.638	0.577
Al-Qa'im	0.9940	1.007	3.568	0.803
Baghdad	0.9953	0.913	1.184	0.740
Ali Al-Garbi	0.9908	1.301	1.247	0.770
Al-Hayy	0.9956	0.905	0.513	0.541
Najaf	0.9767	2.221	3.64	1.62
Diwaniyah	0.9968	0.717	1.030	0.597
Samawah	0.9921	1.203	2.342	1.132
Nasiriyah	0.9977	0.641	0.347	0.474
Basrah	0.9962	0.833	0.462	0.582