Al-Mustansiriyah Journal of Science
ISSN: 1814-635X (print), ISSN:2521-3520 (online) Volume 33, Issue 4, 2022 DOI: http://doi.org/10.23851/mjs.v33i4.1202

Research Article Open Access

Treatment Missing Data of Daily and Monthly Air Temperature in Some Iraqi cities by Using Curve Fitting

Ali Hamid Yassen[1]*, Asraa Khtan Abdulkareem[2]

[1]Weather Forecasting Department, Iraqi Meteorological Organization and Seismology, Baghdad, IRAQ.
[2]Department of Atmospheric Science, College of Science, Mustansiriyah University, Baghdad, IRAQ.

*Correspondent contact: alihamid.1987@yahoo.com

Article Info

Received
04/08/2022

Accepted
06/09/2022

Published
30/12/2022

ABSTRACT

Climate change has become fast and entered a new stage and began to affect all regions of the world. so, the climate must be analyzed and studied accurately. In order to do this, should be available a continuous database without interruptions, to improve the accuracy of forecasts. Therefore, this research aims to treat the missing temperature data for the stations (Baghdad, Hilla, Basra, Nasiriya, and Samawa) by using the curve fitting method. In the monthly treatment for the period (1980-2020), it was observed that the highest match between the real and the treatment values using the Gaussian function and the sine wave function was recorded in the summer months at (100%), and the lowest match was recorded in the winter months. The daily treatment period (2010-2020) recorded the highest match at (97%) in the summer, and the lowest match was recorded in the winter months. In order for the treated values to be close to the real values, it is recommended to use this method for months from April to October. In the winter months, it should be used with caution.

KEYWORDS: Missing data; treatment; temperature; curve fitting method; Iraq.

الخلاصة

أصبح تغير المناخ سريعاً ودخل مرحلة جديدة وبدأ يؤثر على جميع مناطق العالم. لذلك لا بد من تحليل المناخ ودراسته بعناية، لتحسين دقة التنبؤات ،من أجل القيام بذلك ، يجب أن تتوفر قاعدة بيانات مستمرة دون انقطاع ، لذلك يهدف هذا البحث إلى معالجة بيانات درجة الحرارة المفقودة اليومية والشهرية للمحطات (بغداد،الحلة،البصرة ،الناصرية ،سماوة) من خلال طريقة ملائمة المنحنى.في المعالجة الشهرية للفترة (**1980-2020**) لوحظ ان اعلى تطابق بين القيم الحقيقية والمعالجة باستخدام دالة غاوسين ودالة دالة الموجة الجيبية سجل في اشهر الصيف (**100**%) ،واقل تطابق سجل في اشهر الشتاء.اما في المعالجة اليومية للفترة (**2010-2020**) سجل اعلى تطابق (**97**%) في اشهر الصيف، واقل تطابق سجل في اشهر الشتاء. لكي تكون القيم المعالجة مقاربة للقيم الحقيقية ينصح باستخدام هذه الطريقة من شهر ابريل الى شهر اكتوبر.اما في اشهر الشتاء يجب استخدامها بحذر.

INTRODUCTION

Climate change has sparked a serious interest in studying and inferring climate variables and climate-related topics in recent years, due to concerns posed by global warming [1]. So, it became the main goal has always been the prediction of the numbers that tell us the facts about the weather, such as checking the temperature, humidity, predicting rain, predicting natural disasters [2]. therefore, should be a focus on Data collection and management, for the availability of climate records data., so continuous data in the weather record is very necessary for the accuracy of prediction models and research studies [3]. and important building decisions in industry and agriculture [4]. but Through time, weather stations may begin to be out of service or stopped to need repair, and Since the Surface weather station networks are not enough in many regions of the world caused of composition and maintenance costs and the complexity of topography for that reason, through those periods, there will be missing data [5]. Therefore, continuous data observations of any weather element are critical to know the behavior and influence of these elements among themselves, as the weather forecasting model becomes more accurate the more data it is fed [6]. Ceylon Yozgatligil *et al.* (2013) [7] compared some methods for calculating missing values of monthly precipitation and mean temperatures in

Turkey, it was found that the multiple imputation strategies adopted by the Monte Carlo Markov series based on expectations-maximization (EM-MCMC) will reduce uncertainty in the data and give more accurate results. Jaber Rahimi *et al.* (2017) [8] evaluate different missing data reconstruction methods for daily minimum temperature for stations of different elevated, of Iran for the period 1965-2010, Results revealed that Artificial Neural networks stood in priority for the reconstruction of daily minimum temperature data. Sura T. Nassir, et al. (2018) [9] uses the AutoRegressive Integrated Moving Average (ARIMA) model for estimating the missing data (air temperature, relative humidity, and wind speed) for mean monthly variables in three stations in Iraq, which is found that The ARIMA model is accurate but data should be available in sufficiently large numbers to estimate the missing data. Okan Mert Katipoğlu, (2022) [10] studied the Prediction of Monthly average temperature missing data (1968-2017) using different machine learning methods, the most suitable machine learning method was chosen by the adaptive neuro-fuzzy inference system ANFIS to estimate monthly air temperatures in northeastern of Turkey.

This work aims to treate missing temperature data in records of the Iraqi Meteorological Organization and Seismology (IMOS) by using (curve fitting method) to replace traditional methods.

MATERIALS AND METHODS

Study Area and Data Sources

study area represents is Iraq, this area which is mostly dry and semi-arid, one of its prominent characteristics is the extremes of temperatures often on the same day, between day and night, and between summer and winter, and this great contradiction is clear [11]. In the summer, temperatures in the middle and southern sections of the country can exceed 50 degrees Celsius, in contrast while in the winter, temperatures in the north and mountain areas can drop below freezing [12]. The monthly mean temperature varies from (30-45) degrees Celsius in the summer to (5-20) degrees Celsius in the winter [13]. In this study historical records of mean monthly temperature were acquired from the Iraqi Meteorological Organization and Seismology (IMOS) for forty years of the period (1980-2020) and records of mean daily temperature for eleven years of the period (2010-2020). The Long-term data were collected from 5 weather surface stations located in different regions of the Country. Which is Baghdad and Hilla stations, which represent central Iraq, which is geographically located in the northern hemisphere between latitudes (32°27' -33°18' N) north of the equator and Between longitudes (44°24'-44°27'E) east of Greenwich line. As well as some stations located in southern Iraq, Basra, Nasiriya and Samawa which is geographically located in the northern hemisphere between latitudes (30°31' -31°16'N) north of the equator and between longitudes (45°16'-47°47'E) east of Greenwich line. As shown in Table 1 and Figure 1.

Table 1. Location of the study area [14].

Station	Longitude	Latitude	Elevation
Baghdad	44° 24'	33° 18'	32 m
Hilla	44° 27'	32° 27'	27 m
Basra	47° 47'	30° 31'	2 m
Nasiriya	46° 14'	31° 01'	5 m
Samawa	45° 16'	31° 16'	11 m

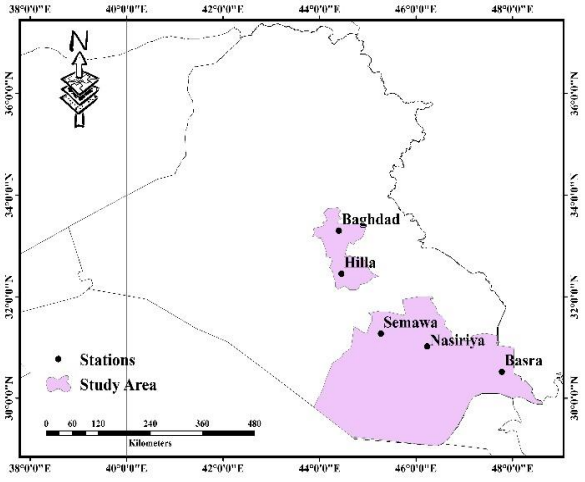


Figure 1. Weather stations that used in the study.

Curve fitting

Regression analysis is a technique for determining the "best fit" line or curve for a set of data points. The curve fit will, in most cases, generate an equation that may be used to find locations along the curve. Curve fitting, on the other hand, is the process of creating a curve that has the best fit for a set of data points, sometimes subject to limits. Fits the data "curve fitting" is a term that refers to the process of determining a curve that goes through a group of points [15]. In this study, many functions have been tested from the curve fitting method (logarithm, exponential, gaussian, and sin wave) by using (sigma plot program), but these two

functions are better than the rest of the functions, because the path of the two functions are in the form of waves, which is similar to the path of temperature data when it is drawn, where it is in the form of waves represented by the minimum temperature (base) and maximum temperature (peak), so we will adopt them in our study.

Peak, Gaussian, for Parameter (G)

$$Tr = y_{\circ} + aexp^{-0.5\left(\frac{x-x_0}{B}\right)^2} \quad (1)$$

Waveform-sine, for Parameter (S)

$$Tr = y_{\circ} + Asin\left[\frac{2\pi X}{B} + C\right] \quad (2)$$

where: Tr= treatment missing data, $y_{\circ}$= mean of available values, (x)=days, $(x_{\circ})$ primary variable, B= period, C= phase shift.

In this study, we assume a special Threshold was used, which was calculated from the general standard deviation (S.D.G) for all months, which was approximately for treatment monthly (±1.5) and daily (±2.5) as shown table for monthly (2), and (3) for daily, and then compared the actual with result from curve fitting values as in the equation 3.

$$R = A_{V-R_c} \quad (3)$$

where R= The result of comparing Actual with curve fitting values, A_v=Actual values, R_c=curve fitting values.

If the result from Equation 3 (daily or monthly period) is higher than the threshold is isolated for these values, but, if the difference within the threshold is considered an acceptable value and calculated the percentage of matching.as equation 4 for monthly:

$$\frac{(Numberofvalueswithinthresholdmonthly)}{allvalues} * 100\% \quad (4)$$

And for daily the percentage of matching:

$$\frac{(Numberofvalueswithinthresholdmonthly)}{allvalues} * 100\% \quad (5)$$

Table 2. Mean monthly temperature and standard deviation for 40 years (1980-2020).

Station	January		February		March		April		May		June	
	Mean	S.D	Mean	S.D	Mean	S.D	Mean	S.D	Mean	S.D	Mean	S.D
Baghdad	9.8	1.6	12.4	1.7	17	1.8	23.1	1.4	29.2	1.2	33.2	1.2
Hella	10.4	1.6	13	1.6	17.6	1.7	23.5	1.7	29.3	1	33.2	0.9
Basra	12.6	1.6	15.1	1.7	19.8	1.6	26.3	1.3	32.8	1.4	36.6	1.5
Nasiriya	12	1.5	14.7	1.7	19.9	1.9	25.7	1.2	32.1	1.2	36	1.2
Samawa	11.4	1.6	13.9	1.7	18.8	2.1	25.1	1.2	31.4	1.2	35.2	1
Station	July		August		September		October		November		December	
	Mean	S.D	Mean	S.D	Mean	S.D	Mean	S.D	Mean	S.D	Mean	S.D
Baghdad	35.5	1.2	34.9	1.6	30.9	1.3	24.9	1.5	16.5	1.5	11.5	1.7
Hella	35.2	1.1	34.7	1.3	31	1	25.3	1.2	16.7	1.3	11.9	1.7
Basra	38.3	1.4	37.8	1.8	34.1	1.3	28.3	1.4	19.8	1.2	14.2	1.7
Nasiriya	37.6	1.4	37.6	1.7	34	1.3	27.9	1.5	19.2	1.2	13.7	1.7
Samawa	37	1.2	36.5	1.7	33	1.2	26.8	1.4	18.3	1.4	13.3	1.7

Table 3. Mean daily temperature and the standard deviation from (2010-2020).

Station	January		February		March		April		May		June	
	Mean	S.D	Mean	S.D	Mean	S.D	Mean	S.D	Mean	S.D	Mean	S.D
Baghdad	11	2.8	13.3	3.5	18.1	3.3	23.7	3.3	29.9	3.2	34.4	2.1
Hella	11.2	2.5	13.7	3.4	18.5	3.2	23.8	3.10	29.7	2.9	33.8	2
Basra	13.9	2.7	16.2	3.3	21.1	3.2	26.9	2.8	33.7	3	38.1	2
Nasiriya	13	2.6	15.7	3.6	20.9	3.5	26.5	2.3	33	3	37.3	2.1
Samawa	12.4	2.6	15	3.5	20.2	3.5	25.8	3.4	32.2	3.5	36.2	21
Station	July		August		September		October		November		December	
	Mean	S.D	Mean	S.D	Mean	S.D	Mean	S.D	Mean	S.D	Mean	S.D
Baghdad	37	2	36.2	2	32.4	2.4	25.7	3.5	17.1	3.6	12	2.9
Hella	36	1.7	35.6	11.8	32	2	25.6	3.32	17	3.7	11	2.6
Basra	40	1.6	39.3	1.5	35.9	2	29.4	3.4	20	3.5	14.9	2.8
Nasiriya	39.3	1.9	39	1.7	35.5	2.2	28.7	3.5	19.5	3.5	14	2.8
Samawa	38.3	2.8	37.7	1.8	34.3	3	30	3.45	18.8	3.5	13.5	2.8

RESULTS AND DISCUSSIONS

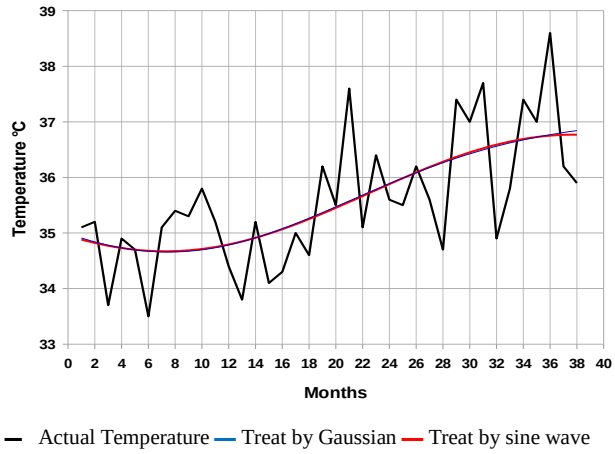
The Equations (1) and (2) were applied to the monthly and daily data set, as shown in Figures 2a and 2b for the mean monthly temperature and Figures 3a and 3b for the mean monthly temperature. The blue line represents the values by

function1(Gaussian), the red line represents the values of Equation 2 (sinewave) and the black line represents the actual mean monthly and daily temperature data in Figures 2 and 3 respectively. Baghdad station was drawn as an example for the rest of the other stations and shows how these two functions work. Almost all study area is similar in terms of percentage of matching as shown in Figure 4, the two functions are nearly matching.

Treatment of monthly missing data

Through Figure 4, the matching of the two equations were observed during the monthly periods in all study stations from April to October. Equation (1) ranged between (72%-100%) and the matching in Equation (2) ranged between (78%-100%). Otherwise, the matching fluctuated between (59% - 89%) in Equation (1) and between (61% - 95%) in Equation (2) for the rest of the months. From the above, both equations work well in the summer months, and Equation (2) is often better than Equation (1) in all months.

(a) Mean monthly (July) temperature data (°C) over Baghdad for the 1980-2020 period.



(b) Mean monthly (December) temperature data (°C) over Baghdad for the 1980-2020 period.

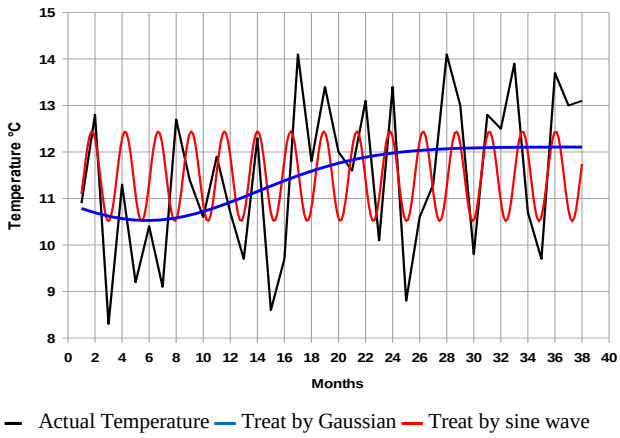
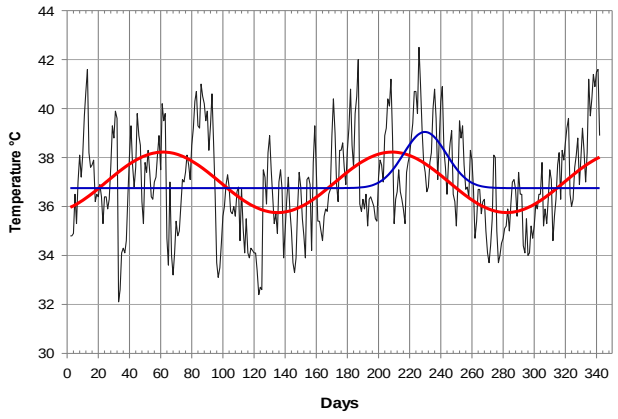


Figure 2. Curve fitting of Mean monthly temperature data (°C) over Baghdad for the 1980-2020 period by two functions Gaussian and sinewave, a (July) and b (December).

(a) Mean daily (July)temperature data (°C) over Baghdad for the 2010-2020 period.



(b) Mean daily (December) temperature data (°C) over Baghdad for the 2010-2020 period

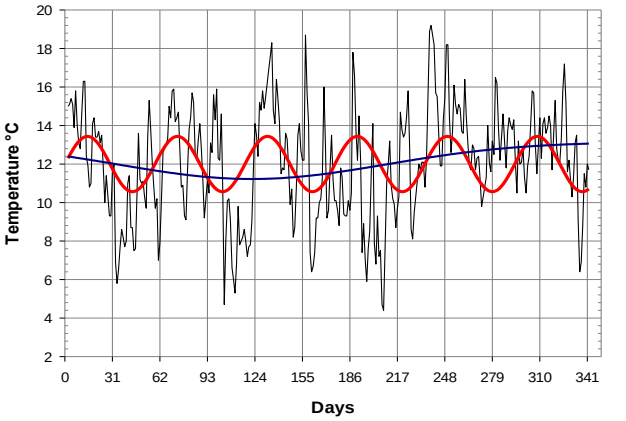


Figure 3. Curve fitting of mean daily temperature data (°C) over Baghdad for the 2010-2020 period by two functions Gaussian and sinewave, a (July) and b (December).

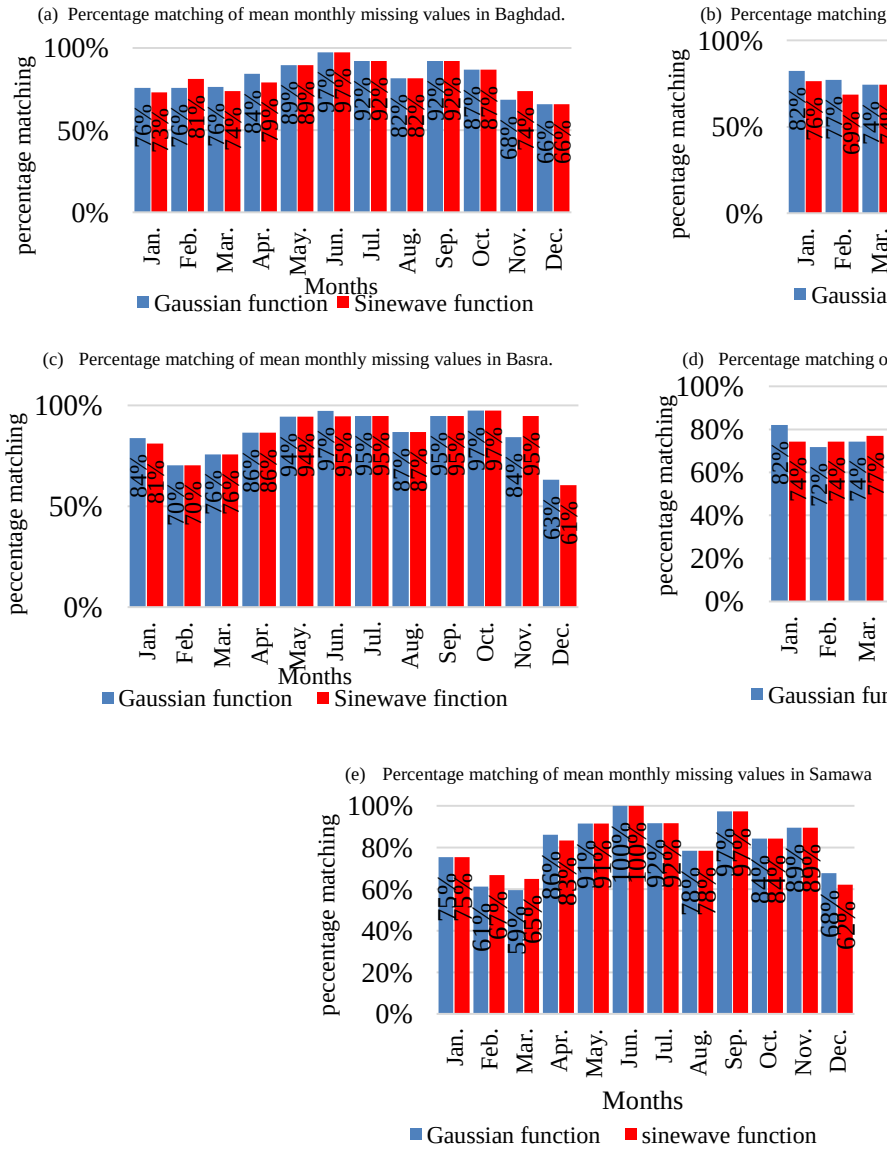


Figure 4. Percentage matching of mean monthly missing values overall study Areas for the (1980-2020) periods: a) Baghdad station, b) Hilla station, c) Basra statin, d) Nasiriya station and e) Samawa station.

Accuracy of Treatment missing monthly data

To know the accuracy of this these methods, the Actual data must be present, so this period was chosen from 2005-2020 because it does not contain missing data in this station as in Table 4.

It was noted that the result of applying this example is matching to all the results of the previous methods when testing them for accuracy, where the actual values for the month of June for the period (2013-2014-2015), its (33, 33.7, 34) respectively. And for December were the actual values for the period (2013-2014-2015), its (12.5, 13.9, 10.7) respectively. These two months were chosen in the test because June represents the best match in all stations, and December is the least match in all stations.

Table 4. Accuracy of Treatment missing monthly data.

June (2013-2015)	**Treat. Temp. by Gaussian**			**Treat. Temp. by sinewave**		
One missing value	34.4			34.5		
Two missing values	34.5	34.6		34.6	34.7	
Three missing values	34.6	34.7	34.8	34.6	34.7	34.8
December (2013-2015)	**Treat.Temp. by Gaussian**			**Treat.Temp. by sinewave**		
One missing value	12			12.2		
Two missing values	11.9	11.9		10.9	11.2	
Three missing values	12	12	12	11	11	12.2

Treatment of daily missing data

Through Figure 5 the matching of the two equations were observed in the daily treatment in all study stations in summer months from June to September. Function (1) ranged between (78%-95%), and function (2), the matching ranged between (80%-97%). while less matching for the rest of the months; ranged between (52%-71%) and between (57%-73%) in functions (1) and (2) respectively.

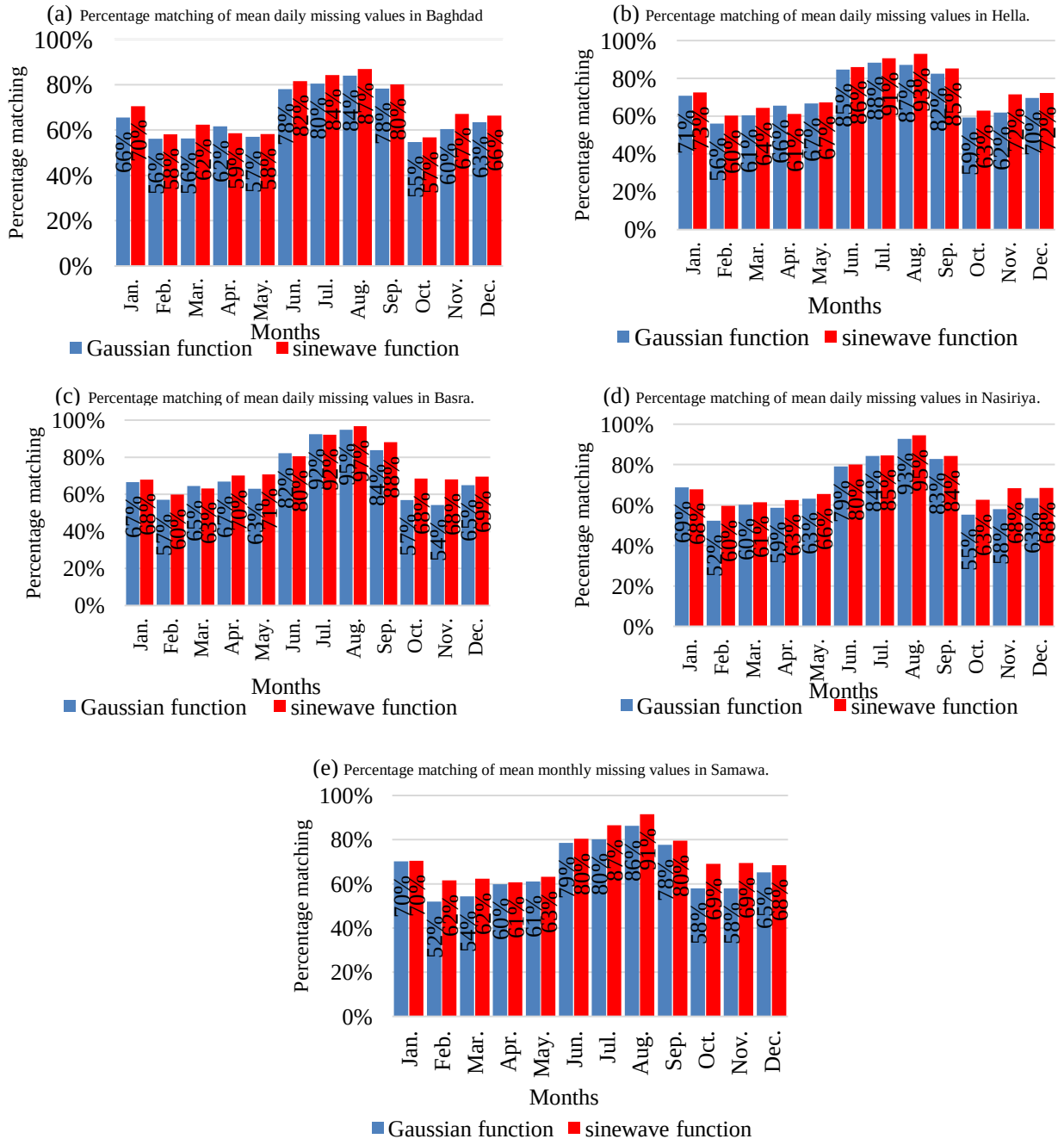


Figure 5. Percentage Matching of mean daily missing values overall study Areas for the (2010-2020) periods: a (Baghdad station), b (Hilla station), c (Basra statin), d (Nasiriya station) and, e (Samawa station).

Accuracy of Treatment missing daily data

Baghdad station was chosen as an example to find daily missing temperature values as shown in Table 5. in the same time Two months were chosen, February and December, because they represent the highest and lowest matching in both methods based on Figure 4.

In the first test period, we selected four days, 8th - 11th February 2015, and the actual values are 16.8 C,16.8 C,18.3 C, and 17.6 C respectively.

And the second test period is from 13th to16th December 2015, where the actual values are 35.9 C, 36.6 C, 37.9 C, and 37.2C respectively.

after applying the treatment of missing daily data and showing the result in Table 5, We find that the sine function is better than the Gaussian function in the treatment of the missing daily temperature data. generally, noted that the accuracy of the two functions increases in general in the summer months due to the stability of weather systems for the study area, and the change in average temperature is almost constant, so the functions represent it well. while the accuracy of two functions in the treatment of missing data for the remaining months was decreased, I think that in the winter months the temperatures are unstable due to the cloud cover as well as the presence of cold air masses, but in the spring and fall they are transitional months in which the temperatures change.

Table 5. Accuracy of Treatment missing daily data.

Date missing (8-11/2/2015)	**Treat. Temp. by Gaussian**				**Treat. Temp. by sinewave**			
One missing value	14.2				16.38			
Two missing values	14.1		14.1		13.3		16.8	
Three missing values	13.9	13.9	13.9		15.6	15.7	15.7	
Four missing values	13.7	13.7	13.7	13.7	11.5	10.4	10	10.2
Date (13-16/8/2015)	Treat. Temp. by Gaussian				Treat. Temp. by sinewave			
One missing value	38.2				35.54			
Two missing values	38.3		38.3		37.4		37.3	
Three missing values	38.3	38.3	38.3		37.4	37.3	37.3	
Four missing values	38.4	38.4	38.4	38.3	37.3	37.2	37.0	36.9

CONCLUSIONS

In this paper, we try to find the missing data on daily and monthly temperature by using curve fitting for some stations in the central and southern regions of Iraq. After several tests on the data to obtain high-accuracy treatment results. It was found that the cutoff should not be at the beginning or the end of the dataset, if one value is missing for a set of ten values the result will be great. And during the monthly treatment, there should be at least 5 values before and after the cutoff data. In the daily treatment, the data must not be cut off more than three consecutive values for a set of 10 values. In general, the sine wave function is
slightly better than the Gaussian function, whether in the monthly or daily treatment. The treatment of missing data is more accurate in the summer months from other months. This method is good in representing the missing data because it treats the missing data from the available data of the same station.

Disclosure and conflict of interest: The authors declare that they have no conflicts of interest.

REFERENCES

[1] S. K. Mondal, R. Chakraborty, S. Choudhury, B. Roy, S. Podder, P. Dey, *et al.*, "Weather Forecasting System," *AJEC*, 2022.

[2] R. B. Alley, K. A. Emanuel, and F. Zhang, "Advances in weather prediction," *Science*, vol. 363, pp. 342-344,2019. https://doi.org/10.1126/science.aav7274

[3] E. Afrifa-Yamoah, U. A. Mueller, S. Taylor, and A. Fisher, "Missing data imputation of high-resolution temporal climate time series data," *Meteorological Applications*, vol. 27, p. e1873, 2020. https://doi.org/10.1002/met.1873

[4] G. Tang, M. P. Clark, and S. M. Papalexiou, "SC-earth: a station-based serially complete earth dataset from 1950 to 2019," *Journal of Climate*, vol. 34, pp. 6493-6511, 2021. https://doi.org/10.1175/JCLI-D-21-0067.1

[5] I. Gad and B. Manjunatha, "Performance evaluation of predictive models for missing data imputation in weather data," in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2017, pp. 1327-1334. https://doi.org/10.1109/ICACCI.2017.8126025

[6] N. ustafsson, N., Janjić, T., Schraff, C., Leuenberger, D., Weissmann, M., Reich, H., Fujita, T. (2018)., "Survey of data assimilation methods for convective-scale numerical weather prediction at operational centres," *Quarterly Journal of the Royal Meteorological Society*, vol. 144, pp. 1218-1256, 2018. https://doi.org/10.1002/qj.3179

[7] C. Yozgatligil, S. Aslan, C. Iyigun, and I. Batmaz, "Comparison of missing value imputation methods in time series: the case of Turkish meteorological data," *Theoretical and applied climatology*, vol. 112, pp. 143-167, 2013. https://doi.org/10.1007/s00704-012-0723-x

[8] J. Rahimi, A. Khalili, and J. Bazr Afshan, "Evaluation of different missing data reconstruction methods for daily minimum temperature in elevated stations of Iran:

comparison with new proposed approach," *Iranian Journal of Soil and Water Research*, vol. 48, pp. 231-239, 2017.

https://dx.doi.org/10.22059/ijswr.2017.62576

[9] S. T. Nassir, A. B. Khamees, and W. T. Mousa, "Estimation the Missing Data of Meteorological Variables In Different Iraqi Cities By using ARIMA Model," *Iraqi Journal of Science*, pp. 792-801, 2018.

[10] O. M. Katipoğlu, "Prediction of missing temperature data using different machine learning methods," *Arabian Journal of Geosciences*, vol. 15, pp. 1-11, 2022. https://doi.org/10.1007/s12517-021-09290-7

[11] W. H. Hassan and B. K. Nile, "Climate change and predicting future temperature in Iraq using CanESM2 and HadCM3 modeling," *Modeling Earth Systems and Environment*, vol. 7, pp. 737-748, 2021. https://doi.org/10.1007/s40808-020-01034-y

[12] H. T. Majeed, W. G. Nassif, and Y. Q. Tawfeek, "Effects of meteorological parameters on surface concentration of carbon monoxide over Iraq," *J. Green Eng*, vol. 10, pp. 5927-5940, 2020.

[13] M. Al-Mukhtar and M. Qasim, "Future predictions of precipitation and temperature in Iraq using the statistical downscaling model," *Arabian journal of geosciences*, vol. 12, pp. 1-16, 2019. https://doi.org/10.1007/s12517-018-4187-x

[14] WMO, "Guide to Instruments and Methods of Observation," ed: World Meteorological Organization Geneva, 2018.

[15] A. M. Abdul-Jabbar and A. K. Abdulkareem, "Conversion of rain data from surface stations to forecast models data," *Mesopotamia Environmental Journal*, vol. 5, 2020.

How to Cite

A. H. Yaseen and A. K. . Abdulkareem, “Treatment Missing Data of Daily and Monthly Air Temperature in Some Iraqi cities by Using Curve Fitting”, *Al-Mustansiriyah Journal of Science*, vol. 33, no. 4, pp. 34–41, Dec. 2022.