ABUAD Journal of Engineering and Applied Sciences (AJEAS)

ISSN: 2992-5584

Volume 3, Issue 1, 135-144

Analysis of Times Series Clear Sky Global Horizontal Irradiance dataset for Application in Photovoltaic System Design

Anyanime Tim UMOETTE[1*], Joan Benedict UMOH[1], Imo Edwin NKAN[1] and Mbetobong Charles ENOH[1]

[1]*Department of Electrical and Electronic Engineering, Akwa Ibom State University, Nigeria*

libertycoast@yahoo.com, joan.umoh@gmail.com, imoenkan2018@gmail.com, festus_mb@yahoo.com

Correspondence: libertycoast@yahoo.com; Tel.: +2348038898633

Date Submitted: 25/05/2025
Date Accepted: 22/06/2025
Date Published: 30/06/2025

Abstract: *In this study, analysis of the times series clear sky global horizontal irradiance dataset for application in photovoltaic (PV) system design is presented. Specifically, the Clear Sky Global Horizontal Irradiance (CKGHI) dataset spanning from 2003 to 2023 for Akwa Ibom State University, Obio Akpa campus, located at latitude and longitude 4.964487 and 7.759773, is considered. The data is retrieved from the National Aeronautics and Space Administration (NASA) Data Access Viewer (DAV) online platform for meteorological data. The case study of the 21-year daily CKGHI dataset has 7670 daily data records with a daily mean value of 6.2549 kW-hr/m^2/day and a 0.4961 standard deviation, and all the 40 potential outliers identified occurred in the values below the lower fence of 4.8, and the outliers ranged from 3.996 to 4.76. The results showed that as the time span for the moving average increases from 3, 5, 7 and up to 365 days, the lower boundary value for outliers increases, whereas the upper boundary values decrease. The two values tend towards the mean daily value of the entire dataset, which is about 6.2549. In all, the results showed that with the use of moving average and outlier determination, it is possible to develop heuristic mechanisms that can be used to select a suitable solar radiation value for sizing a solar power system so that the desired loss of load specification can be achieved for the given days of power autonomy.*

Keywords: Descriptive statistics, time series dataset, solar radiation, photovoltaic power system, Z-score outlier method.

1. INTRODUCTION

Nowadays, in Nigeria, the national grip power supply system has been privatized and so the cost of energy from the national grid has increased tremendously [1, 2]. As such, many clients are searching for cost effective alternative energy sources. The readily available alternative energy option for most Nigerians is the solar photovoltaic (PV) power which is easy to deploy using solar panels and rechargeable storage battery bank [3, 4, 5].

One major concern with the design and use of PV solar power is the variability in the solar radiation which determines the energy yield of the PV panels [6]. As such, designers and users rely on carefully selected PV values for the sizing of their PV power solar panels and battery bank for any given days of power autonomy and tolerable loss of load specifications. In practice, designers rely on the use of the mean of the daily solar radiation data computed using daily dataset for a given period which is mainly several years [7, 8, 16, 17]. In other case, the minimum daily or minimum monthly mean values is selected to mitigate loss of load [9, 10]. However, the two approaches fail to meet the desired PV power system performance in most cases. Accordingly, in this study, statistical analysis approach that can enable designers of PV power system to use statistical parameters along with moving average and outlier determination to carefully select the solar radiation data value for application in the sizing of the PV power system is presented. The study shows how the moving average and outlier determination can be used to identify the possible point of power outages in the solar power system. By employing the ideas presented in this work, solar power designers can design power system that effectively meet their load demand with the desired days of power autonomy and loss of load specifications.

2. METHODOLOGY

2.1 Description of the Study Site and Time Series Dataset

In this work, statistical analysis is conducted for a time series dataset of Clear Sky Global Horizontal Irradiance [11, 12, 15], CKGHI for Akwa Ibom State University, Obio Akpa campus located at latitude and longitude of 4.964487 and 7.759773 respectively, as shown in Figure 1. The CKGHI dataset spanning from 2003 to 2023 was downloaded from National Aeronautics and Space Administration (NASA) online platform available at the URL of https://power.larc.nasa.gov/data-access-viewer/which is the NASA Data Access Viewer (DAV) portal for meteorological data [13, 14, 16].

Specifically, the study seeks to evaluate the dataset for application in the design of solar power system for the university campus. In this wise, the statistical analysis conducted included measure of dispersion, determination of missing data and outliers, as well as data distribution and computation of moving average using different time scales. In addition, histogram, scatter plots and box plots are also used for the data analysis. The online tool at https://www.statskingdom.com/ was used for the various analysis. The essence of the analysis is to preprocess the dataset for application in analytical or machine learning modelling of the CKGHI dataset. The study also presents different options for the choice of the solar radiation value that can be used in the solar power system design to satisfy a given load profile based on a given performance specification [18, 19].

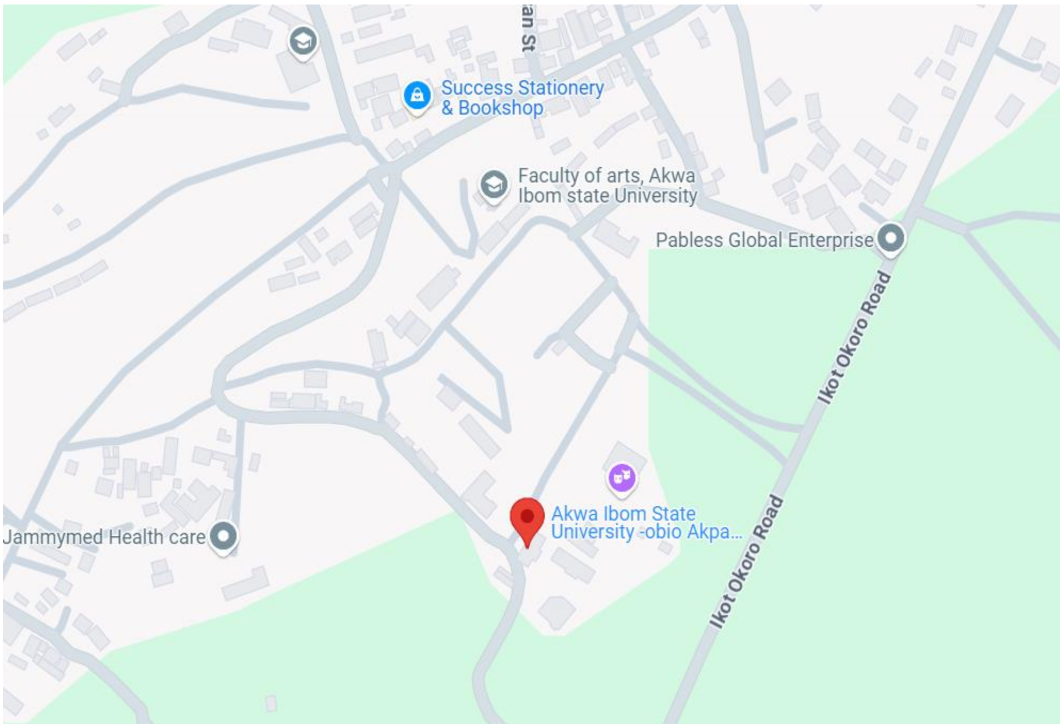

Figure 1: The sunshine duration was finally defined as the period during which direct solar irradiance exceeds a threshold value of 120 W/m^2 pdf

2.2 Determination of the Outliers

One of the factors that affect data utilization for modelling purposes is outlier data. The outliers need to be identified and possibly removed from the dataset before the dataset is employed in the modelling process. In this study, the outliers are identified using the Z-score method. The Z-score method requires the mean and the standard deviation of the dataset along with a value of k for setting the upper and lower limit of acceptable data values. The mean, $\bar{X}$of the dataset is given as Equation (1).

$$\bar{X} = \frac{\sum_{j=1}^{j=n}(x_j)}{n} \tag{1}$$

The standard deviation, s of the dataset is given as Equation (2).

$$s = \sqrt{\left(\frac{\sum_{j=1}^{j=n}(\bar{X}-x_j)^2}{n-1}\right)} \tag{2}$$

Given the value of k, the upper limit for the outlier, XU, is given as Equation (3).

$$XU = \bar{X} + k(s) \tag{3}$$

The lower limit for outlier, XL is given as Equation (4).

$$XL = \bar{X} - k(s) \tag{4}$$

The outliers base on Z-score is then conducted using the procedure given in algorithm 1.

ALGORITHM 1

Step 1: Obtain the n data items, x_j where j = 1, 2, 3, …, n
Step 2: Determine the mean, X ̅as given in Equation (1)
Step 3: Determine the standard deviation, as given in Equation (2)
Step 4: Set the value of k ; in this case, k =3
Step 5: Determine the upper limit for outlier, XU, as given in Equation (3)
Step 6: Determine the lower limit for outlier, XL as given in Equation (4)
Step 7: Determine the outliers, 〖XO〗_i where;
Step 7.1: i = 1
Step 7.2: j = 1
Step 7.3: If (x_j>XU) or (x_j<XL)Then
〖Step 7.4: XO〗_i=x_j
Step 7.5: Output, i, 〖XO〗_i

Step 7.6: i = i +1
Step 7.7: GOTO Step 7.8
Step 7.8: Else Endif
Step 7.9: j = j +1
Step 7.10: If (j≤n)Then
Step 7.11: GOTO Step 7.3
Step 7.12: Else
Step 7.13: Endif
Step 7.14: End

2.3 Computation of the Moving Average

In the design of solar power system, the annual mean value of the solar radiation is normally selected. However, the annual mean value seems to be high causing excess loss of load in those months in the raining season will low solar radiation values. In order to avoid loss of load, some solar power system designers use the minimum solar radiation value which can be selected from the minimum daily or monthly solar radiation values.

Analytically, the moving average can be used to determine the minimum solar radiation values that can be used for a given days of power autonomy. In this case, for a three-day power autonomy, 3 point moving average of the solar radiation dataset can be used and then determine the minimum of the 3-point moving average solar radiation data. The expression for the k point moving average computed on x_j solar radiation dataset, where j = 1, 2, 3, …, n is given as follows in Equation (5).

$$\overline{Xmv(K)} = \frac{\left(x_j+x_{j+1}+x_{j+2}+\cdots,+x_{j+(K-1)}\right)}{k} \text{ for } 1, 2, 3, \ldots, n, k = 1, 2, \ldots, n \quad (5)$$

After the computation of the moving average for selected number of K values, some descriptive statistical parameters are determined in respect of $\overline{Xmv(K)}$. Some of the statistical parameters considered are the minimum, maximum, range , mean , standard deviation among others.

3. RESULTS AND DISCUSSION

3.1 The Descriptive Statistics Results for the Daily Mean CS-GHI Dataset

The results of the descriptive statistics conducted on the 21 years (2003 to 2023) daily mean Clear sky Global Horizontal Irradiance (CS-GHI) dataset is shown in Table 1. The line graph is also shown in Figure 2. The results showed a 21 year daily mean value of 6.2549 kW-hr/m^2/day with 0.4961 standard deviation.

Table 1: Some descriptive statistics parameters for the daily mean CS-GHI Dataset

S/N	Groups	Daily mean CS-GHI(kW-hr/m^2/day)
1	Number of observations	7,670
2	Number of missing values	0
3	Minimum	3.96
4	Maximum	7.15
5	Range	3.19
6	Mean (x̄)	6.2549
7	Standard Deviation (S)	0.4961
8	Q1	5.91
9	Median	6.33
10	Q3	6.65
11	Interquartile range	0.74
12	Skewness	-0.6842
13	Excess kurtosis	0.1111

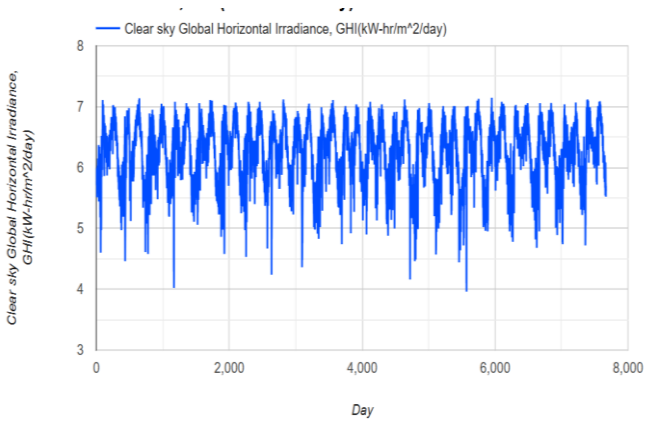

Figure 2: The line graph of 21 years (2003 to 2023) daily mean CS-GHI dataset

3.2 The Results of the Box, the Outlier's Determination and the Histogram Chart for the Daily Mean CS-GHI Dataset

The box and whisker plot in Figure 3 shows that there are outliers. The outliers are determined using the Z-score method and the results are presented in Table 2 and Figure 4. The Z-score method gave upper fence value as 7.15 and the lower fence value as 4.8. The results in Figure 3 and Figure 4 show that all the potential outliers identified occurred in the values below the lower fence of 4.8 and the outliers ranged from 3.996 to 4.76. Although these values are identified as potential outliers, they may not be discarded during data analysis. They show the extreme lower values that do occur in certain seasons in the case study area which points to the fact that design values for the solar radiation should accommodate these extreme low values that are seen as possible outliers.

In addition, the box plot confirmed the negative skewness (-0.6842 shown in Table 1) of the data, which means that there are higher values above the mean than the smaller values below the mean. The predominance of the higher values above the mean (6.2549) is highlighted by the histogram of the dataset, as shown in Figure 5.

In essence, the case study site is predominantly high on the solar radiation above the mean value. However, based on the outlier results, there are extreme low values that are about 40 in number which amounts to 0.052% of the total data records.

Table 2: The potential 40 outliers identified using the Z-score method

Row/ Column	1	2	3	4	5	6	7	8	9	10	Summary of outlier results	Values
1	4.67	4.6	4.46	4.61	4.58	4.51	4.29	4.46	4.02	4.22	Number of observations	40
2	4.58	4.53	4.67	4.24	4.36	4.76	4.48	4.61	4.76	4.74	Minimum	3.96
3	4.36	4.16	4.6	4.74	4.46	4.63	4.49	4.57	4.64	4.44	Maximum	4.76
4	4.59	4.64	4.72	4.1	4.22	3.96	4.68	4.74	4.72	4.76	Range	0.8

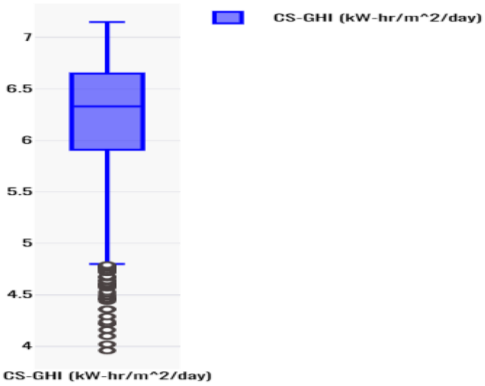

Figure 3: The box and whisker plot including outliers for the daily mean CS-GHI dataset

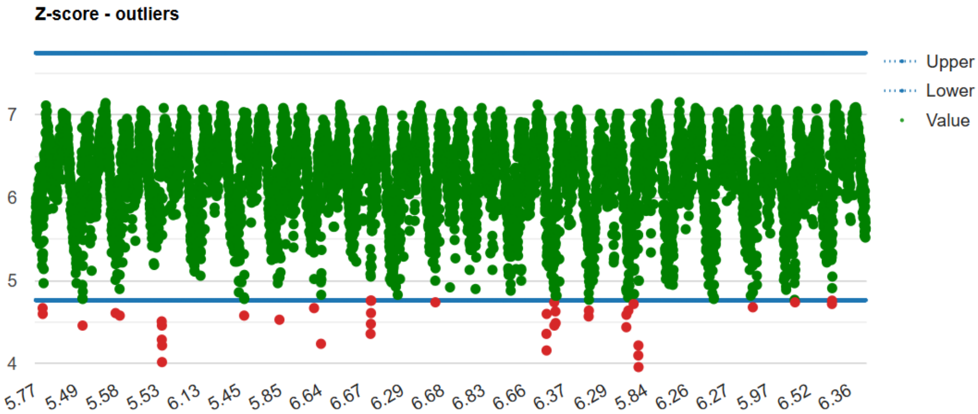

Figure 4: The scatter plot showing the normal daily mean CS-GHI points in green with the potential outliers in red

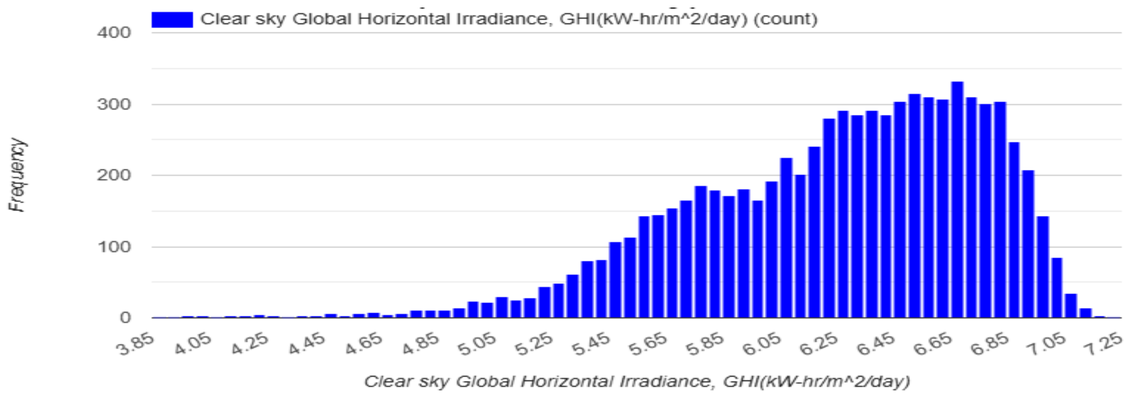

Figure 5: The histogram of the daily mean clear sky global horizontal irradiance

3.3 The Results of the Moving Average at Different Time Span and the Corresponding Outliers

The results of the outlier evaluated for moving average computed at different time scales of 1, 3, 5, 7, 30 and 365 days are shown in Table 3. The scatter plot with line for the lower boundary (lower fence for outlier) and the upper boundary (upper fence for outlier) versus the number of points considered in the moving average is shown in Figure

6 whereas the line graph of the same result is shown in Figure 7. Furthermore, the line graph for number of outliers for the moving average at the various selected time span of 1, 3, 5, 7, 30 and 365 is presented in Figure 8.

The results showed that as the time span for the moving average increases from 3, 5, 7 and up to 365 days, the lower boundary value for outlier increases whereas the upper boundary values decrease. The two values tend towards the mean daily value of the entire dataset which is about 6.2549. This is evident in the scatter plot in Figure 6 and the line graph in Figure 7. In Figure 8, it can be seen that for each moving average time span, the number of outliers decreases as the time span of the moving average increases. The outlier number by the Z-score method dropped to zero with the moving average time span of 30. The scatter plot showing the outlier count for the different moving average time span are shown in Figure 9 for the 3-point moving average, in Figure 10 for the 5-point moving average, in Figure 11 for the 7-point moving average, in Figure 12 for the 30-point moving average, and in Figure 13 for the 365-point moving average.

Table 3: The results of the outlier evaluated for moving average computed at different time scales of 1, 3, 5, 7, 30 and 365 days

Parameter	**CS-GHI for 1 point moving average**	**CS-GHI for 3 point moving average**	**CS-GHI for 5 point moving average**	**CS-GHI for 7 point moving average**	**CS-GHI for 30 point moving average**	**CS-GHI for 365 point moving average**
Number of observations	7,670	7,670	7,670	7,670	7,670	7,670
Minimum	3.96	4.09	4.3	4.6814	5.2083	6.1415
Maximum	7.15	7.08	7.046	7.0414	6.943	6.372
Range	3.19	2.99	2.746	2.36	1.7347	0.2305
Mean ($\bar{x}$)	6.2549	6.2549	6.2548	6.2548	6.2545	6.2562
Standard Deviation (S)	0.4961	0.4747	0.4629	0.4545	0.4112	0.05036
Q1	5.91	5.91	5.908	5.91	5.9107	6.2192
Median	6.33	6.32	6.323	6.3221	6.31	6.2676
Q3	6.65	6.65	6.644	6.64	6.6007	6.2838
Interquartile range	0.74	0.74	0.736	0.73	0.69	0.06458
Lower boundary (lower fence for outlier)	4.7666	4.8308	4.8662	4.8914	5.0211	6.1051
Upper boundary (upper fence for outlier)	7.7433	7.679	7.6434	7.6181	7.4879	6.4073
Number of outliers	40	35	23	13	0	0
Number of outliers with respect to 4.7666 and 7.7433 boundaries	40	28	10	2	0	0

The essence of studying the outliers below the lower boundary is to identify those instances where there is likely going to be low solar radiation such that the energy yield will not be adequate to sustain the desired days of power autonomy. In this case, with 3 point moving average, it is expected that the derived solar radiation based on the 3-points are values that can be used sizing of the solar power for three days power autonomy. The incidence of outliers below the lower boundary means that those data points identified as outlier are too low that the three days power autonomy may not be achieved at those days. Hence, the use of the outlier and moving average analysis can help in developing heuristic mechanisms for selecting the appropriate solar radiation value that is suitable for a given day of power autonomy

.

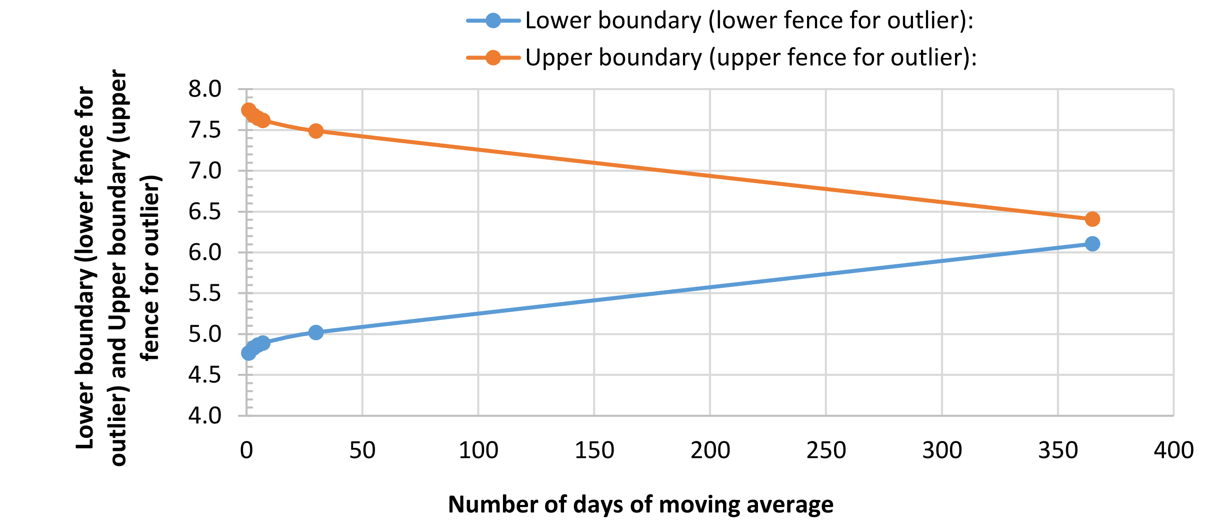

Figure 6: The scatter plot with line for the lower boundary (lower fence for outlier) and the upper boundary (upper fence for outlier) versus the number of points considered in the moving average

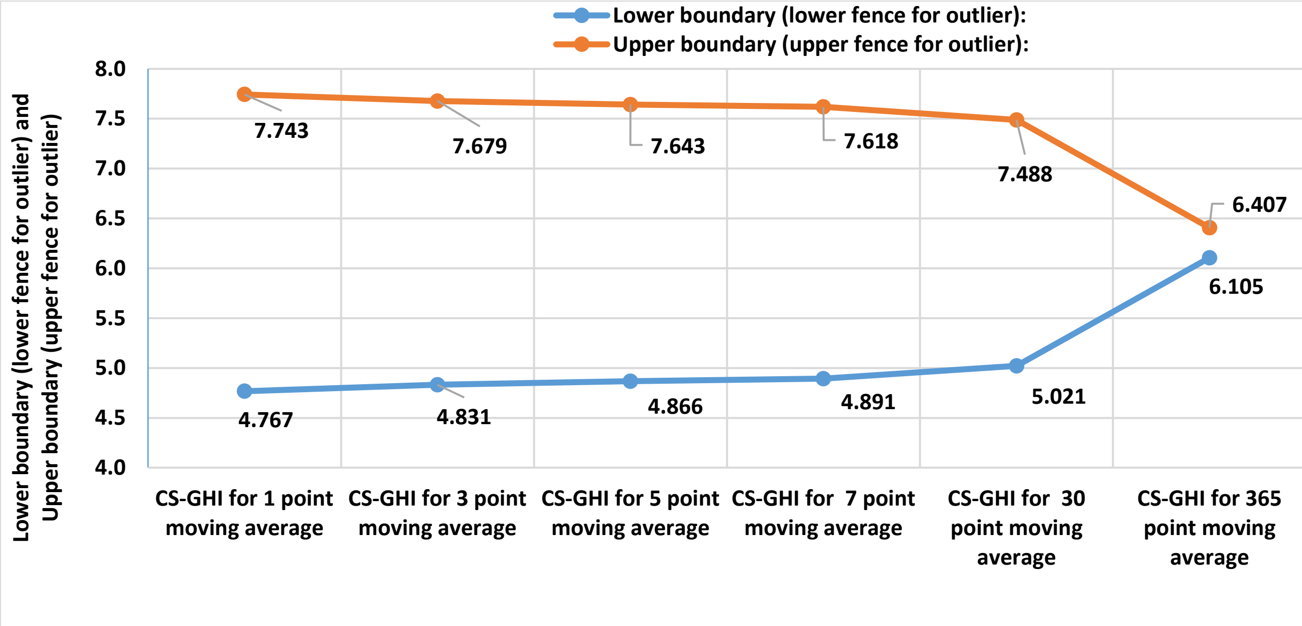

Figure 7: The line graph for the lower boundary (lower fence for outlier) and the upper boundary (upper fence for outlier) versus the number of points considered in the moving average

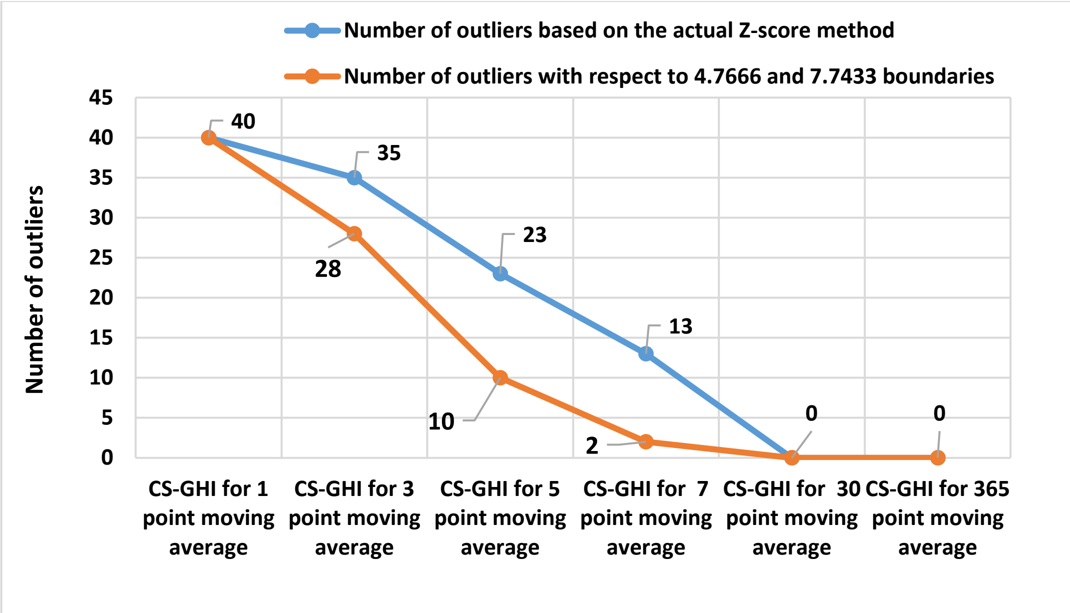

Figure 8: The line graph for number of outliers for the moving average at the various selected time span of 1, 3, 5, 7, 30 and 365

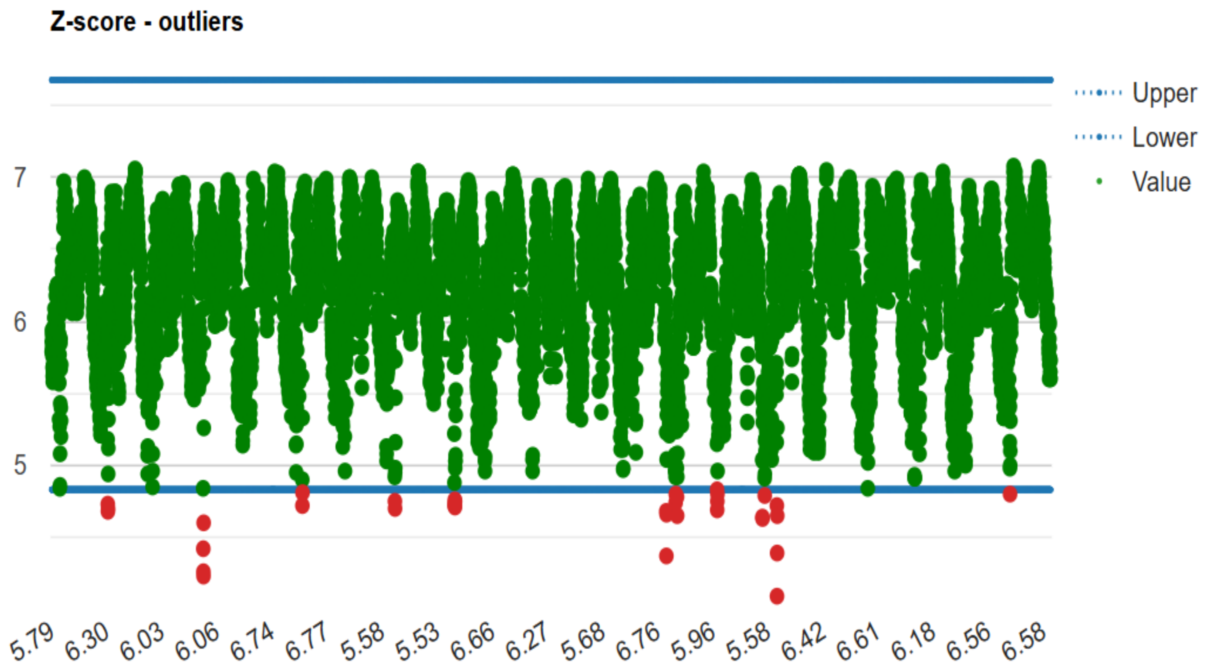

Figure 9: The scatter plot showing the 3-point moving average mean CS-GHI points in green with the 35 potential outliers in red

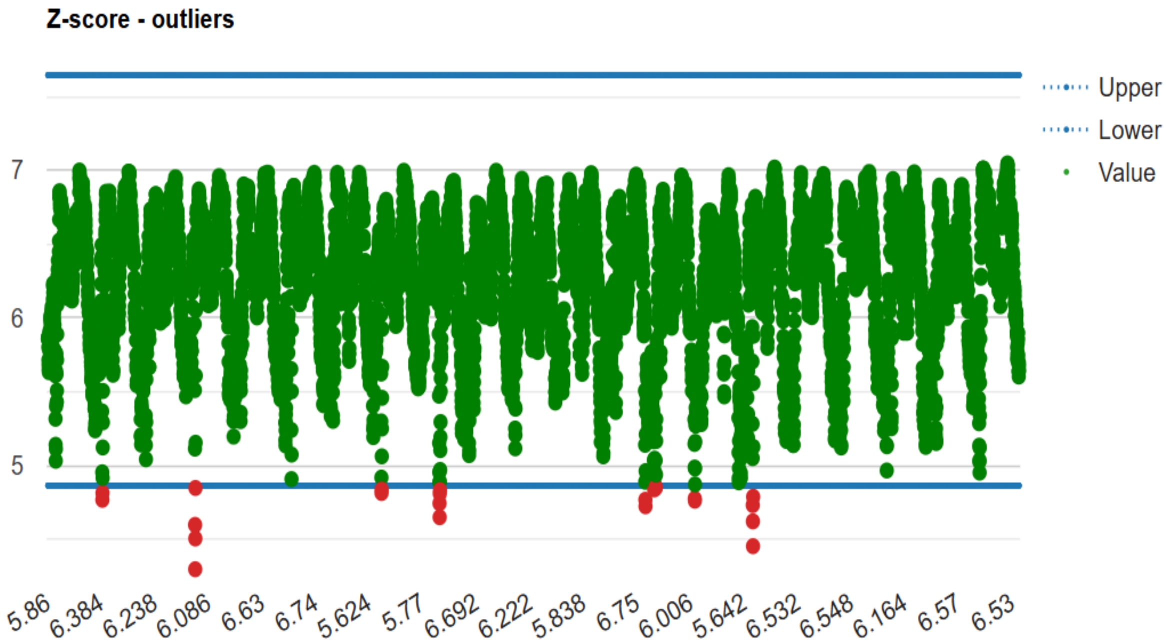

Figure 10: The scatter plot showing the 5-point moving average mean CS-GHI points in green with the 35 potential outliers in red

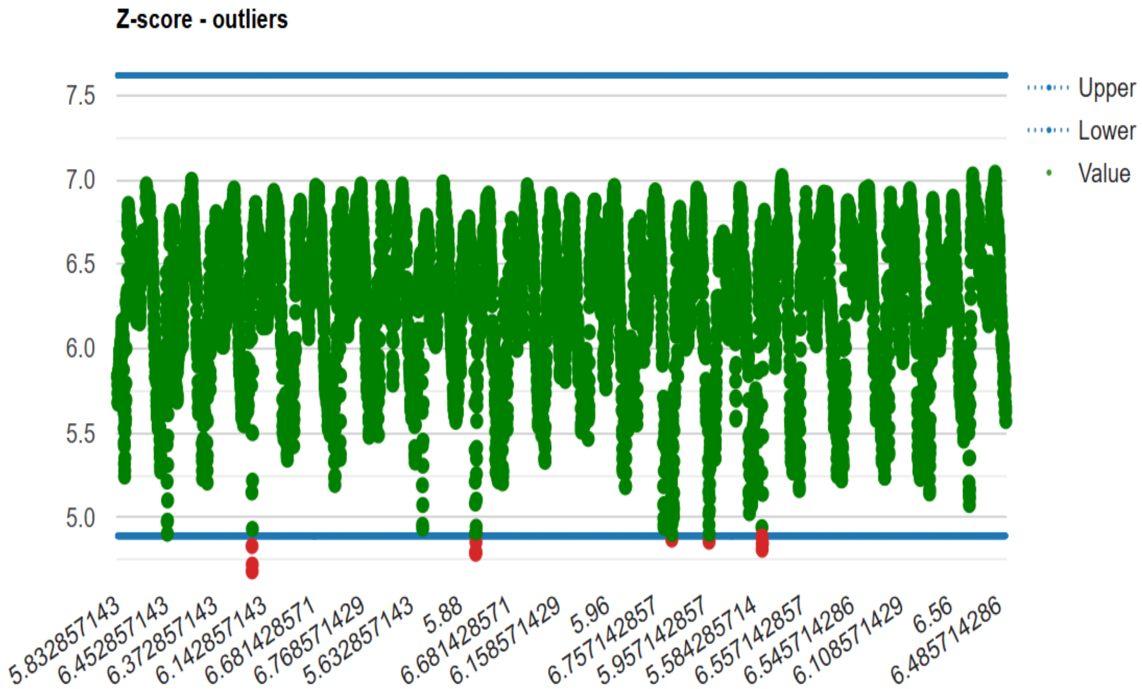

Figure 11: The scatter plot showing the 7-point moving average mean CS-GHI points in green with the 35 potential outliers in red

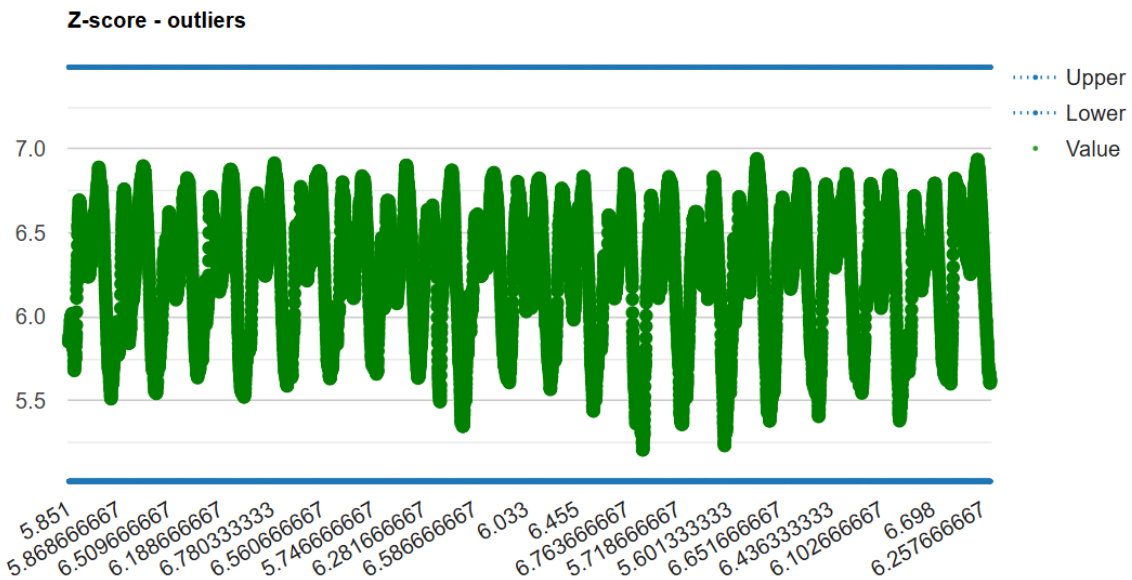

Figure 12: The scatter plot showing the 30-point moving average mean CS-GHI points in green with the 35 potential outliers in red

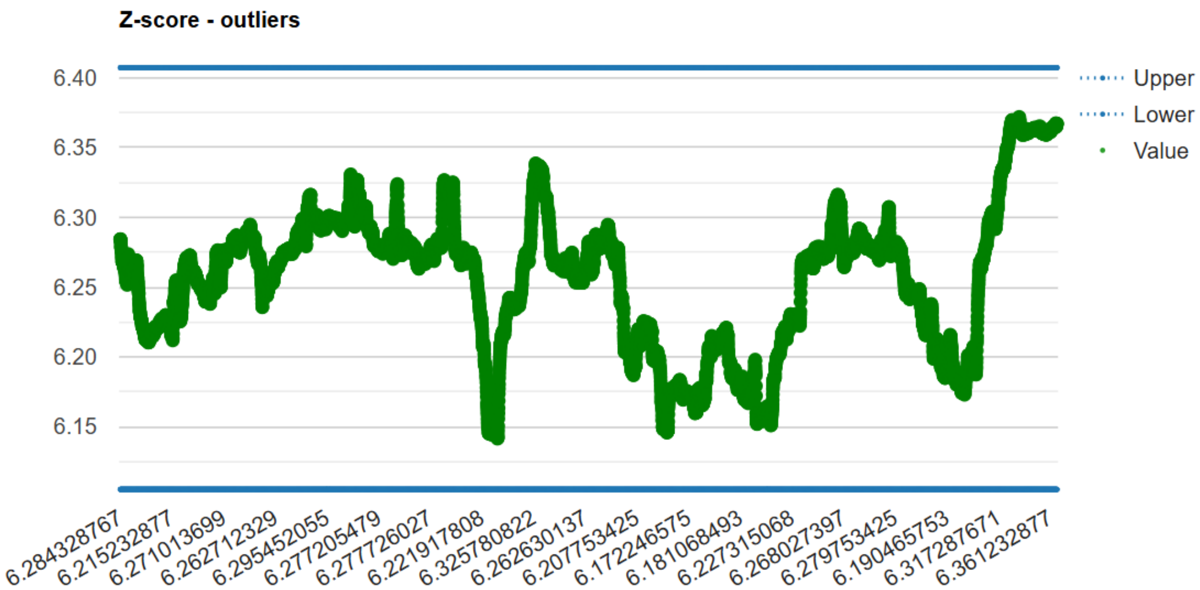

Figure 13: The scatter plot showing the 365-point moving average mean CS-GHI points in green with the 35 potential outliers in red

4. CONCLUSION

The statistical analysis of global solar radiation incident on a horizontal surface on earth under clear sky condition is presented. The analysis is meant to determine the various statistical parameters of the dataset which will aid in the selection of the most suitable solar radiation value for use in the solar power system sizing. The study is based on the daily mean solar radiation values acquired over a period of 21 years. The data spread, distribution and outliers are studied and the moving average is computer for different time scales to examine the possible days of loss for any given days of power autonomy. In all, the results showed that with the use of moving average and outlier determination, it is possible to develop heuristic mechanisms that can be used to select suitable solar radiation value for sizing solar power system for that the desired loss of load specification can be achieved for the given days of power autonomy.

REFERENCES

[1]. Adebayo, A. V., and Ainah, P. K. (2024). Addressing Nigeria's Electricity Challenges: Past, Present, And Future Strategies. *American Journal of Applied Sciences and Engineering*, *5*(2), 1-16.

[2]. Ukommi, U., Okon, B. B., Awaka-Ama, J., Umoette, A., and Ubom, E. A. (2024). Design and Performance Analysis of 500KVA Pumped-Water-Energy Storage Solar Integrated Power Plant. *ABUAD Journal of Engineering Research and Development*, *7*(1), 252-259.

[3]. Anyanime Tim Umoette, Emmanuel A. Ubom, Ibiangake Etie Akpan. Comparative Analysis of Three NOCT-Based Cell Temperature Models.*International Journal of Systems Science and Applied Mathematics.* Vol. 1, No. 4, 2016, pp. 69-75. doi: 10.11648/j.ijssam.20160104.16

[4]. Silas A. F., Umoette A.T., and Ekanem B.D. "Energy Analytical Model for The Selection of Peak Sun Hour and Days of Power Autonomy for PV Solar Power System Components Sizing". Journal of Multidisciplinary Engineering Science and Technology (JMEST) ISSN: 2458-9403 Vol. 11 Issue 3, March – 2024

[5]. Chanchangi, Y. N., Adu, F., Ghosh, A., Sundaram, S., and Mallick, T. K. (2023). Nigeria's energy review: Focusing on solar energy potential and penetration. *Environment, Development and Sustainability*, *25*(7), 5755-5796.

[6]. Ukommi, U. S., Okon, B. B., and Awaka-Ama, J. (2024). Analysis Of 400KW Photovoltaic Plant Array Layout Design with Fixed Tilt Angle and Internal Shading Mitigation Row Spacing. *Journal of Multidisciplinary Engineering Science and Technology (JMEST)*, *11*(3), 16948-16957.

[7]. Festus, U., Ukommi, U., and Ubom, E. (2023). Real-Time control of solar pv system by adaptive extremum seeking technique. *International Multilingual Journal of Science and Technology (IMJST)*, *8*(2), 6903-6911.

[8]. Soomar, A. M., Hakeem, A., Messaoudi, M., Musznicki, P., Iqbal, A., and Czapp, S. (2022). Solar photovoltaic energy optimization and challenges. *Frontiers in Energy Research*, *10*, 879985.

[9]. Umoette A.T. , Udofia Joshua Paul and Emmanuel A. Ubom, Energy Audit and Standalone Solar Power Generation Design for Akwa Ibom State University Main Campus. International Multilingual Journal of Science and Technology (IMJST) ISSN: 2528-9810 Vol. 8 Issue 6, June – 2023

[10]. Kumar, R., Rajoria, C. S., Sharma, A., and Suhag, S. (2021). Design and simulation of standalone solar PV system using PVsyst Software: A case study. *Materials Today: Proceedings*, *46*, 5322-5328.

[11]. Umoette, A. T. Okoro O. I. and Davidson, I. E. "Performance Analysis of a 10hp Three phase Induction Motor Using Classical and Finite-Elements for Varying Load Conditions," 2021 IEEE PES/IAS Power Africa, 2021, pp.1-5.

[12]. Uche, C. K. A., Udoka, E. V. H., Kisakye, A., Maxwell, K. F., and Wisdom, O. O. (2023). Design of a Solar Powered Water Supply System for Kagadi Model Primary School in Uganda. *Journal of Engineering, Technology, and Applied Science (JETAS)*, *5*(2), 67-78.

[13]. Khamisani, A. A. (2019). Design methodology of off-grid PV solar powered system (A case study of solar powered bus shelter). *Goolincoln Avenue Charleston, IL: Eastern Illinois University*.

[14]. Umoette A.T. Friday S.A, Clement . B., Clearness Index–Based Computation and Evaluation of Mean Daily Insolation and Optimal Fixed Tilt Angle for PV Installation In Uyo, Akwa Ibom State International Multilingual Journal of Science and Technology (IMJST) ISSN: 2528-9810 Vol. 9 Issue 3, March – 2024

[15]. Al-Shamani, A. N., Othman, M. Y. H., Mat, S., Ruslan, M. H., Abed, A. M., and Sopian, K. (2015). Design and sizing of stand-alone solar power systems a house Iraq. *Design and Sizing of Stand-alone Solar Power Systems A house Iraq Ali*, 145-150.

[16]. Kwarikunda, N., and Chiguvare, Z. (2021). Performance analysis of clear sky global horizontal irradiance models: Simple models adapted for local conditions. *Journal of Renewable Energy*, *2021*(1), 4369959.

[17]. Kamil, R., Garniwa, P. M., and Lee, H. (2021). Performance assessment of global horizontal irradiance models in all-sky conditions. *Energies*, *14*(23), 7939.

[18]. Abubakar, I. M., and Idi, B. Y. (2024). Statistical analysis of NASA POWER meteorological data for the assessment of climate variability in Adamawa State. *Environmental Technology and Science Journal*, *15*(2), 119-129.

[19]. Gebeyhu, B. (2024). Power Data Access Viewer-Based Meteorological Drought Analysis and Rainfall Variability in the Nile River Basin. *Advances in Meteorology*, *2024*(1), 9985773.