Open Data Quality Assessment Behaviour and a Boomerang Effect Investigation using the SAURAN Weather Station Network

Donald Fitzgerald
Center for Renewable and Systainable Energy Studies, Faculty of Engineering,
Stellenbosch University
Stellenbosch, South Africa
don@sun.ac.za

Bernard Bekker
Electrical and Electronic Engineering Department
Stellenbosch University
Stellenbosch, South Africa
bbekker@sun.ac.za

***Abstract*— Accessible high-quality weather data is important in enabling renewable energy and climate-based research in Africa, and across the world. However, caution must be extended to open data sets with respect to data quality, as unchecked poor-quality data produces poor results.**

This paper investigates the data quality assessment behavior of researchers that have used data acquired from the South African Universities Radiometric Network (SAURAN). This is done by performing a review of related published literature over the last eight years. A review is also done on international open weather data platforms to establish the best practices worldwide.

The results show that although SAURAN data quality is variable over time, only 25% of publications in the last eight years mention quality checking on data used. A case study is then done by reviewing a 2022 publication which makes direct use of poor-quality SAURAN data to produce poor-quality results and conclusions.

Keywords—data cleaning, data quality assessment, research behavior, renewable energy

I. Introduction

The renewable energy industry is growing rapidly, aligned with efforts to mitigate the climatic challenges arising from traditional fossil fuel based power generation. This is especially relevant in Southern Africa, which is currently facing adequacy problems in electricity supply. Supporting research in the field of renewable energy is imperative in addressing these challenges.

Accessible high-quality weather data is important in enabling renewable energy and climate-based research in Africa, and across the world. However, caution must be extended to open data sets with respect to data quality, as unchecked poor-quality data produces poor results. This is specifically applicable to researchers in developing African countries who are typically more reliant on open data due to funding limitations, and struggle with challenges in energy generation and availability.

One initiative developed through a collaboration between the Centre for Renewable and Sustainable Energy Studies (CRSES) at Stellenbosch University and the Group for Solar Energy Thermodynamics (GSET) at the University KwaZulu-Natal is the Southern African Universities Radiometric Network (SAURAN) [1], which provides high-resolution, ground-based solar radiometric and meteorological data across the Southern African region through an online open data model. The mandate of this initiative is to increase the availability of high-quality data in Southern Africa with the objective of boosting research in relevant fields such as renewable energy. However, since SAURAN is published as open data, there is no guarantee of data quality. The data is partially assessed rather than fully assessed, which will be explained in further detail later in this study. Although there is information on data quality assessments and related tools available on the SAURAN web application, the responsibility of data quality checking falls on the researchers downloading the data, which is thoroughly explained on the application.

This study firstly establishes a snapshot of the historical data quality of the SAURAN network, with a description of the network and the challenges it experiences that affect quality. The data quality assessment behaviour of researchers is then investigated by performing an online review of related published literature; and by looking at the download history of the quality checking resources on the SAURAN web application. This leads to a discussion of the possible boomerang effect (an intervention that has the opposite effect to what was intended) [2] that SAURAN could be unintentionally causing. Finally, the investigation will look at the best practices of other open data available online, specifically with respect to data quality control.

This paper is valuable for researchers, policymakers, and product owners of open databases both worldwide and especially in developing African countries. It is imperative that interventions such as SAURAN be monitored extensively after implementation to avoid the boomerang effect. There needs to be improved communication or accessibility to information relating to data quality control between the product owners and the researchers so that research in related fields provides policy makers with more accurate results, allowing them to make better decisions, improving things such as the healthy uptake of clean energy.

II. SAURAN Description and Data Snapshot

The SAURAN program was initiated with the objective of making high-resolution, ground-based weather data freely available from stations located across the Southern African region, including South Africa, Namibia, and Botswana. Its stations provide high quality radiometric data, using state of the art Kipp & Zonen radiometers, as well as a range of meteorological data. Data is available in minute, hourly and daily time averaged intervals through an online web application (https://sauran.ac.za). At the time of this study, there are ten stations recording live data and 14 more offering historical data.

2023 IEEE PES/IAS PowerAfrica | 979-8-3503-3755-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/POWERAFRICA57932.2023.10363266

979-8-3503-3755-6/23/$31.00 ©2023 IEEE

All stations (apart from one that has since been decommissioned) include a Kipp & Zonen Solys 2 solar tracker, fitted with two Kipp & Zonen CMP11 pyranometers and one CHP1 pyrheliometer. These instruments provide measurements of Global Horizontal Irradiance (GHI), Diffuse Horizontal Irradiance (DHI) and Direct Normal Irradiance (DNI). Most stations also include a mix of meteorological data such as air temperature, barometric pressure, total rainfall, relative humidity and wind speed and direction. Instruments are subject to regular maintenance, being calibrated every two years, and all glass domes and windows are cleaned multiple times each week to maintain data quality.

The product owner of SAURAN is the CRSES, while station maintenance and data collection are outsourced to GeoSUN Africa (GA). GA establishes stations and measured data is pushed to their database, which they then use to do primary data assessment to evaluate if the stations are operating effectively. They may then flag potential station concerns on an online ticketing system. CRSES staff receive these tickets and contact the station hosts to help resolve them. The data is then pulled to a database at Stellenbosch University (where the CRSES is based), which also hosts the SAURAN web application and thus is available for download. Data is collected by GA and pulled to Stellenbosch University 3 times a day. GA also handles instrument calibration and more technical station fixes.

One major challenge to the SAURAN program is limited funding, as no revenue is generated from the data itself or from advertising on the web application. The program therefore relies on grant funding, the stations themselves being hosted voluntarily with most hosts being incentivised by an interest in the data at their location. Some hosts are renumerated a small monthly fee for the regular cleaning and maintenance of the station instruments, to help increase maintenance regularity (currently only two of the ten).

Experience has shown that dynamic instruments, such as the solar tracker, tend to encounter more frequent problems. Most commonly the trackers tend to lose alignment with the sun, resulting in incorrectly measured DNI and DHI values. This is due to weather (high winds or hail), faulty instrumentation or human / animal interference. GA flags misaligned trackers by monitoring the DNI values: if it deviates too far from what is expected, a ticket is logged and the CRSES contacts the station host to check the station's alignment. Other major challenges with the stations include power loss, station communication loss and instrument failure due to weather (corrosive salt build up or lightning strikes) or due to old age. Since these stations are located far apart from each other (Fig. 1), and the hosts are not equipped to troubleshoot and fix major station issues, these problems can result in the station going down for extended times (data loss) or failing to track the sun for extended times (incorrectly measured data), decreasing the quality of the data recorded.

The quality of the data is thus variable over time, and due to the preliminary checks GA does on the SAURAN data, some level of quality assessment is done. For the purposes of this study, four tiers of data quality control are defined:

1. **Raw:** data that is not assessed or assured at all. This data is just provided as is.
2. **Partially assessed:** this data includes some preliminary data quality checks but does not provide flags for missing or poor-quality data. It is the user's responsibility to check the data and replace poor quality data with secondary or synthetic data.
3. **Assessed:** data that includes flags which highlight questionable data points for the user. The user can then easily replace this data at their discretion.
4. **Assured:** this data has already been checked for quality and any questionable data has been replaced. The data provider assures the user that the datasets don't have missing or poor-quality data.

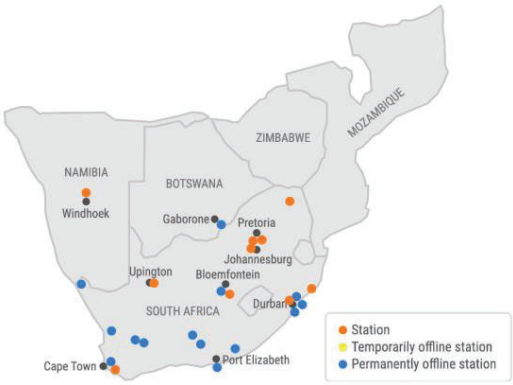

Fig. 1. SAURAN weather station map showing the current and historical weather station locations (https://sauran.ac.za/)

A challenge with quantifying the exact quality of the data collected by SAURAN stations is that there is nothing to compare the data to. Although not perfect, there is a way to quantify the radiometric data quality by assessing the DNI error. Out of the three parameters measured by the solar tracker, GHI and DHI mostly rely on the tracker being level and the sensors clean to measure good quality data, while the DNI value is also highly dependent on the tracker pointing directly at the sun. This naturally results in the GHI and DHI values being of a higher quality than the DNI measurement. Therefore, the GHI and DHI measurements are used to calculate the expected DNI value (DNI_{calc}) using the sun's zenith angle (θ) as shown in (1):

$$DNI_{calc} = \frac{GHI - DHI}{cos(\theta)} \quad (1)$$

The DNI_{calc} can then be compared to the measured DNI to form the error parameter DNI_{error} as shown in (2):

$$DNI_{error} = \left| \frac{DNI(measured) - DNI_{calc}}{DNI_{calc}} \right| \quad (2)$$

This parameter is an indicator of radiometric data quality. This error point is constantly monitored, and if it exceeds 4% [3] [4], GA logs a ticket and the CRSES contacts the station hosts to go and check the tracker alignment. The DNI_{calc} is included in the data downloaded by the user, and therefore the SAURAN data is classified as partially assessed.

Fig. 2 shows the monthly average of DNI_{error} for all the currently active stations over the last five years, with the 4% error line shown as the dotted red line. Datapoints were only considered for strong daylight hours between 10:00 and 15:00, non-low light conditions above 5 W/m^2, and months that are not missing more than half of their respective datapoints. This

strongly illustrates the importance of performing data quality checking on any data downloaded from the SAURAN network before implementing it in research. Other than variation in DNI quality, the figure also shows gaps in data, which would need filling by secondary or synthetic data.

For example, the large orange spike seen in Fig. 2 at the Namibia University of Science and Technology station was due to a solar tracker breakdown. The tracker stopped tracking the sun, but continued to measure radiometric data, resulting in incorrect readings. Due to a lack of funding and the replacement tracker being away for maintenance, the issue took about five months to fix.

The SAURAN program has established initiatives to both inform and aid researchers in performing quality assessments on their data. In May 2019, a link was added to the web application beneath each "Download" button, where a document could be downloaded that detailed the importance of quality checking, as well as how to perform this manually in MS Excel. In October 2021, this link was improved to include a more detailed document, as well as an online tool that was designed to perform some quality assessment calculations on SAURAN data fed in by researchers. To further focus researchers' attention towards quality checking, a "terms of use" tick box was made mandatory to the activation of each download button, of which the "terms of use" again states the importance of quality checking and provides both the document and the online tool.

Website analytics over the past year were used to investigate the number of unique downloads of the quality checking resources, which is shown to be significant (Fig. 3).

III. Publication Scrape

Various related online databases, such as Google Scholar, Scopus, and the Stellenbosch University Library Database, were used to identify publications that mention "SAURAN" since 2015. This search yielded 114 publications, only two of which access could not be gained to. The following attributes were identified in the remaining 112 publications:

1) *Year published*
2) *Publishing institution*
3) *Type of publication*
4) *Was SAURAN data used directly in the research?*
5) *Is there any mention of data quality assessment?*

Out of the 112 publications, 89 made use of SAURAN data in their research, of which 77 publications were from South African institutions (87%), with the remaining 12 being international (13%). The results of this publication review are illustrated in Fig. 5. It must be noted that the publications were searched for any mention of quality assessment on the SAURAN data used, and it is possible that the researcher still performed quality checking and did not mention it in the publication. Out of the 89 publications considered, only 22 mentioned some sort of data quality assessment (25%).

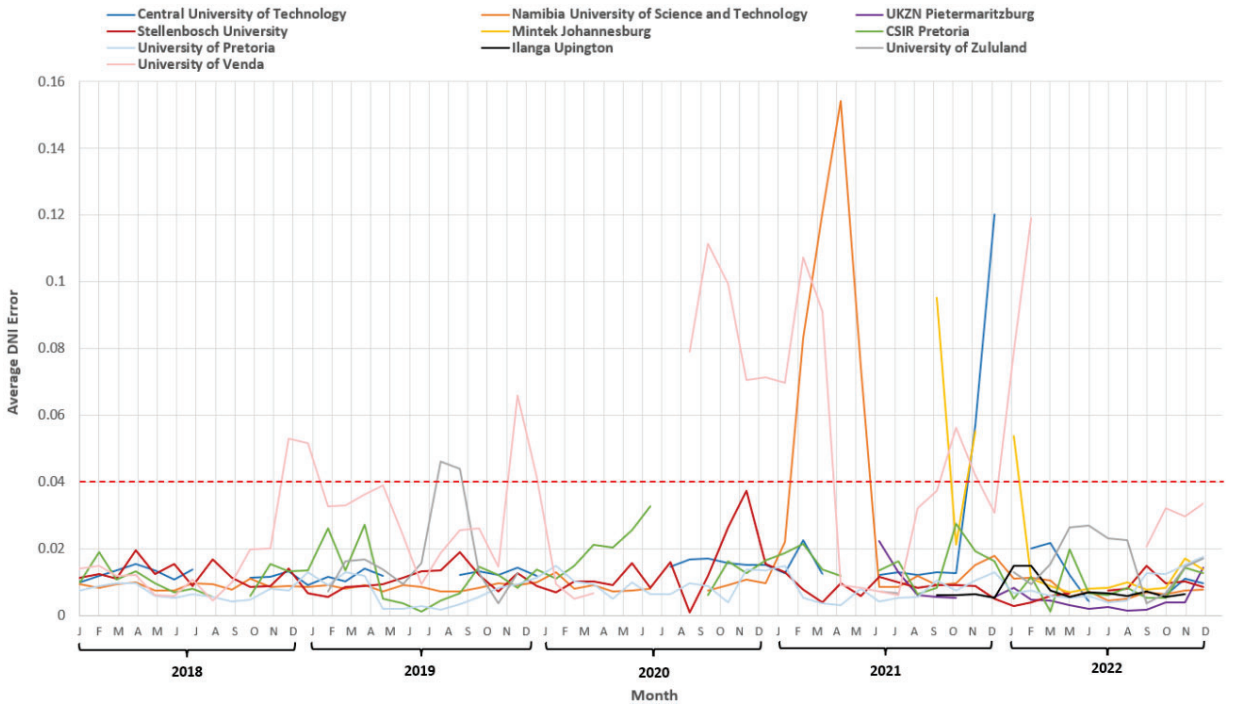

Fig. 2. Monthly average DNI_{error} (difference between measured and calculated DNI) between 2018 and 2022 for all currently active SAURAN weather stations (only considered for strong daylight hours between 10:00 and 15:00, non-low light conditions above 5 W/m^2, and months that are not missing more than half of their respective datapoints). The dotted red line represents 4% DNI_{error}, above which the station host is asked to check the solar tracker.

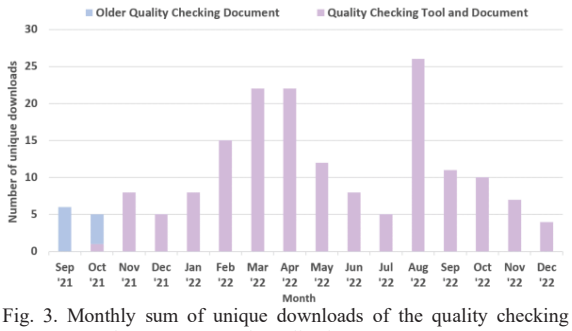

Fig. 3. Monthly sum of unique downloads of the quality checking resources on the SAURAN web application

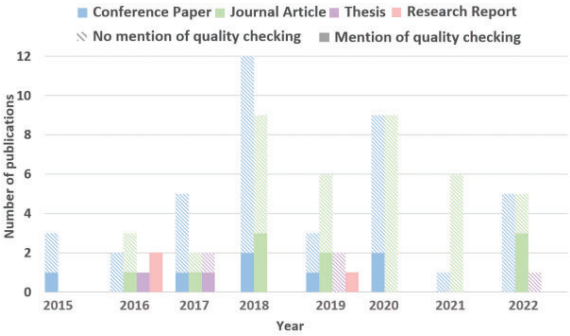

Fig. 4. SAURAN data quality assessment behavior of researchers by year and publication type, categorized by if the publication made any mention of quality checking

IV. The Boomerang Effect

In social psychology, the boomerang effect refers to when an intervention unintentionally causes the opposite result of what was intended [2]. In the case of the SAURAN case study discussed above, the intension of the initiative was to support research aimed at strengthening the renewable energy sector. The boomerang effect occurs if the network's data is of poor quality, and the researchers using the data do not perform adequate data quality assessment procedures, resulting in poor research that instead of strengthening and supporting, rather weakens the renewable industry.

As seen in this investigation, only 25% of researchers make mention of some sort of quality checking in their publications since 2015. This leaves 67 publications that may have poor quality results due to poor quality input data, some of which are peer-reviewed journal articles and theses. Whether or not this constitutes evidence for the boomerang effect is still subjective, as the exact quality of the input data of each of the 67 publications in most cases cannot be assessed.

One example, where the lack of adequate quality assessment of SAURAN data resulted in poor results is a paper presented at the Southern African Universities Power Energy Conference (SAUPEC) conference, titled: Solar Irradiation Forecasting for the City of Durban Using Time Series Analysis [5]. This conference paper is currently published on the Institute of Electrical and Electronics Engineers (IEEE) Xplore web portal. The author's objectives were to analyse the SAURAN station data between 2016 and 2019, use mathematical models to achieve irradiance forecasting, and draw conclusions on the state of solar irradiance in Durban for solar power resource potential. The data used in this study was downloaded from the SAURAN web application and focused on daily averaged GHI values only (shown in Fig. 4). Although the study filled in missing data, such as the gap seen in October 2016 (Fig. 4), the data appeared to be otherwise unchanged. This station had a solar tracker malfunction and it stopped tracking midway through 2019, which is very noticeable in the data. This was a very old tracker and later the station was removed from recording live data due to equipment reaching end-of-life.

Instead of doing a quality assessment procedure on the data, which should have seen bad datapoints in the latter half of 2019 replaced by secondary or synthetic data, the researchers used it as is and directly based their conclusions on the results. The publication focuses on the forecasting of data in the latter half of 2019 and compares it to the "true" (measured) data. This leads to poor results and conclusions that are not helpful to the renewable energy sector [5].

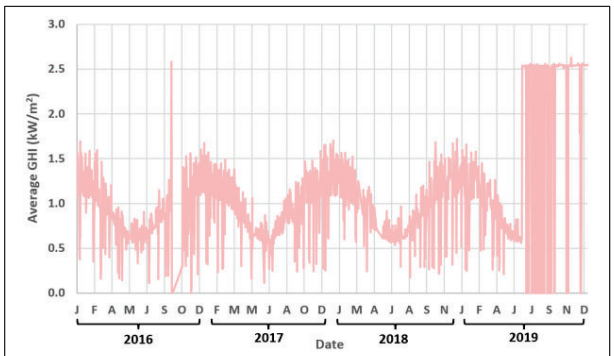

Fig. 5. Daily average GHI measured at the University of KwaZulu-Natal Howard College Campus SAURAN weather station from 2016 to 2019

V. Best Practices

Four programs that offer open radiation and/or meteorological data were investigated with respect to their data quality checking procedures to help establish an idea of the best practices for such data.

A. Baseline Surface Radiation Network

The World Radiation Monitoring Centre (WRMC) provides a central archive of the Baseline Surface Radiation Network (BSRN) [6]. This database provides open radiation measurement data from both ground stations and upper-air meteorological observations (satellite data). The network has 74 stations worldwide, with 58 currently active. The data is available in high time resolutions (1 to 3 minutes).

The website displays a disclaimer stating that the station scientists are responsible for the quality of their submitted data. The data is format-checked and visualized before it is made public. Obvious errors are not imported to the public domain but returned to the station scientists that submitted it. Although this ensures some basic data integrity, it is clearly stated that imported data is not quality checked in detail and quality flags have not been applied. The website includes the following warning banner:

"WRMC highly recommends that all users do their own quality checks of the data after extracting BSRN-data!!!"

The WRMC offers a software tool to help researchers perform quality checks on their data (BSRN-Toolbox). The website also includes an "Errata and Updates" page that lists known errors with datasets. Under the categorization of this study, the platform offers partially assessed data.

B. Australian Government – Bureau of Meteorology

The Australian Government Bureau of Meteorology provides open meteorological data through an online web application [7], having thousands of grounds stations across Australia. Some of the ground stations are automated, while others are managed by full-time staff, contract observers or volunteer observers (around 6000 volunteers).

Quality control information was difficult to find on the website. Gross errors on all datasets are checked automatically as data comes in from weather stations. They state that additional quality checking is done on low resolution data, with a priority on daily rainfall and daily maximum and minimum temperatures. These additional checks are not done on one-minute observations. Checks include "common sense" checks, consistency with nearby sites and consistency over time [7].

The Bureau has quality checking operators that monitor these data checks and attach a quality flag to each observation. Their main objective is to identify faulty instrumentation. Since these weather stations comprise of static instruments, their quality varies less than with dynamic instruments such as a solar tracker. This platform offers some data that is assessed and some that is partially assessed.

C. Surface Radiation Budget Network

Surface Radiation Budget Network (SURFRAD) was established through the support of the National Oceanic and Atmospheric Administration's (NOAA) Office of Global Programs (United States of America Government) [8]. The network includes eight ground stations across the USA, most of which were installed in the 1990's. The stations include a solar tracker, measuring infrared, DNI, DHI and GHI; as well as other static instruments measuring UVB, photosynthetically active radiation, aerosol optical depth, cloud cover, upwelling solar, upwelling infrared, temperature, pressure, relative humidity and wind speed and direction.

They state that "quality assurance built into the design and operation of the network, and good data quality control ensure that a continuous, high-quality product is released." The exact extent of quality control is not explained. They have a page that lists all problems with datasets, but they do not suggest that the user should do their own quality checking on the data acquired. This platform provides assured data.

D. Solar Radiation Monitoring Lab

The Solar Radiation Monitoring Lab (SRML) is a regional solar radiation network in the USA. It consists of four primary and ten secondary ground stations [9]. The primary stations include solar trackers with high-accuracy instruments, while the secondary station include static instruments, with higher uncertainties. The network also offers historical data for 20 stations that have since been decommissioned.

According to the website, the data is checked for errors that have been automatically flagged. After the errors have been corrected, any gaps in the data are filled in (it is not mentioned with what secondary or synthetic data). The data files are then edited to eliminate errors arising from misalignment of the tracker or ice build-up on the instruments. The finalised versions of the data are then made public. These quality checking procedures are not done on the secondary station data [9]. This platform provides assured data through their four primary stations.

VI. Conclusion

Accessible high quality ground station data is undoubtably important for the stimulation of renewable energy and climate-based research in Southern Africa. However, caution must be extended to open data with respect to data quality, as unchecked poor-quality data produces poor quality results, which may cause unintended harm to the research field.

As a case study, SAURAN is seen to offer partially assessed open data, which still requires a full quality assessment to be done before implementation in research. However, it is shown that only 25% of publications in the last eight years mention quality checking on the SAURAN data they have used. Even though quality checking resources are accessible on the web application, and downloads of these resources are shown to be significant over the last year, it seems to have made no positive difference in researcher behaviour with regards to data quality control. A 2022 publication was used as a case study to show the effects of using poor quality data to yield poor research results.

Compared to other platforms that offer open weather data internationally, SAURAN does a good job in providing partially assessed data and having quality assessment resources more visible than most other platforms. This raises the question of the state of research based on open data, that is not assured, worldwide. This research is crucial for the development of African countries, who are more likely to rely on open data for research due to funding restrictions. It is imperative that interventions such as SAURAN be monitored extensively after implementation to avoid the boomerang effect [2]. There needs to be improved communication between product owners and researchers so that policy makers are provided with more accurate results, allowing them to make better decisions, improving things such as the healthy uptake of clean energy.

References

[1] M. J. Brooks, S. du Clou, J. L. van Niekerk, P. Gauche, C. Leonard, M. J. Mouzouris, A. J. Meyer, N. van der Westhuizen, E. E. van Dyk and F. Vorster, "SAURAN: A new resource for solar radiometric data in Southern Africa," *Journal of Energy in Southern Africa,* pp. 26, 2-10, 2015.

[2] A. Levy and Y. Maaravi , "The boomerang effect of psychological interventions," *Social Influence,* vol. 13:1, pp. 39-51, 2018.

[3] J. van Jaarsveldt, Interviewee, *Technologist at GeoSUN Africa.* [Interview]. 31 05 2023.

[4] A. Forstinger, S. Wilbert, A. R. Jensen, B. Kraas, C. F. Peruchena, C. A. Gueymard, D. Ronzio, D. Yang, E. Collino and J. P. Martinez, "Expert quality control of solar radiation ground data sets," in *SWC 2021: ISES Solar World Congress*, Online, 2021.

[5] S. Ntlela and I. Davidson, "Solar Irradiation Forecasting for the City of Durban Using Time Series Analysis," in *Southern African Universities Power Engineering Conference*, Durban, 2022.

[6] Alfred-Wegener-Institute, "WRMC-BSRN," 20 October 2022. [Online]. Available: https://bsrn.awi.de/.

[7] Commonwealth of Australia, "Australian Government," 20 October 2022. [Online]. Available: http://www.bom.gov.au/.

[8] National Oceanic and Atmospheric Administration, "Global Monitoring Laboratory - SOLRAD Network," 20 October 2022. [Online]. Available: https://gml.noaa.gov/grad/solrad/.

[9] UO Solar Radiation Monitoring Laboratory, "Solar Radiation Monitoring Lab," 20 October 2022. [Online]. Available: http://solardat.uoregon.edu/index.html.