Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2021 doi:10.20944/preprints202101.0545.v3

Evaluation of high-resolution satellite-derived solar radiation data for PV Performance Simulation in East Africa

Diane Palmer *[a] and Richard Blanchard [a]
[a] Centre for Renewable Energy Systems Technology, Loughborough University, LE11 3TU, UK
* Corresponding Author D.Palmer@lboro.ac.uk Tel. +44 (0)1509 635604
R.E.Blanchard@lboro.ac.uk

Abstract: Access to reliable, clean, modern cooking enhances life chances. One option is photovoltaic cooking systems. Accurate solar data is needed to ascertain to what extent these can satisfy the needs of local people. This paper investigates how to choose the most accurate satellite derived solar irradiance database for use in Africa. This is necessary because there is a general shortage of ground measurements for Africa. The solar data is needed to model the output of solar cooking systems, for instance, a solar panel, battery and electric pressure cooker. Four easily accessible satellite databases are validated against ground measurements using a range of statistical tests. Results demonstrate the impact of the mathematical measure used and the phenomenon of balancing errors. Fitting of the satellite model to appropriate climate zone and/or nearby measurements improves accuracy, as does higher spatial and temporal resolution of input parameters. That said, all the four databases reviewed were found to be suitable for simulating PV yield in East Africa.

Keywords: Solar radiation, Satellite-derived irradiance, Global Horizontal Irradiance, Clear sky model, ground stations, validation.

Nomenclature

Abbreviation	Definition
BSRN	Baseline Surface Radiation Network
CAMS	Copernicus Atmosphere Monitoring Service
CMSAF	Climate Monitoring Satellite Application Facility
ETR	Extraterrestrial irradiation
EUMETSAT	European Organisation for the Exploitation of Meteorological Satellites
GHI	Global horizontal irradiation
h	Solar Elevation
JRC	Joint Research Centre
MECS	Modern Energy Cooking Services Programme
nMBE	normalised Mean Bias Error
MERRA-2	Modern Era Retrospective Analysis for Research and Applications, Version 2
MVIRI	Meteosat Visible Infra-Red Imager
nRMSE	normalised Root Mean Error

PMCC	Pearson Product-Moment Correlation Coefficient
PV	Photovoltaic
PVGIS	Photovoltaic Geographical Information System
QC	Quality Control
SARAH	Surface Solar Radiation DataSet-Heliosat
SEVIRI	Spinning Enhanced Visible and Infrared Imager
SoNG	Solar Nano Grids
SPECMAGIC	SPECtral Mesoscale Atmospheric Global Irradiance Code
SRTM	Shuttle Radar Topography Mission
WRMC	World Radiation Monitoring Centre

1. Introduction

Accurate knowledge of incoming solar radiation at specific locations is very important for many applications. In the context of this research, it is required for modelling PV yield as input to solar cooking systems for the Modern Energy Cooking Services project (https://mecs.org.uk/). Worldwide, nearly three billion people rely on solid fuel for cooking and heating. This has health and environmental implications. Women and children especially are exposed to smoke resulting in respiratory illnesses, cataracts, heart disease and cancer. Much time and human energy is expended in firewood collection. Reliance on wood fuel contributes to climate change and local forest degradation. The Modern Energy Cooking Services Programme (MECS) is investigating how to rapidly transition from biomass to genuinely 'clean' cooking (e.g. with electricity). The aim of this ongoing research is to investigate the possibility of developing a solar power support system that can support individual electric cooking systems in off-grid situations. Such a system might comprise a solar panel, battery, and a cooking device such as an electric hob or electric pressure cooker. A detailed solar resource assessment is necessary to discover to what extent such solar enabled cooking can supply people's needs.

Europe has a relatively dense network of well-maintained weather stations which provide publicly available data. In Africa, the situation is quite different. There are proportionately few ground sensors and a dearth of accessible weather measurements (Meyer, 2015). Thus, an alternative source of information must be sought. Satellite derived radiation datasets are widely regarded as the most accurate alternative. However, not all solar datasets are created equal. Moreover, as yet there is no standardised approach for choosing the best solar irradiation dataset (Solargis, 2021).

It is difficult to select a dataset from published validation statistics. These use different locations, temporal resolutions, methods of error calculation, data filtering and data aggregation processes. Yang and Bright (2020) suggest that due to uncertainties in ground records, it is better to ask if the database under investigation is sufficient for the intended purpose, or if one dataset performs better than another, rather than relying on error and bias values.

The goal of this article is to determine which of four easily available satellite-derived global horizontal products is to be preferred for modelling PV output in East Africa. This is novel because solar satellite databases have only previously been compared in South Africa (Amillo et al., 2018) where the solar market is established.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2021 doi:10.20944/preprints202101.0545.v3

There are three specific objectives:

1. To evaluate and compare satellite solar datasets in East Africa with a view to advising which database to use where. Satellite data is mostly verified against data from the archive of the Baseline Surface Radiation Network (BSRN), based at the World Radiation Monitoring Centre (WRMC). However, there are just three BSRN monitoring stations on the continent of Africa (Algeria, Namib Desert and South Africa), as opposed 13 in the USA, and 11 in Europe. Additionally, the West has many other ground stations, which although accurate, do not belong to the BSRN network. Africa is very short of ground-based solar radiation sensors in general.
2. To establish whether the different clear sky models utilised by satellite derived solar radiation datasets affect the outcomes of the dataset values in East Africa. Clear sky models differ in complexity of algorithm, atmospheric inputs, temporal and spatial resolution of atmospheric inputs, and location where the model was fitted.
3. To compare and contrast solar satellite data with measurements from ground stations in East Africa.

The paper is organised as follows: Section 2 summarises the methods used. Section 3 describes the satellite derived global horizontal irradiation databases compared in this research. Section 4 investigates the extent to which comparative accuracy of databases can be ascertained without weather station data. Section 5 describes the ground station data obtained for comparative validation by this project. Section 6 explains and discusses the results of a multiplicity of statistical tests used to differentiate between the four satellite datasets. Finally, Section 7 presents the conclusions and main messages of this research.

2. Methodology

There are many satellite-derived solar radiation databases. For this research, up-to-date, high temporal resolution ones where required. It was also necessary to select those which cover Africa, as some are confined to India, Europe, or the USA. Suitable candidate datasets include free products e.g. Solemi (https://wdc.dlr.de/data_products/SERVICES/SOLARENERGY/description.php), available upon request. There are also paid-for services: Meteonorm (meteonorm.com), Reuniwatt (https://reuniwatt.com/en/), SoDa (http://www.soda-pro.com/web-services/radiation/helioclim-3-archives-for-pay), SolarAnywhere (https://www.solaranywhere.com/), Solargis (solargis.com), 3E (https://www.3e.eu/data-services/solar-resource-data/), and 3Tier (https://www.3tier.com/en/support/solar-online-tools/). The four databases selected for use in this paper (detailed in Section 3) were instantly downloadable and free. (Except for Solcast, which has a generous free allowance for researchers.)

The ground measurements used for validation were the only ones available to the authors during the COVID-19 pandemic when this paper was written. The ground data locations are described in Section 3.

All the database time series values were averaged or rounded to the nearest time period end so that inner database joins could be performed to enable subsequent analysis. The global horizontal irradiance values in all databases have compatible units, being recorded in Wh/m^2, except for the ground measurements from Galu and Munje (see below). These daily values were divided by 10 (the average number of daylight hours in the day in Kenya) to convert them to Wh/m^2.

The following statistical tests were used to validate the satellite data against the ground values: normalised Root Mean Error (nRMSE), normalised Mean Bias Error (nMBE), hourly average, hourly

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2021 doi:10.20944/preprints202101.0545.v3

standard deviation, trendlines, Pearson Product-Moment Correlation Coefficient, average global horizontal irradiation per hour of day, average global horizontal irradiation per day of year, and frequency distribution.

3. The satellite derived GHI databases used in this research.

Two of the solar radiation products under investigation here are produced by the Climate Monitoring Satellite Application Facility (CMSAF) of the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT). The Joint Research Centre (JRC) Photovoltaic Geographical Information System (https://re.jrc.ec.europa.eu/pvg_tools/en/tools.html) versions are used.

The first of these is SARAH, the Surface Solar Radiation DataSet-Heliosat. This data is available at hourly interval and at a spatial resolution of 0.05° (5.6 km). Extensive validation has been performed by Urraca et al. (2017). SARAH employs observations from the Meteosat Visible Infra-Red Imager (MVIRI) and the Spinning Enhanced Visible and Infrared Imager (SEVIRI) instruments carried by EUMETSAT geostationary Meteosat satellites. The Heliosat-2 algorithm is utilised. It subtracts cloud properties recorded by the satellite sensor from clear-sky irradiance. Clear-sky radiation is obtained via the SPECMAGIC method (SPECtral Mesoscale Atmospheric Global Irradiance Code—http://gnu-magic.sourceforge.net/). Inputs to SPECMAGIC comprise aerosol properties, total column water vapour and ozone, in the form of a monthly look-up table for processing speed. SPECMAGIC was fitted at two European sites.

The second is CMSAF. The data is supplied at 15 minute, hourly, daily, and monthly intervals, with 0.05° spatial resolution. CMSAF uses the same instruments, inputs, and algorithms as SARAH, but the look-up table is updated continuously with 3-hourly satellite-derived values of atmospheric inputs (Amillo et al. 2018).

The third satellite derived solar database examined is CAMS (Copernicus Atmosphere Monitoring Service) (http://www.soda-pro.com/web-services/radiation/cams-radiation-service). Temporal resolution is one minute to one month. It is spatially interpolated to the point of interest. Again, CAMS uses Meteosat / SEVIRI, but this time the Heliosat-4 model (Qu et al, 2017) is applied. Heliosat-4 combines inputs from the McClear clear sky model and the McCloud cloud properties model (Schroedter-Homscheidt et al. 2017). The McClear model (also used by this research) takes as inputs the solar position, ground reflectance, ground elevation (Shuttle Radar Topography Mission (SRTM)), and atmospheric particulates (with 3- hour temporal resolution) zoned according to simplified Köppen Climate Classification (tropic, mid-latitude or sub-Arctic). The McCloud algorithm divides cloud into four types (low, medium, high, or thin ice) and treats these separately.

The last solar irradiance product studied in this paper is Solcast (https://solcast.com/). This is a paid-for service. (The author suggests the PVsyst version is chosen, regardless of intended software, for ease of analysis.) Satellite inputs include those from Meteosat. Like CAMS, a clear sky model (REST2v5, parameterised in the US) and a cloud model (proprietary in this case) are used (Bright, 2019). Atmospheric inputs are from MEERA-2 reanalysis (https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/). MEERA-2 temporal resolution is hourly, but the spatial resolution is 50 km. The ground altitude data incorporated is likewise of low spatial resolution.

All the satellite derived GHI databases reviewed here use input data from the same satellite sensors. All are semi-empirical (fitted to ground measurements somewhere to some extent). CMSAF might be expected to be more accurate than SARAH because atmospheric data is three-hourly rather than monthly. This also applies to CAMS. Differences may also arise from the different clear sky models

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2021 doi:10.20944/preprints202101.0545.v3

(SPECMAGIC versus McClear). Solcast has high temporal resolution of atmospheric variables but these have low spatial resolution, as do the ground elevation data inputs. The four satellite-derived databases analysed by this research are summarised in Table 1.

Table 1: Models and Data Inputs of satellite derived GHI datasets under review

Database	Satellite Model	Clear sky model	Temporal resolution of clear sky inputs	Spatial resolution of clear sky inputs
SARAH	Heoliosat-2	SPECMAGIC	Monthly	125 km
CMSAF	Heoliosat-2	SPECMAGIC	3-hourly	125 km
CAMS	Heoliosat-4	McClear	3-hourly	125 km
Solcast	Proprietary	REST2v5	Hourly	50 km

4. Selection of satellite database without ground validation

Initially an attempt to choose a suitable database to simulate PV output for a cooking system was made without the support of ground-based measurements, which is normally the situation throughout most of Africa. Two databases were selected for comparison: SARAH and CMSAF. Ten years of data (2007-2016) were analysed. The port of Dar es Salaam, on the Tanzanian coast, was taken as the example. Dar Es Salaam is Köppen Climate Classification subtype "Aw" (Tropical Savannah Climate). It is located on the coast of the Indian Ocean, at an elevation of 10 – 60 m, in the southern hemisphere.

Direct comparison methods were selected from the many statistical metrics available, as being well known and simple to apply. As can be seen in Table 2, overall summations and averages do little to distinguish between the two satellite databases. Yearly totals and hourly averages are almost the same. Mean standard deviation of each hourly global horizontal irradiation (GHI) value over 10 years is not particularly high, at around 10% of maximum hourly values, although SARAH does vary more than CMSAF.

Table 2: Overall Comparison of two satellite derived GHI datasets at Dar es Salaam

Statistical Measure	CMSAF	SARAH	% difference
Average annual in-plane irradiation kWh/m^2	1650.67	1664.12	-0.81
10-year average hourly GHI Wh/m^2	188.28	189.81	-0.81
Mean std dev of each hourly GHI value Wh/m^2	73.98	82.10	-9.90

More useful for distinguishing between the datasets are average hour of day values which show that SARAH is nearly always greater than CMSAF, except at the end of the day (Figure 1). So here, SARAH is overestimating (or CMSAF is underestimating), although the differences are not very big except for the last daylight hour.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2021 doi:10.20944/preprints202101.0545.v3

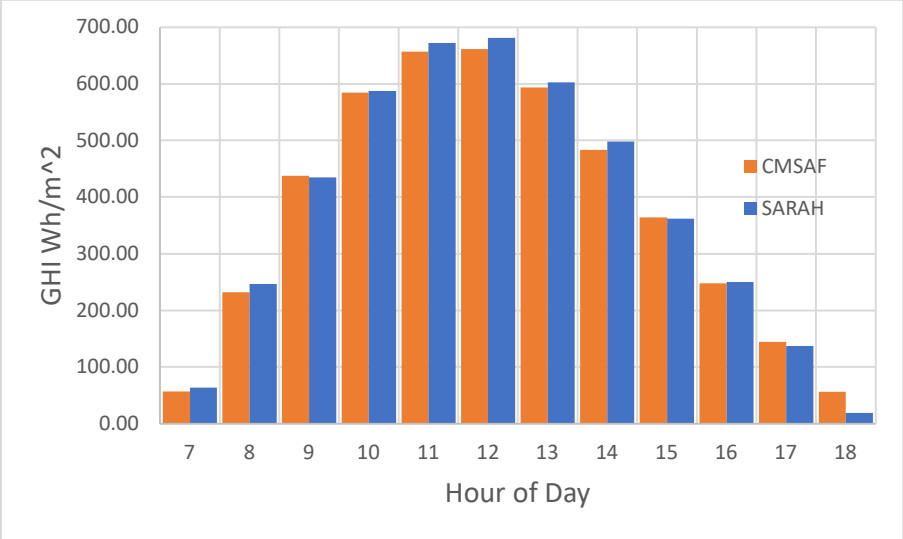

Figure 1: Average GHI per hour of day for Dar es Salaam

Monthly differences tell us that SARAH is less than CMSAF in May - November when precipitation is at its lowest. SARAH is more than CMSAF in the other months (Figure 2). Therefore, one of the models is not responding to cloud cover as well as the other. Again, the differences are not very big (average monthly difference 3.4%).

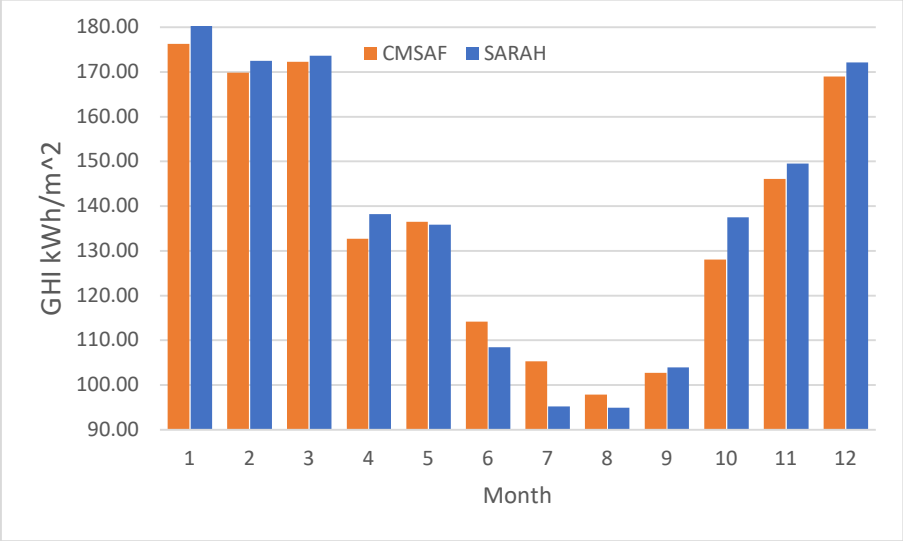

Figure 2: Average GHI kWh/m^2 for each month over ten years at Dar es Salaam

Daily averages also indicate that SARAH is less than CMSAF in May - October (cooler dry season) (Figure 3).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2021 doi:10.20944/preprints202101.0545.v3

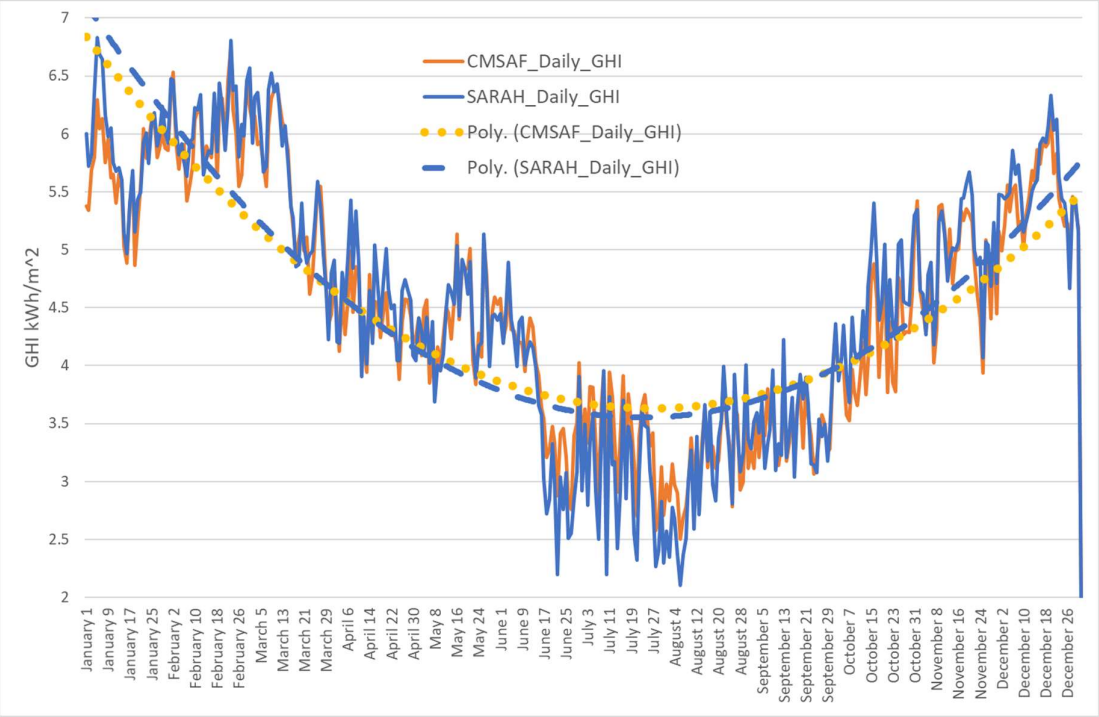

Figure 3: Daily Average GHI over ten years for Dar es Salaam

Looking at the frequency chart (Figure 4), CMSAF is greater than SARAH for GHI between 226 – 426 Wh/m^2, and 476 – 550 Wh/m^2. That is, SARAH is less than CMSAF at low-medium GHI values. Probably these are occurring in the dry season (from the daily and monthly graphs) and 5 pm – 6pm from the daily graph.

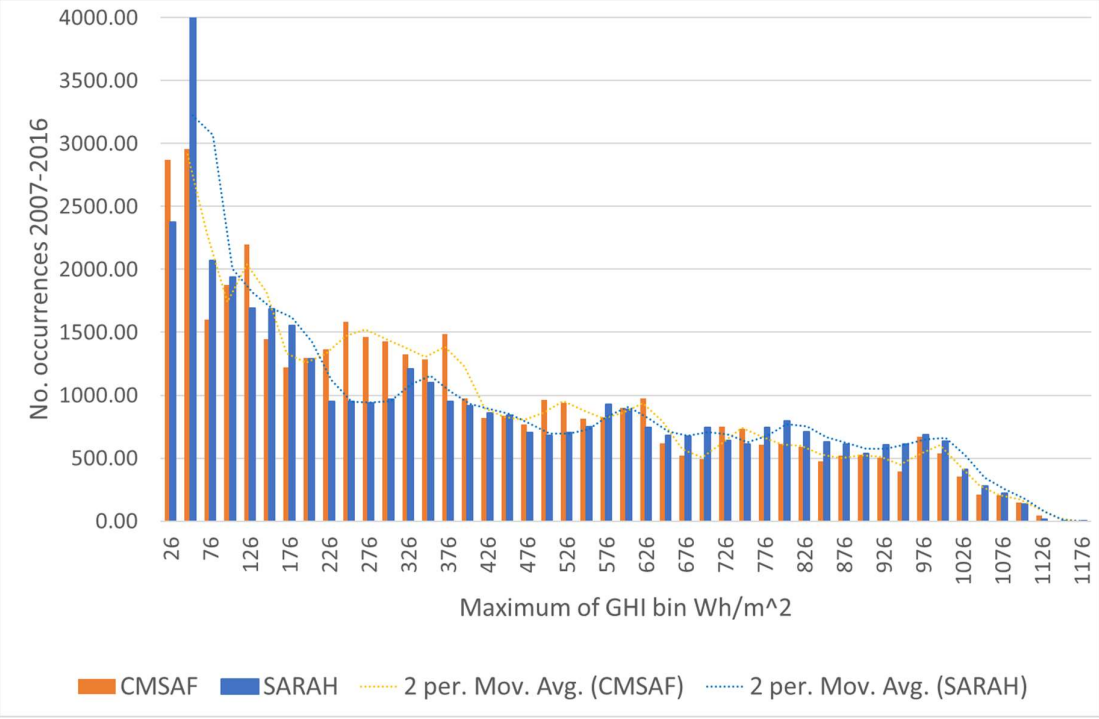

Figure 4: Frequency distribution of hourly GHI satellite values for Dar es Salaam

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2021 doi:10.20944/preprints202101.0545.v3

The statistical tests so far indicate that there is a difference between the two satellite datasets but do not give any guidance on which is preferable for the intended purpose. SARAH has lower irradiance values in the dry season, suggesting that possibly it is less representative of the true situation, but the evidence for this is weak.

The problem of model validation without measurement data was been discussed in the discipline of hydrology (https://www.researchgate.net/post/Is_there_a_way_of_calibrating_and_validating_sediment_yield_model_without_observed_sediment_data) but there are no references on this topic in the solar PV field, despite it being a very common problem. In hydrology, nearby data is used, but solar data changes rapidly over short distances (Palmer et al, 2018). Therefore, the second suggestion of using values from the literature is adopted in the following investigation. Theoretical clear sky values from Meteonorm are compared to the two satellite databases at Dar es Salaam (Figure 5).

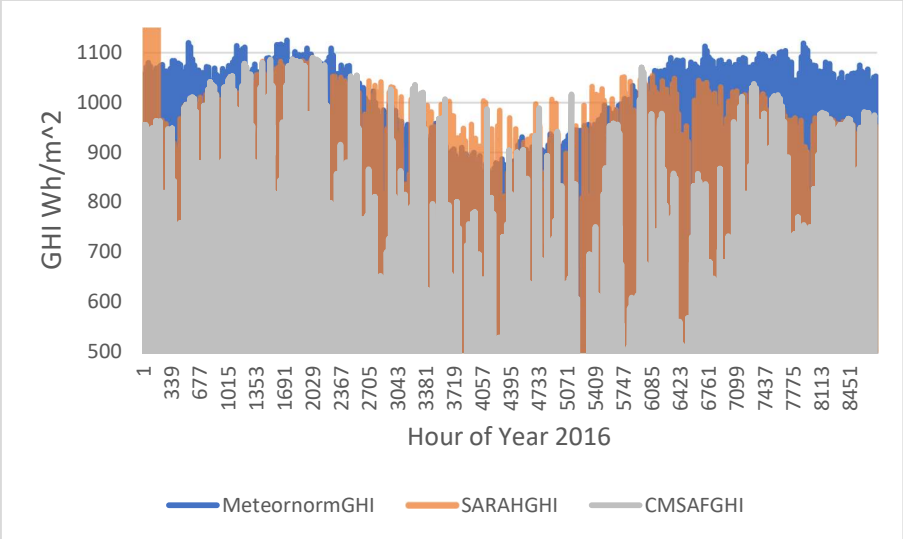

Figure 5: Comparison of Meteonorm (Clear sky), SARAH and CMSAF Hourly GHI for Dar es Salaam

It may be seen that SARAH tracks the clear sky values more closely. The general trend is for SARAH values to be higher than those of CMSAF. Whereas this would indicate accuracy in a desert, Dar es Salaam has a tropical savannah, also known as a tropical wet and dry, climate. Therefore, variance well below clear sky values in the wet season (November – May) is anticipated. Turning to the dry season (May – November), both satellite databases occasionally exceed the clear sky value, SARAH 16% of the year, and CMSAF 9% of the year.

The foregoing comparison with theoretical values again indicates that there is a difference between the two satellite datasets. There is weak evidence to suggest that SARAH is less accurate.

To conclude this paper section, it may be deduced that some idea of which database better relates to reality may be obtained by comparing them to local climate descriptions and seasonal behaviour. Comparison with clear sky values is another alternative. In both cases, any inference reached is somewhat arguable, and there appears to be a strong need for validation with ground measurements.

5. The Ground-Based Data used in this research.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2021 doi:10.20944/preprints202101.0545.v3

The ground-based data measurements used in this research are from two sources. The first is from two locations for a solar nano-grids project: UK Engineering and Physical Sciences (EP/L002612/1) Research Project: Solar Nano Grids (SoNG) http://songproject.co.uk/. The second author was involved in this project.

The two villages in Kenya where the measurement instruments are located are:

1. Lemolo B (latitude: -0.006861; longitude: 36.041456), in a semi-arid region of Kenya (Köppen-Geiger climate classification AW, tropical savannah).
2. Echareria (latitude: -0.348151; longitude: 36.2243068) with a Köppen-Geiger climate classification of Csb. That is, it enjoys a "Mediterranean" climate with a dry summer and mild wet winter.

Data is available from Lemolo B for July 2016 to December 2017, and from Echareria for September 2016 to October 2017. The data logging interval varies slightly but is generally 7 consecutive one-second values at the end of each minute (UTC). One-, five-, fifteen-, sixty-minute and daily averages were calculated for the purposes of this analysis. The measurement instrument was a CS300 (SP-110) APOGEE PYR-P silicon photovoltaic detector. The data was quality controlled as described in Appendix A.

The second source of ground data measurements is daily global horizontal irradiance data for two locations in a ground water management project: Gro for GooD: Groundwater Risk Management for Growth and Development (https://upgro.org/consortium/gro-for-good/). The data was downloaded from: https://metadata.bgs.ac.uk/geonetwork/srv/eng/catalog.search#/metadata/5cfd5112-e0c0-41cb-e054-002128a47908.

The details of the two villages in Kwale County, Kenya, where measurement took place are:

1. Galu: latitude -4.350, longitude 39.567
2. Munje: latitude -4.5101, longitude 39.4572

Both are Köppen-Geiger climate classification Af, tropical rainforest. They are coastal near Mombasa. The measurement equipment is Maplin Professional Solar Powered Wi-Fi Weather Station (Maplin N23DQ), which records solar radiation every 5 minutes. This is aggregated to daily totals before being made public. There is no way of obtaining any further information about this data.

During the COVID-19 pandemic when this article was written, these were the only ground GHI measurements it was possible to obtain. There is little solar data for Africa in any eventuality. Comparisons with all satellite databases will be affected, so none of them will be unfairly disadvantaged.

6. Comparison of satellite databases to ground-based measurements

6.1.1. Accuracy of the satellite databases

To commence, the measure of differences between the ground-based data measurements and the four satellite-derived databases under investigation was determined by calculating the normalised Root Mean Square Error (nRMSE), normalised by mean of inputs.

Looking at the highest temporal resolution data first, the only possible comparison was between CAMS and McClear (the clear-sky model) because these are the only databases (of those investigated) for which one-minute interval data is available. The average value per minute was

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2021 doi:10.20944/preprints202101.0545.v3

calculated for this purpose from the ground-based data. As would be expected, CAMS performs better than the clear-sky model at this level, because it accounts for cloud fields, although both deliver a very good nRMSE (Table 3), considering the time interval.

Table 3: nRMSE of Satellite model and clear-sky model at Lemolo, one-minute interval data

ONE-MINUTE	No. values Lemolo	nRMSE Lemolo %
CAMS	1143	76
McClear	1143	162

Moving up to five-minute interval data, CAMS and Solcast data were compared. The CAMS values and ground-based measurements were calculated as the average of the period. The Solcast data was downloaded directly at this temporal resolution. It may be seen from Table 4 that Solcast is more accurate than CAMS at this timing. The same may be observed for 15-minute interval data, which is directly available at this resolution from both CAMS and Solcast (Table 5). Solcast has almost the same value of nRMSE for 15-minute data as for 5-minute, whereas CAMS has a different value. This suggests that method of aggregation is having an impact.

Table 4: nRMSE of Satellite models at Lemolo, five-minute intervals

5-MINUTE	No. values Lemolo	nRMSE Lemolo
CAMS	72611	166
Solcast	72611	47

Table 5: nRMSE of Satellite models at Lemolo, 15-minute interval data

15-MINUTE	No. values Lemolo	nRMSE Lemolo
CAMS	45487	125
Solcast	45487	45

Juxtaposition of more satellite databases and another Song site (Echareria) was possible for hourly data, because of greater data availability at this resolution. The results are illustrated in Figure 6. The ground measurements were averaged to 60-minute intervals, but all the other datasets were available ready-prepared at this granularity. It is evident that the SARAH database performs poorly, being no better (Echareria) or worse (Lemolo) than the clear-sky model. At Lemolo, next best is Solcast, with CMSAF and CAMS being the most accurate, with little between them. At Echareria, CMSAF is third best, Solcast second and CAMS slightly outperforms Solcast, to give overall best accuracy. nRMSE values are lower for all databases at Lemolo due to its semi-arid climate. CAMS would be anticipated to deliver good results because it is a more modern model, but not to better Solcast.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2021 doi:10.20944/preprints202101.0545.v3

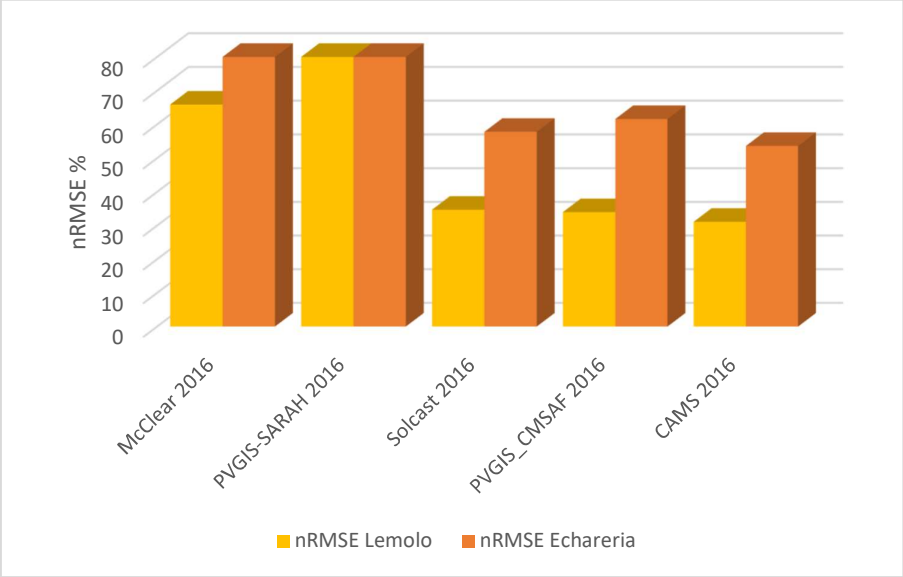

Figure 6: nRMSE of Satellite models and clear-sky model at Lemolo and Echareria, hourly data

The raw numbers upon which Figure 6 is based are given in Appendix B. Figure 6 is based on 2016 data only because this is the last year for which SARAH and CMSAF are currently available. However, the same pattern is observable between Solcast and CAMS if 2017 data is included to take advantage of the remaining ground measurements (Appendix B, Table B.1). Comparison of the same hours for both Song sites also gives the same order of performance (Appendix, Table B.1).

Normalised mean bias error (nMBE) values for the two sites additionally reveal virtually the same pattern of accuracy between databases (Table 6). Positive nMBE results demonstrate (on average) under-estimation in all cases. All nMBEs are low due to cancellation (mitigation of positive and negative values).

Table 6: nMBE of Satellite models at Lemolo and Echareria, 60-minute interval data

HOURLY	No. values Lemolo	nMBE Lemolo	No. values Echareria	nMBE Echareria
PVGIS-SARAH 2016	3489	0.35	715	0.47
Solcast 2016	3489	0.16	715	0.34
PVGIS_CMSAF 2016	3489	0.18	715	0.39
CAMS 2016	3489	0.16	715	0.31

Moving on to daily data granularity allowed the inclusion of two more Kenyan locations, Galu and Munje (Upgro project). Figure 7 shows that SARAH fares the worst at this interval at Lemolo and Echareria. There is less to choose between the other databases at Lemolo and Echareria. Having said that, Solcast performs well at this timescale, being best at three of four sites. Note: this graph was based on 235 days of 2016 data only, because this was all that matched in the satellite dataset and the logger records at Lemolo, Echareria and Galu. Thus, observations may be subject to anomaly caused by the low quantity of data. (2017 data only was available for Munje (182 days), obviating the use of SARAH and CAMS data). The raw numbers upon which Figure 7 is based are given in

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2021 doi:10.20944/preprints202101.0545.v3

Appendix B (Table B.2). Galu and Munje are at sea level, whereas Lemolo and Echareria are situated at 1961 m and 1594 m, respectively.

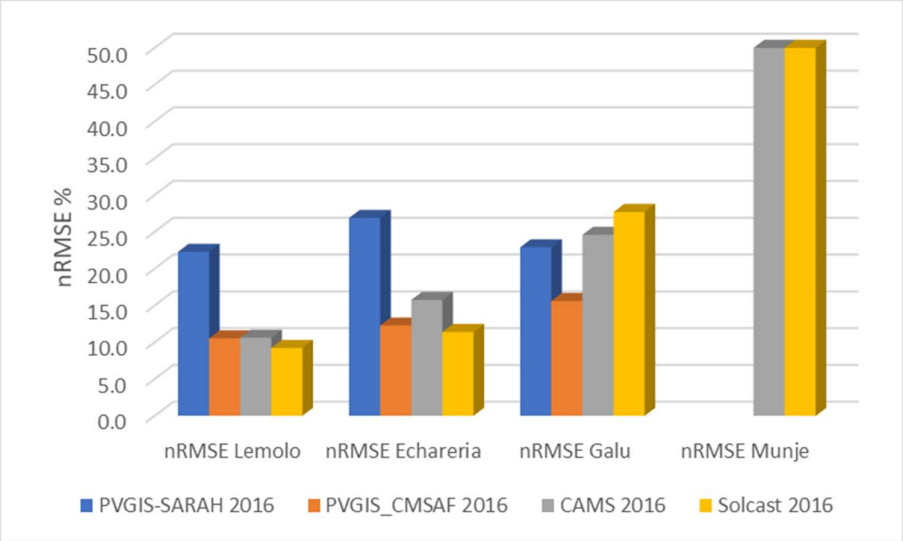

Figure 7: nRMSE of Satellite models four Kenyan sites, daily data

Thus, it appears that accuracy of satellite-derived databases is dependent on climate, temporal resolution, height above sea level and method of deriving one-minute, five-minute and hourly data from the original 15-minute satellite interval.

6.1.2. Instantaneous Accuracy of the satellite databases

The data in satellite databases is generally taken as representative of the whole time period of its resolution e.g. 15 minutes. However, satellite images are taken at an instant in time and in fact only reflect that instant. Therefore, a further comparison was made between the satellite values and the ground-based one-second value closest to the end time of those values, rather than with the average of ground-based readings for the whole period, as detailed above. The end time of the satellite 15-minute interval was used as the best compromise. In fact, the satellite image may be taken any time in the 15-minute interval, but for prepared global horizontal irradiation values, this time is not stored.

The results of this analysis are given in Tables 7 and 8 below. At the hourly resolution, there is little to choose between databases, except for SARAH. Solcast outperforms CAMS for 15-minute data.

Table 7: Instantaneous nRMSE of Satellite models and clear-sky model at Lemolo, 60-minute interval data

HOURLY	No. values Lemolo	nRMSE Lemolo
CAMS	1161	47
Solcast	1161	47
PVGIS_CMSAF	364	53
McClear	1161	60
PVGIS-SARAH	364	295

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2021 doi:10.20944/preprints202101.0545.v3

Table 8: Instantaneous nRMSE of Satellite models at Lemolo, 15-minute interval data

15-MINUTE	No. values Lemolo	nRMSE Lemolo
Solcast	4916	52
CAMS	4916	130

Although the foregoing discussion demonstrates that selection of the most accurate satellite model is not clear-cut, depending on location, resolution and method of ascertaining accuracy, the CAMS model would seem to be a good choice for most Kenyan sites. It is free to download and current.

6.1.3. Managing changing uncertainties and preserving the temporal pattern of the data

The nRMSE and nMBE measures employed above utilise the sum of squared residuals which assumes the size of the error term does not differ across values. This is does not hold true for the satellite databases under investigation, as is obvious from the frequency charts (see later). Additionally, these methods consider each data value at each time separately. They lose any pattern which may exist between previous and subsequent values. A performance metric capable of respecting the relationship between data points is the Pearson Product-Moment Correlation Coefficient (PMCC) (Bennett et al., 2013).

PMCC draws a trendline through a scatterplot of two data variables. Its value, *r*, is an indication of how well the data matches the line of best fit. *r* ranges between 0 (no relationship between the two data sets) and 1 (a perfect relationship) (Laerd Statistics).

The PMCC values for hourly GHI data for Lemolo and Echareria are shown in Table 9.

Table 9: PMCC values for hourly GHI satellite data (2016) for Lemolo and Echareria

Database	Lemolo PMCC	Echareria PMCC
SARAH	0.875	0.877
CMSAF	0.978	0.905
CAMS	0.978	0.946
Solcast	0.971	0.907

According to this metric, CAMS and CMSAF jointly have the best accuracy at Lemolo, followed by Solcast, with SARAH last. (That is, the same as the nRMSE ranking). At Echareria, CAMS is best, followed by Solcast, then CMSAF, with SARAH coming last. Again, the same as nRMSE comparison.

6.1.4. Statistics for the ground-based measurements and satellite databases

Having determined the accuracy of the four satellite databases under investigation in this research, the effect on solar irradiation values is explored. Table 10 details findings for Lemolo. SARAH is closest in terms of overall solar radiation sum and hourly average to the logger measurements, due to its smaller standard deviation. Compensating errors are occurring more frequently than for the other satellite databases. CMSAF and CAMS are remarkably different, considering their similar nRMSE values. All the databases tend to over-estimate, using this measure, with CMSAF being the worst at this. However, they perform this over-estimation in just one-third of daylight hours, under-estimating for most of the time.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2021 doi:10.20944/preprints202101.0545.v3

Table 10: General Statistics for Lemolo

GHI for 3489 hours Wh/m^2	Logger	SARAH	CMSAF	CAMS	Solcast
Sum	842684	835475	910815	888160	881534
Hourly average	242	240	261	255	253
% difference sum/avg to Logger		- 1	8	5	5
Hourly Std dev	322	337	358	345	347
% difference std dev to Logger		5	11	7	8
% of hours under-estimating		70	63	69	68

Analysis of values for Echareria generates a contrasting set of observations (Table 11). Solcast has the most accurate overall solar radiation sum and hourly average with the smallest standard deviation, and SARAH has the greatest mismatch at this site. Again, all the databases tend to under-estimate in most hours.

Table 11: General Statistics for Echareria

GHI for 715 hours Wh/m^2	Logger	SARAH	CMSAF	CAMS	Solcast
Sum	157929	189617	172963	179699	167135
Hourly avg	221	266	242	252	234
% difference sum/avg to Logger		20	10	14	6
Hourly Std dev	314	360	354	350	335
% difference std dev to Logger		15	13	11	7
% of hours under-estimating		65	67	67	64

Thus, accuracy and usefulness of satellite database appears to vary from location to location in the same African country. Good performance at one site cannot be taken as a guide for the country as a whole.

The data is now decomposed for closer examination. In the case of percentage difference of satellite value to logger per hour, Solcast shows the greatest similarity at Lemolo. The other databases cluster closely together, further away from the logger and Solcast. Figure 8 demonstrates this observation in the form of trendlines. The busy data series plots are hidden for clarity. At Echareria, CMSAF and CAMS are jointly closest to the logger, with the trendlines of Solcast and SARAH being at greater distances ((Appendix B, Figure B.1).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2021 doi:10.20944/preprints202101.0545.v3

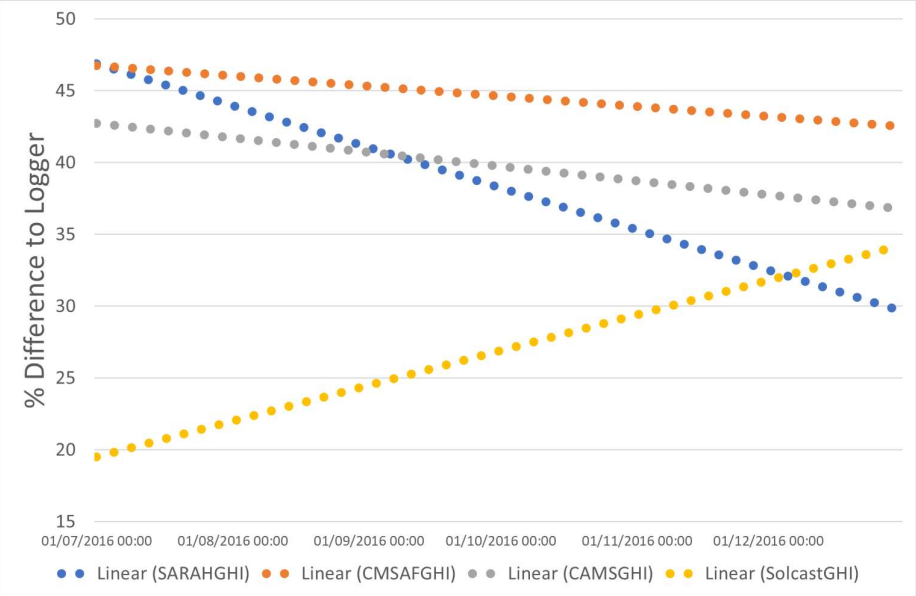

Figure 8: Trendlines of Percentage Difference to Logger for four satellite derived databases of Average Hourly GHI at Lemolo

Looking at the average hourly GHI value for each discrete daylight hour from each data source for Lemolo (Figure 9), compared to the logger, all datasets over-estimate, except that SARAH tracks the logger closely at midday. (Note: this observation is not the consequence of incorrect time stamps. This has been tested and all databases aligned to the nearest hour (nn:00). 60-minute CAMS reports at nn:00, as does Solcast (PVSyst version), SARAH at nn:06 and CMSAF at nn:51.)

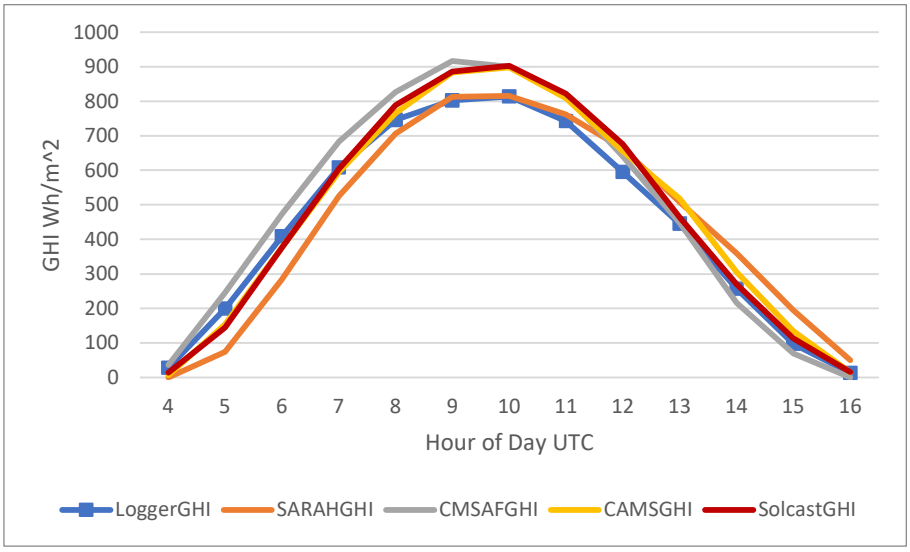

Figure 9: Average GHI per hour of day (satellite values and ground measurements) for Lemolo

In the case of Echareria, all databases over-estimate, noticeably at noon. Solcast tracks the logger the closest (Appendix B, Figure B.2).

However, looked at in terms of percentage differences, all the satellite databases are only around 10% different from the ground data value in the early afternoon hours at Lemolo i.e. the most productive hours for PV (Figure 10), although using this measure, CAMS is frequently most accurate.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2021 doi:10.20944/preprints202101.0545.v3

This suggests that any of them may function well for the purpose of PV performance simulation. All databases also have similar comparative hourly differences to logger readings at Echareria ((Appendix B, Figure B.3).

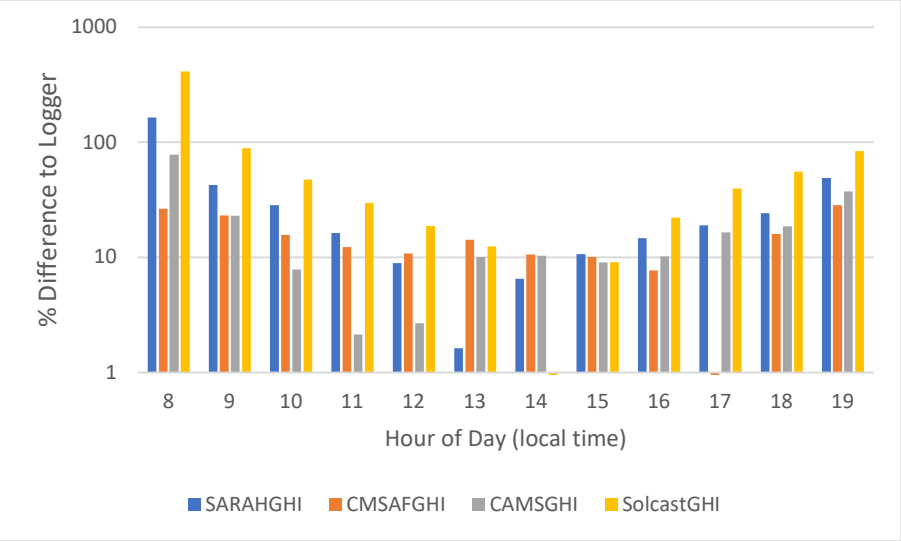

Figure 10: Absolute Percentage Differences in average GHI per discrete hour between four satellite derived GHI databases and ground measurements for Lemolo

On a daily basis, SARAH has most similar values to the ground measurement at Lemolo on average (Table 12). (Solcast has the nearest value at Echareria.) At Lemolo, SARAH under-estimates in summer and over-estimates in winter (Appendix B, Figure B.4). The other databases are only inclined to this trend to a minimal degree (Figure 11). The average values in Table 12 hide the observation that Solcast really has the closest daily values to the ground measurements, with CAMS and CMSAF also performing well, and SARAH less so (Figure 11).

Table 12: Average Daily GHI difference for both Song sites

% Avg daily difference to logger	No. Days	SARAHGHI	CMSAFGHI	CAMSGHI	SolcastGHI
Lemolo	152	0	8	6	4
Echareria	22	22	12	16	4

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2021 doi:10.20944/preprints202101.0545.v3

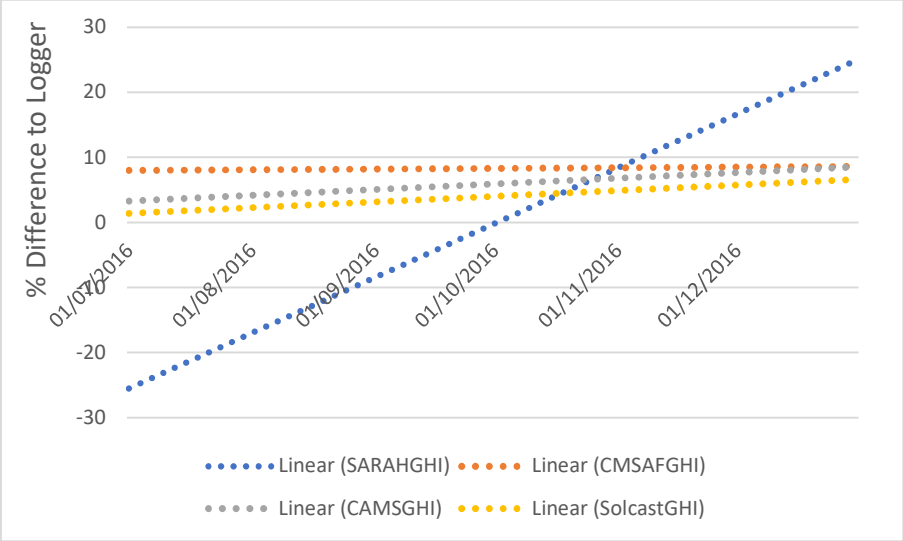

Figure 11: Trendlines of Percentage Difference to Logger for four satellite derived databases of Average GHI per Day of Year for Lemolo

A study of the frequency distribution of GHI at Lemolo shows that SARAH has too many low values as compared to the logger, and CAMS has too few. All the databases mirror the logger reasonably well between 100 Wh/m^2 and 1000 Wh/m^2. SARAH tracks it the best, then Solcast and CMSAF, with CAMS coming last. All have too many very high values (Figure 12). There are also too many high values in all satellite databases at Echareria (Appendix B, Figure B.5).

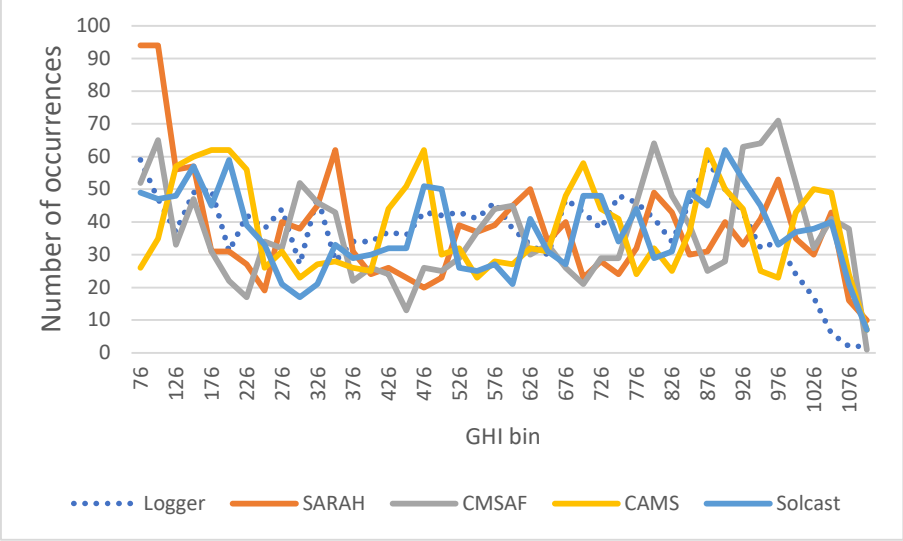

Figure 12: Frequency distribution of hourly GHI (satellite values and ground measurements) for Lemolo

Study of further statistics does not clarify the issue of choice of satellite GHI database to any extent. CAMS has some of the best nRMSE and nMBE values but SARAH has the most realistic frequency distribution. SARAH can have either the best or worst accuracy in respect of sums and averages depending on location. Solcast has trendlines of percentage difference to logger closest to zero (i.e. best match) for hourly and daily values at both Lemolo and Echareria. Solcast also does well on a

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2021 doi:10.20944/preprints202101.0545.v3

daily basis. All databases are likely to overestimate in the middle of the day by around 10% when most PV production occurs, so in this regard they are evenly matched.

Photovoltaic system designers, including those with electric cooking loads, need to consider the accuracy of solar databases and not take them verbatim. For example, it might be prudent to consider a 10% buffer in the design of a photovoltaic domestic cooking system to account for the variation and uncertainty in the results. Whilst in the past this might have added to system costs, the rapid fall in the price of photovoltaic modules mean that this should not be an additional economic burden and has the potential to ensure greater end-user service satisfaction. But, further investigation is needed with more ground stations in other parts of Africa.

Which database performs best is site dependent. In the case of all the databases, it is a matter of how well equal and opposite errors balance, rather than few errors. Some have many small errors, others have fewer large ones, and this varies from location to location, season to season and time of day.

Looking at the resolution of the atmospheric inputs (Table 1), it would be expected that SARAH would have the worst performance. A review of the different analytical tests and four Kenyan sites investigated here shows that it sometimes does but not always. Occasionally it can exhibit the best accuracy of the satellite databases examined. From the inputs table (Table 1), Solcast should be best. It is at Echareria but not at Lemolo. The accuracy of the satellite databases is thus influenced rather complexly by both resolution of atmospheric inputs and performance of the clear sky model utilised. It depends how well the clear sky model performs in the particular climate zone (or percentage of clear, almost clear, partly clear, totally cloudy days) at the site of interest. That is, it is site dependent. In addition to this, it has been found that some clear sky models are more sensitive to uncertainties in inputs than others (Polo et al, 2014).

To clarify the findings of this paper, Appendix B, Table B.3 summarises the accuracy ranking of each database in relation to the other three for each performance metric for each temporal resolution for each site. Looking at the hourly and daily data at Lemolo, SARAH achieves the best accuracy most times, but it also achieves the worst accuracy most times. At Echareria, Solcast has the highest number of best scores, with SARAH having the highest number of worst scores. Taking both sites together, CAMS and Solcast jointly perform the best and SARAH the worst. An alternative is to calculate the average rank (Appendix B, Table B.4). (Lowest score is best). Looking at hourly data only, CAMS is best at both Lemolo and Echareria. Taking both sites together, CAMS scores the highest accuracy followed by Solcast in second place, then CMSAF, and lastly SARAH.

7. Conclusion

It is not feasible to verify a satellite derived GHI model without ground measurements although an informed guess as to which model is likely to perform satisfactorily may be made via comparison with knowledge of local seasons and climate and/or clear sky data.

The comparative accuracy, and therefore the selection of, satellite derived GHI databases has been shown to be site dependent. Therefore, those datasets which by chance have been fitted to ground measurements close to the site of interest, or datasets which employ greater numbers of ground stations data in their construction, are likely to display superior performance. Performance is evidently influenced by climate and height above sea level, although the role these factors play is not clear from the analysis carried out here. Temporal resolution and method of deriving one-minute, five-minute and hourly data from the original 15-minute satellite interval are also playing a part. Further factors are the compatibility of the clear sky model to the climate zone of the site of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2021 doi:10.20944/preprints202101.0545.v3

interest, spatial and temporal resolution of clear sky model input parameters, and susceptibility of the clear sky model to imperfections in input.

Based on the preceding analysis, the CAMS model, as a publicly available, up-to-date, and fairly accurate resource, appears to be a good option for PV simulation in East Africa. However, in general, all the databases deliver figures around 10% of the ground measurement values in the middle of the day. This contrasts with findings in the UK where one database clearly outranked the others (Palmer et al, 2018). Whether this level of accuracy is sufficient to model provision of energy for pressure cookers/hotplates etc has yet to be determined in a subsequent publication. A simple investigation, using the PVGIS Performance of Off-Grid PV Systems tool to model the type of solar cooking system usage envisaged by the MECS project (300 W solar panel, battery 24 V / 75 Ah, 1.0kWh daily consumption for two meals) at Lemolo, revealed some difference between databases. SARAH models 25 days per year with an empty battery, and 283 days with a full one. CMSAF predicts 11 days per year with an empty battery, and 304 days with a full one. SARAH generally anticipated a lower state of battery charge, suggesting longer cooking times.

Impression of accuracy is determined by which mathematical measure is employed. Solar radiation publications commonly use average hourly/daily nRMSE or nMBE but these aggregate values can cover trends. Equal and opposite errors may counterbalance. Hourly trendlines of percentage differences and nRMSE values rank the satellite databases in the same order of accuracy at Echareria (Appendix B, Figure B.1 and Figure 6) but not at Lemolo (Figure 6 and Figure 8).

Whereas making up the difference over the long-term is acceptable for calculating the profitability of a solar farm, it is of little use when investigating if an individual solar panel can power a cooking device at a specific time. If a battery is used in the cooking system, daily data becomes applicable. Viewed on an hour of day (Figure 10), or daily basis (Figure 7), none of the satellite derived GHI databases largely outperform the others. In any one hour, one will be more accurate than the others, but there is no consistency as to which one this is.

The location dependence of GHI databases accuracy means that superior precision at one site (or a small number of sites) cannot be taken as a guide for East Africa or one country there as a whole. Ground-based measurements (e.g. for one year) are necessary to select the more accurate satellite database at each specific location. It is hoped the next steps in the MECS project will include setting up a ground station in the region.

If it was known which clear sky model performs best in each climate, it would be possible to select a satellite GHI database appropriately. This would overcome the problem that satellite databases are used where there are no ground measurements, but ground measurements are needed to choose the most accurate satellite database. Only one study has investigated this (Sun et al, 2019). But here the 29 Koppen zones are simplified into 5. This is not enough because Kenya is arid and REST2 is reported as best, whereas it has not been found to be so at all sites in this study.

Finally, the goal of this research was to establish which satellite-derived solar irradiance dataset is most suitable for simulating PV yield in East Africa. The initial findings presented here suggest that all the four databases reviewed are suitable for this task. Future work will include comparing modelled PV output based on satellite datasets to actual output. This may enhance understanding of suitability further.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2021 doi:10.20944/preprints202101.0545.v3

Funding

The project was funded by Modern Energy Cooking Solutions (MECS), UK, P.O. number 500230223.

Conflicts of Interest

The authors declare no conflict of interest.

Data Availability

The datasets related to this article can be found as follows:

1. PVGIS-SARAH at https://re.jrc.ec.europa.eu/pvg_tools/en/tools.html, an open-source online data repository hosted at Joint Research Centre EU Science Hub (Urraca et al. 2017)).
2. PVGIS-CMSAF at https://re.jrc.ec.europa.eu/pvg_tools/en/tools.html, an open-source online data repository hosted at Joint Research Centre EU Science Hub (Amillo et al. 2018).
3. CAMS at http://www.soda-pro.com/web-services/radiation/cams-radiation-service, an open-source online data repository hosted at the research center O.I.E. of Mines ParisTech (Center Observation, Impact, Energy) (Qu et al., 2017).
4. Solcast at https://solcast.com/, a commercial data supplier with support for researchers (Bright, 2019)
5. McClear clearsky model at http://www.soda-pro.com/web-services/radiation/cams-mcclear, an open-source online data repository hosted at the research center O.I.E. of Mines ParisTech (Center Observation, Impact, Energy) (Schroedter-Homscheidt et al. 2017).
6. Ground-based Kenyan data at Solar Nano Grids (SoNG) http://songproject.co.uk/, a research partnership. Data available on contact.
7. Ground-based Kenyan data at Gro for GooD: Groundwater Risk Management for Growth and Development (https://upgro.org/consortium/gro-for-good/). The data was downloaded from: https://metadata.bgs.ac.uk/geonetwork/srv/eng/catalog.search#/metadata/5cfd5112-e0c0-41cb-e054-002128a47908.

APPENDIX A: Quality Control of Lemolo and Echareria Data

Data quality control checks summarised in Table 1 were applied. The chosen tests were selected from Journée and Bertrand (2011) and Laitia et al (2014). These are based on guidance from Baseline Surface Radiation Network (BSRN) from the World Radiation Monitoring Centre (WRMC). The procedures were chosen with regard to availability of data (i.e. no beam or diffuse irradiation data was available).

Table 1: Quality Criteria of global horizontal irradiation data used in temporal drift tests (TD), physical threshold tests (PT), step tests (S), persistence tests (P) and spatial consistency tests (SC).

Type of Test	Test Name	Test Description	Quality Criteria
TD	Temporal Drift	Clock drift detection	i. Comparative hourly plots between datasets ii. Comparative hourly plots between datasets and clear sky values[1].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2021 doi:10.20944/preprints202101.0545.v3

PT	Upper Limit	Upper bound when comparing surface solar radiation data against the extraterrestrial solar radiation[2].	GHI/ETR < 1 if h > 2°
PT	Upper Clear-sky Limit	Upper bound when comparing surface solar radiation data against the clear-sky solar radiation[1].	GHI/Clear-sky irradiance <= 1.1 if h > 2°
PT	Lower Limit	Lower bound for heavily overcast conditions with low atmospheric transparency.	GHI ≥ 0.03 * ETR
PT	Clear-sky hours	Number of clear-sky hours[3].	
PT	Daily Lower Limit	Lower bounds for GHI in heavily overcast conditions with low atmospheric transparency. The daily mean μ is calculated from data when the sun is above the horizon (daylight hours).	μ (GHI/ETR) ≥ 0.03
S	Step	Plausible rate of change between two successive timestamps.	$\left(\frac{GHI(t)}{ETR\ (t)} - \frac{GHI(t-1)}{ETR\ (t-1)}\right) < 0.75$ If h > 2°
S	Shadow	• Shadow contamination: rapid drop of values followed by sudden increase.	$\left(\frac{GHI(t)}{ETR\ (t)} - \frac{GHI(t-1)}{ETR\ (t-1)}\right) > 0.1$ If h > 2°
P	Persistence	Check for variability of measurements/sensor failure. The daily mean μ and standard deviation σ are calculated from data when the sun is above the horizon (daylight hours).	$\left(\frac{1}{8}\right).\mu\left(\frac{GHI}{ETR}\right) \leq \sigma\left(\frac{GHI}{ETR}\right) \leq 0.35$
SC	Spatial Consistency / Sum	Comparison of the sum of GHI for 990 hours in the period under review when both weather instruments report data.	
SC	Completeness of data	Percentage of hours in the measurement period for which data exists.	

[1] The McClear clear-sky model was used because of its easy accessibility (download from http://www.soda-pro.com/web-services/radiation/cams-mcclear). It is a physical model and employs a look-up table on satellite-derived aerosols, water vapour and ozone data.

[2] Extraterrestrial irradiation and solar elevation angle were obtained from the solaR package in R software (Perpiñán 2012).

[3] Clearsky periods were identified from the simple model of Collares-Pereira and Rabi (1979) (GHI/ETR > 0.6) due to lack of measured diffuse irradiance.

For the temporal drift test, there is no evidence of incorrect timestamp values. The results of most of the other tests are presented in Table 2. Mostly, these are very good, with only a few errors around sunrise and sunset.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2021 doi:10.20944/preprints202101.0545.v3

Table 2: Percentage of hours containing data which failed QC Limit, Step, Shadow and Persistence tests for the Lemolo and Echareria.

Test Name	Lemolo B	Echareria
Upper Limit	22	15
Upper Clear-sky Limit	0.9	7
Lower Limit	9	9
Daily Lower Limit	0	4
Step	0.1	0
Shadow	2.7	7
Persistence	0	0

Proceeding to the Spatial Consistency test, both villages achieve similar values, allowing for the difference in climate. There is a high level of completeness of data for Lemolo on an hourly basis. Echareria data is fragmentary but available.

Table 3: Results of Spatial Consistency tests

Test Name	Lemolo B	Echareria
Percentage of Clearsky hours in test period	54	50
Average GHI of clear-sky hours Wh/m^2	700	500
Sum of GHI for 990 hours when both data loggers report data kWh/m^2	237	219
Percentage completeness of data 2016 – 2017 Percentage completeness of data 2017 only	93 100	11 6

In general, the results of the tests indicate that Lemolo B and Echareria data loggers have produced data of good quality. There are few outliers, little shading and nothing to suggest instrument failure.

On the other hand, the loggers at Lemolo and Echareria may or may not be absolutely vertical. However, there is no way of obtaining any further information. The measurement is in millivolts so a tiny difference will have a large impact at low values i.e. morning and evening hours.

APPENDIX B

Table B.1: nRMSE of Satellite models and clear-sky model at Lemolo and Echareria, 60-minute interval data

HOURLY	No. values Lemolo	nRMSE Lemolo	No. values Echareria	nRMSE Echareria	nRMSE Lemolo 715 values
PVGIS-SARAH 2016	3489	82.3	715	80.8	65.7
Solcast 2016	3489	37.5	715	57.8	50.2
PVGIS_CMSAF 2016	3489	34	715	68.6	35.4
CAMS 2016	3489	31.1	715	53.6	33.3
Solcast 2017	8102	38.6	386	68.5	
Solcast 2016 & 17	11590	37.5	1100	61.6	
CAMS 2017	8102	38.6	386	65.7	

CAMS 2016 & 17	11590	36.5	1100	57.9	
McClear 2016	3489	65.9	715	99.8	77.5
McClear 2017	8102	73.7	386	100.6	
McClear 2016 & 17	11590	71.3	1100	100.1	

Table B.2: nRMSE of Satellite models four Kenyan sites, daily data

DAILY	nRMSE Lemolo	nRMSE Echareria	nRMSE Galu	nRMSE Munje
PVGIS-SARAH 2016	22.3	26.9	22.9	
PVGIS_CMSAF 2016	10.5	12.3	15.6	
CAMS	10.6	15.7	24.6	52.2
Solcast	9.2	11.4	27.5	53.2

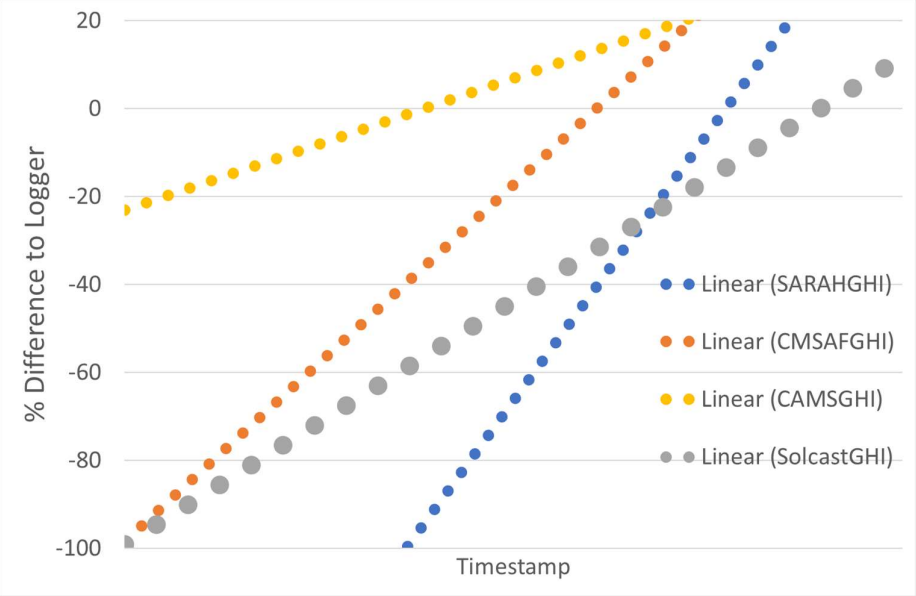

Figure B.1: Trendlines of Percentage Difference to Logger for four satellite derived databases of Average Hourly GHI at Echareria

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2021 doi:10.20944/preprints202101.0545.v3

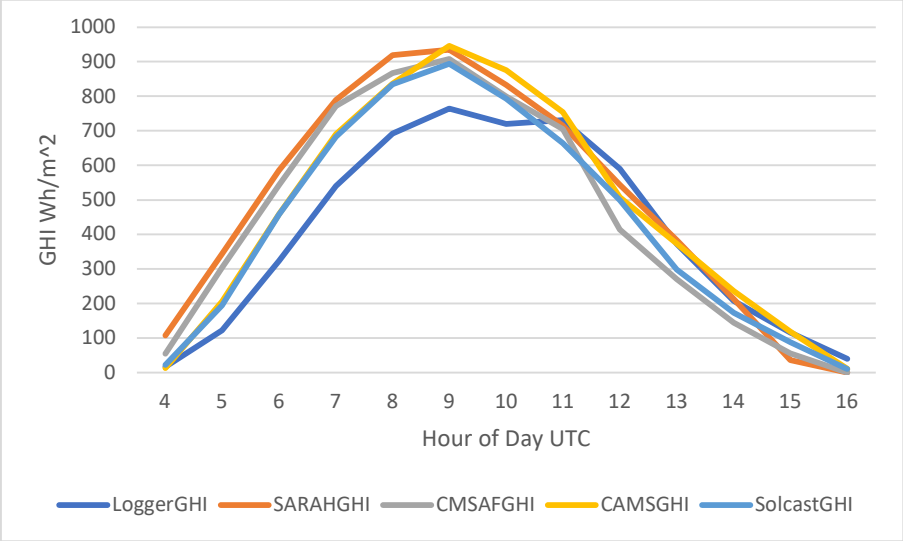

Figure B.2: Average GHI per hour of day (satellite values and ground measurements) for Echareria

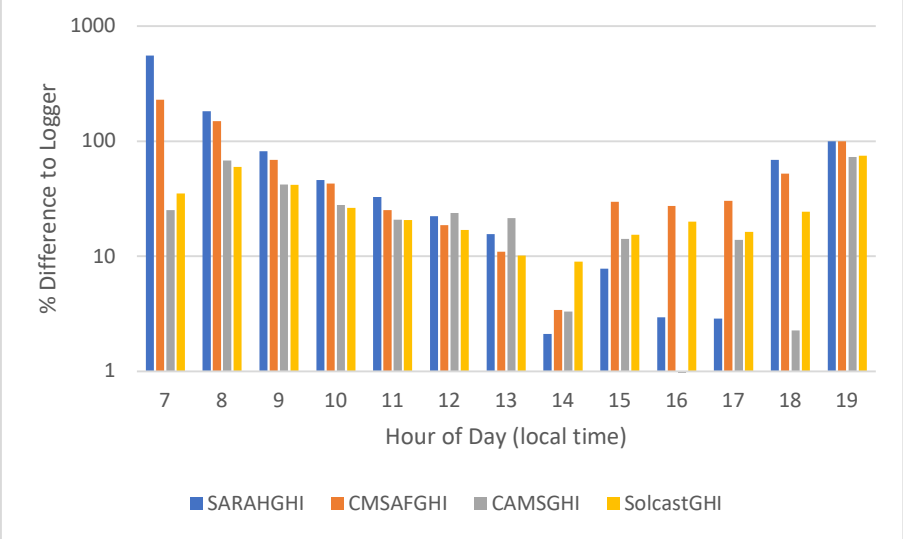

Figure B.3: Absolute Percentage Differences in average GHI per discrete hour between four satellite derived GHI databases and ground measurements for Echareria

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2021 doi:10.20944/preprints202101.0545.v3

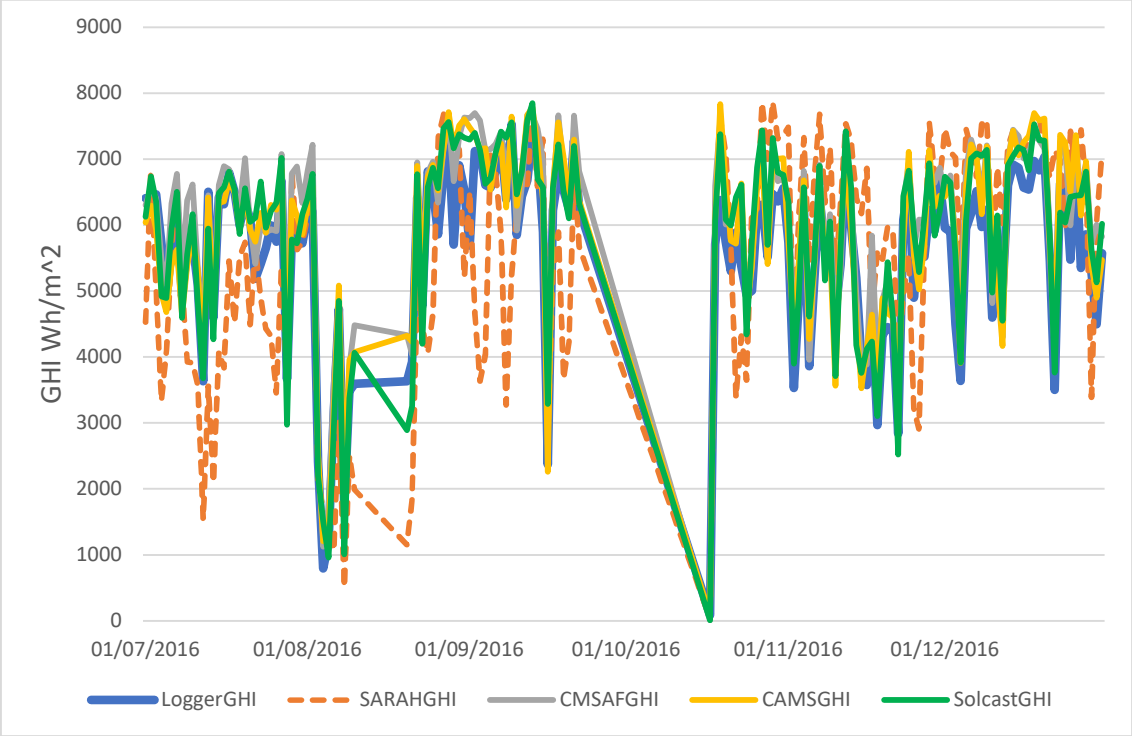

Figure B.4: Average GHI per Day of Year (satellite values and ground measurements) for Lemolo

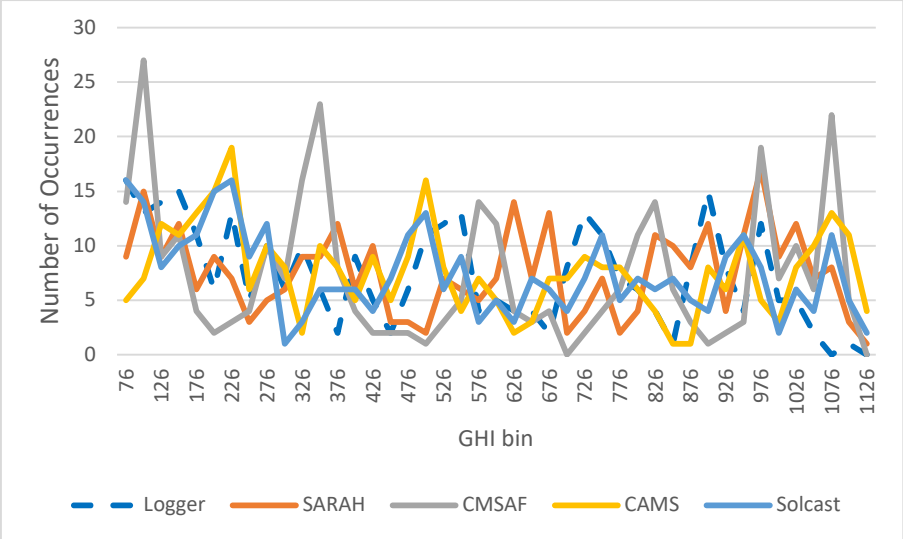

Figure B.5: Frequency distribution of hourly GHI (satellite values and ground measurements) for Echareria

Table B.3: Accuracy ranking of each database for each performance metric for each temporal resolution for each site.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2021 doi:10.20944/preprints202101.0545.v3

		Lemolo					Echareria					Galu				Munje	
Time Interval	Test	Best	2nd Best	3rd Best	Worst of 4	Worst of 5	Best	2nd Best	3rd Best	Worst of 4	Worst of 5	Best	2nd Best	3rd Best	Worst of 4	Best	2nd Best
One minute	nRMSE	CAMS	McClear														
5 minute	nRMSE	Solcast	CAMS														
15 minute	nRMSE	Solcast	CAMS														
15 minute	Instant nRMSE	Solcast	CAMS														
60 minute	nRMSE	CAMS	CMSAF	Solcast	McClear	SARAH	CAMS	Solcast	CMSAF	SARAH	McClear						
60 minute	Instant nRMSE	CAMS	Solcast	CMSAF	McClear	SARAH											
60 minute	nMBE	CAMS	CMSAF	Solcast	SARAH		CAMS	Solcast	CMSAF	SARAH							
60 minute	Hourly Average	SARAH	Solcast	CAMS	CMSAF		Solcast	CMSAF	CAMS	SARAH							
60 minute	Hourly Std Dev	SARAH	CAMS	Solcast	CMSAF		Solcast	CAMS	CMSAF	SARAH							
60 minute	Trend closest to logger	Solcast	CAMS	CMSAF	SARAH		CAMS	CMSAF	Solcast	SARAH							
60 minute	Pearson	CAMS	CSMSAF	Solcast	SARAH		CAMS	Solcast	CMSAF	SARAH							
60 minute	Average GHI per hour of day	SARAH	CAMS	Solcast	CMSAF		Solcast	CAMS	CMSAF	SARAH							
Daily	nRMSE	Solcast	CAMS	CMSAF	SARAH		Solcast	CMSAF	CAMS	SARAH		CMSAF	SARAH	CAMS	Solcast	CAMS	Solcast
Daily	Daily Average	SARAH	Solcast	CAMS	CMSAF		Solcast	CMSAF	CAMS	SARAH							
Daily	Average GHI per Day of Year	Solcast	CAMS	CMSAF	SARAH												
	Frequency Distribution	SARAH	Solcast	CMSAF	CAMS												

Table B.4: Calculation of average rank of each database across all performance metrics for hourly data

	Lemolo					Echareria					Both
Rank	1	2	3	4	Mean	1	2	3	4	Mean	Mean
CAMS	4 x 1	3 x 2	1 x 3	0	3.25	4 x 1	2 x 2	1 x 3	0	2.75	3
CMSAF	1 x 1	3 x 2	2 x 3	3 x 4	6.25	0	2 x 2	5 x 3	0	4.75	5.5
Solcast	1 x 1	2 x 2	5 x 3	0	5	3 x 1	3 x 2	1 x 3	0	3	4
SARAH	2 x 1	0	0	5 x 4	5.5	0	0	0	7 x 4	7	6.25

References

Amillo, A.M.G.; Ntsangwane, L.; Huld, T.; Trentmann, J. Comparison of satellite-retrieved high-resolution solar radiation datasets for South Africa. 2018, Journal of Energy in Southern Africa, 29(2):63-76, DOI: 10.17159/2413-3051/2017/v29i2a3376

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsilli-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D. and Andreassian, V. (2013) Characterising performance of environmental models. Environmental Modelling and Software 40, 1-20. https://doi.org/10.1016/j.envsoft.2012.09.011

Available from: https://www.researchgate.net/publication/285693132_Characterising_performance_of_environmental_models [accessed Jan 12 2021].

Bright, J. Solcast: Validation of a satellite-derived solar irradiance dataset. Solar Energy. 189. 435-449. 2019. 10.1016/j.solener.2019.07.086

Collares-Pereira M, Rabi A, The average distribution of solar radiation – correlations between diffuse and hemispherical and between daily and hourly insolation values, Solar Energy 1979, 22:155

Gracia-amillo, A.M.; Ntsangwane, L.; Huld, T.; Trentmann, J. Comparison of satellite-retrieved high-resolution solar radiation datasets for South Africa. Journal of Energy in Southern Africa. 2018, 29(2), 63-76

Journée M, Bertrand C. Quality control of solar radiation data within the RMIB solar measurements network, Solar Energy 2011; 85:72-86.

Laitia L, Andreis D, Zottele F, Giovannini L, Panziera L, Toller G, Zardi D. A solar atlas for the Trentino region in the Alps: quality control of surface radiation data, Energy Procedia 2014; 59:336-343

Laerd Statistics. Available from: https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php [accessed Jan 12 2021].

Meyer,R. Industry insight: On-site measurements for PV projects – is it really necessary? ESI Africa – Africa's Power Journal, 2015, Available from: https://www.esi-africa.com/industry-sectors/renewable-energy/industry-insight-on-site-solar-measurements-for-solar-pv-projects-is-it-really-necessary/ [accessed Jan 12 2021].

Palmer, D.; Koubli, E.; Cole, I.; Betts,T.; Gottschalg, R. Satellite or ground-based measurements for production of site specific hourly irradiance data: Which is most accurate and where?, 2018, Solar Energy, 165, 240-255

Perpiñán O, Solar Radiation and Photovoltaic Systems with R, Journal of Statistical Software 2012;50(9): 1-32. URL http://www.jstatsoft.org/v50/i09/.

Polo, J. & Antonanzas-Torres, F. & Vindel, J.M. & Ramirez, L., 2014. "Sensitivity of satellite-based methods for deriving solar radiation to different choice of aerosol input and models," Renewable Energy, Elsevier, vol. 68(C), pages 785-792. Available from: https://ideas.repec.org/a/eee/renene/v68y2014icp785-792.html [accessed Jan 12 2021].

Qu, Z.; Oumbe, A.; Blanc, P.; Espinar, B.; Gesell, G.; Gschwind, B.; Klüser, L.; Lefèvre, M.; Saboret, L; Schroedter-Homscheidt, M.; and Wald L. Fast radiative transfer parameterisation for assessing the surface solar irradiance: The Heliosat-4 method, Meteorol. Z., 26, 33-57, doi: 10.1127/metz/2016/0781, 2017

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Schroedter-Homscheidt, M.; Hoyer-Klick, C.; Killius, N.; Lefevre, M.; Wald, L.; Wey, E.; Saboret, L. User's Guide to the CAMS Radiation Service - Status December 2017. URL: https://www.researchgate.net/publication/324542911_User%27s_Guide_to_the_CAMS_Radiation_Service_-_Status_December_2017, accessed 10/12/2020.

Solargis. How to choose the right dataset for evaluation of solar projects — the MASTER approach, 2020. Available from: https://solargis.com/ebook-how-to-choose-solar-resource-data [accessed Jan 12 2021].

Sun, X.; Bright, J.; Gueymard, C.A.; Acord, B.; Wang, P.; Engerer, N. Worldwide performance assessment of 75 global clear-sky irradiance models using Principal Component Analysis. Reviews, 111, 2019, 550-570. https://doi.org/10.1016/j.rser.2019.04.006. Available from: https://www.sciencedirect.com/science/article/abs/pii/S1364032119302187 [accessed Jan 12 2021].

Urraca, R.; Gracia-amillo, A.M.; Koubli, E.; Huld, T.; Trentmann, J.; Riihelä, A.; Lindfors, A.V.; Palmer, D.; Gottschalg, R.; Antonanzas-torres, F. Remote Sensing of Environment Extensive validation of CM SAF surface radiation products over Europe. Remote Sens. Environ. 2017, 199, 171–186.

Yang, D. and Bright, J.M. Worldwide validation of 8 satellite-derived and reanalysis solar radiation products: A preliminary evaluation and overall metrics for hourly data over 27 years. 2020, Solar Energy, 210, 3-19, DOI: 10.1016/j.solener.2020.04.016