Expert quality control of solar radiation ground data sets

Anne Forstinger[1], Stefan Wilbert[2], Adam R. Jensen[3], Birk Kraas[1], Carlos Fernández Peruchena[4], Chris A. Gueymard[5], Dario Ronzio[6], Dazhi Yang[7], Elena Collino[6], Jesús Polo Martinez[8], Jose A. Ruiz-Arias[9], Natalie Hanrieder[2], Philippe Blanc[10], Yves-Marie Saint-Drenan[10]

[1] CSP Services GmbH, Cologne (Germany)
[2] German Aerospace Center, Almería (Spain)
[3] Technical University of Denmark, Kgs. Lyngby (Denmark)
[4] Centro Nacional de Energías Renovables, Sevilla (Spain)
[5] Solar Consulting Services, Colebrook (United States of America)
[6] Ricerca sul Sistema Energetico – RSE S.p.A., Milan (Italy)
[7] Harbin Institute of Technology, Harbin (China)
[8] Ciemat, Madrid (Spain)
[9] University of Málaga, Málaga (Spain)
[10] MINES ParisTech, Sophia Antipolis cedex (France)

Abstract

Ground-based radiation measurements are required for all large solar projects and for evaluating the accuracy of solar radiation models and datasets. Ground data almost always contain low-quality periods caused by instrumental issues, logging errors, or maintenance deficiencies. Therefore, quality control (QC) is needed to detect and eventually flag or exclude such suspicious or erroneous data before any subsequent analysis. The few existing automatic QC methods are not perfect, thus expert visual inspection of the data is still required. In this work, we present a harmonized QC procedure, which is a combination of various available methods, including some that include an expert visual inspection. In the framework of IEA PVPS Task 16, these tests are applied to 161 world stations that are equipped with various radiometer models, and are candidates for an ongoing benchmark of irradiance datasets derived from satellite or weather models. Because the implementation of these methods by experts, and their subsequent decisions, might lead to different QC results, the independently obtained results from nine evaluators are compared for two test sites. The QC results are found similar and more stringent than purely automated tests, even though some deviations exist due to differences in manual flagging.

Keywords: *solar irradiance, solar resource, quality control, quality inspection, visual inspection*.

1. Introduction

Ground-based radiation measurements constitute the basis for solar energy projects and applications. Such measurements are required to evaluate and improve the quality of radiation data derived from satellite retrievals or numerical weather models, and to monitor the performance of solar installations, in particular. The availability of such data is necessary to reach the required accuracy in solar resource data for utility-scale solar projects and for many other applications in the realms of solar power plant operation, solar forecasting, etc. Ground-based data almost always contain periods of low quality and erroneous measurements caused by instrumental errors, maintenance deficiencies, or environment-related inadvertent issues. Therefore, a stringent data quality control (QC) procedure is needed to detect such erroneous (or potentially erroneous) data and ultimately exclude them in high-accuracy applications. QC methods of various kinds have been proposed, e.g., (Espinar et al. 2011; Long and Dutton 2002; Maxwell et al. 1993; Long and Shi 2008). Each one of them consists of a suite of automatic tests. However, automatic tests alone are insufficient as they typically miss certain types of errors and often mislabel valid data. In the vast majority of cases, an expert visual inspection step must be added to automatic QC to obtain the best possible results.

doi:10.18086/swc.2021.38.02 Available at http://proceedings.ises.org

In this work, a harmonized QC procedure is developed in the form of a *"best-of"* method based on a combination of a variety of tests that have already been published and widely recognized, while adding expert visual inspections. Such a QC method is required as an essential first step of an ongoing benchmarking exercise, in which the quality and accuracy of solar irradiance estimates derived from satellite imagery and atmospheric reanalysis are evaluated through preprocessed ground-based measurements. To be meaningful, the benchmarking results must be obtained by comparison with measurements having the lowest possible uncertainty, which can only be obtained after a rigorous QC. The benchmarking exercise is an ambitious project currently being carried out under the auspices of the International Energy Agency's PV Power Systems (PVPS) Task 16. The database of ground measurements consists of several years of data from 161 ground stations, most of which using thermopile pyranometers for global and diffuse horizontal irradiance (GHI, DIF) and an automatically tracked pyrheliometer for direct normal irradiance (DNI). In order to perform QC of such a large database, a group of solar radiation experts from Task 16 was gathered, and several radiometric stations were assigned to each participant. Their QC results may differ to a certain degree for different reasons; for instance, experts might have different opinions on what constitutes a bad data point, they might have varying experience with a specific instrument model or with unusual measurement situations, or, more pragmatically, coding errors may be inadvertently introduced in some cases.

This study reports on the exemplary results obtained by nine different evaluators using data from two radiometric stations selected randomly. The spread of expert-assisted QC results is established to quantify how much they differ from the automatic tests. This makes it possible to evaluate the specific impact of the expert decisions on the overall QC process. In turn, this can help determine to what extent expert assistance contributes to the quality of validation datasets and should be considered as either optional or necessary in practice. The evaluators implemented the QC method individually so that any difference in implementation can be traced back for further improvements and better documentation. The deviations of the results of different evaluators are then compared to the variation in the fraction of usable data for the 161 stations.

The improved QC methodology, including all automatic and visual tests, is presented in Section 2. Section 3 compares the results of the QC undertaken by nine evaluators using data from two randomly chosen stations. In Section 4, conclusions are drawn and a summary is provided.

2. QC Methodology

The QC methodology consists of a combination of tests selected from the literature. In addition to these automatic tests, the evaluators also reviewed and reported the available information on instruments, calibration, maintenance, and records of any special events at each station if such detailed information was available. The visual inspection of such a large database (686 station-years of 1-min data from 161 stations) was an important accomplishment of this QC exercise, at a scale never attempted before.

There are several sets of QC tests that were specifically designed for historic radiation databases, such as BSRN (Long and Dutton 2002), SERI QC (Maxwell, Wilcox and Rymes, 1993), QCRad (Long and Shi 2008), MESOR (C. Hoyer-Klick et al. 2008; Carsten Hoyer-Klick et al. 2009), ENDORSE (Espinar et al. 2011), RMIB (Journée and Bertrand 2011), or MDMS (Geuder et al. 2015). In these methods of the literature, the various tests use different limits for the three individual irradiance components—diffuse horizontal (DIF), direct normal (DNI), and global horizontal (GHI)— as well as parameters derived from these components together with additional quantities such as solar position angles or clear-sky irradiance. The types of limits are:

- physical possible limits
- rare limits
- extremely rare limits.

The existing QC tests have been compared and critically discussed by experts in the framework of IEA PVPS Task 16. Considering the diversity of monitoring stations currently existing in the world, separate methods have been devised, (i) for the ideal case when measurements of all three irradiance components (GHI, DIF, DNI) are available; and (ii) for the case when only two components are directly measured (GHI and DIF or DNI). The latter case is typical of remote solar resource stations that are equipped with a rotating shadowband irradiometer (RSI); see details in Sengupta et al. (2021).

The quality control procedure developed here consists of the combination of automatic tests and a detailed visual inspection performed by an expert. Each test generates a specific flag per timestamp. Each flag can take one of

three possible values: “data point seems ok”, “data point seems problematic”, or “test could not be performed”. The latter situation can occur because of a missing timestamp/data or because the test conditions were not met, and thus the test could not be applied.

The visual inspection of the data is of importance to detect bad data and manually assign a specific flag. This step also includes checking the metadata, if available (logbook with maintenance schedules, issues found, calibration information, comments, etc.). Visual inspection can also help determine if the timestamps refer to the start or the end of the averaging interval (e.g., 1-min, 10-min or 1-h averaging), since this information is not always provided in practice (or can be erroneous). The correct interpretation of the timestamps is essential for practically all QC tests but also for any validation or benchmarking exercise. Furthermore, errors in the correct time zone and coordinates can also only be identified through visual examination by an expert.

The applied QC tests are defined and described in detail below. All test results are visualized using appropriate public-domain software and provide automatically generated flags. Manual flagging is also permitted, thus providing a way to flag data that passed the automated QC tests. Some tests are not automated, but rather consist of purely visual inspections by the evaluators. The applied tests are:

- Missing timestamps
- Missing values
- K-Tests (Geuder et al. 2015; Gueymard 2017)
- BSRN’s closure tests (Long and Dutton 2002)
- BSRN’s extremely rare limits test (Long and Dutton 2002)
- BSRN’s physically possible limits test (Long and Dutton 2002)
- Tracker-off test, improved from (Long and Shi 2008)
- Visual inspection, including shading assessment, closure test, AM/PM symmetry check for GHI, and calibration check using the clear-sky index.

All the automatic tests, as well as the visual review, are discussed in more detail in the following sections.

2.1 Missing timestamps

Skipped timestamps, which might occur during a data logger reset or data acquisition failure, are identified and filled in with the “not a number” date type (NaN). This ensures that, at the end of the QC procedure, all data files provide a continuous information flow.

2.2 Missing values

After filling missing timestamps with NaNs, the total number of missing data can be determined to give an overview of the data completeness of each station.

2.3 *K*-Tests

Various studies, e.g., (Geuder et al. 2015; Gueymard 2017), have defined a number of tests for physical limits and to detect possible tracker issues. These tests are based on the clearness indices K_n, K, and K_t, and their physical relationships. These quantities are defined as

$$K_n = \frac{\mathrm{DNI}}{\mathrm{ETN}} \qquad \text{(eq. 1)}$$

$$K = \frac{\mathrm{DIF}}{\mathrm{GHI}} \qquad \text{(eq. 2)}$$

$$K_t = \frac{\mathrm{GHI}}{\mathrm{ETN}\cdot\cos\left(\mathrm{SZA}\right)} \qquad \text{(eq. 3)}$$

where ETN is the extraterrestrial irradiance at normal incidence, and SZA is the solar zenith angle. The suite of *K*-Tests is applied within each appropriate domain; the corresponding flag names are indicated in Tab. 1. If the condition is not fulfilled within the appropriate domain for one data point, the point is flagged with the corresponding flag name. Because the measured GHI at 1-minute resolution can be much higher than the corresponding clear-sky value during cloud-enhancement periods (Gueymard, 2017), the upper threshold for K_t is adjusted here for the use of 1-min data. It might have to be decreased somewhat for usage at 5-min or 10-min resolution.

Tab. 1: Performed *K*-Tests. ALT denotes the elevation of the site in m (a.m.s.l.).

Condition	Domain	Flag name
$K_n < K_t$	($\mathrm{GHI} > 50\ \mathrm{W/m^2}$ and $K_n > 0$ and $K_t > 0$)	flagKnKt
$K_n < (1100 + 0.03 * \mathrm{ALT})/\mathrm{ETN})$	($\mathrm{GHI} > 50\ \mathrm{W/m^2}$ and $K_n > 0$)	flagKn
$K_t < 1.35$	($\mathrm{GHI} > 50\ \mathrm{W/m^2}$ and $K_t > 0$)	flagKt
$K < 1.05$	($\mathrm{SZA} < 75°$ and $\mathrm{GHI} > 50\ \mathrm{W/m^2}$ and $K > 0$)	flagKlowSZA
$K < 1.1$	($\mathrm{SZA} \geq 75°$ and $\mathrm{GHI} > 50\ \mathrm{W/m^2}$ and $K > 0$)	flagKhighSZA
$K < 0.96$	($K_t > 0.6$ and $\mathrm{GHI} > 150\ \mathrm{W/m^2}$ and $\mathrm{SZA} < 85°$ and $K > 0$)	flagKKt

2.4 BSRN's closure tests

To test the coincidence between the GHI, DNI and DIF irradiance components, i.e., any deviation from the ideal closure, the BSRN closure tests are applied (Long and Dutton 2002). If the conditions described in Tab. 2 are not fulfilled in the noted domain, the data point is flagged with the corresponding flag.

Tab. 2: Performed Three-Component tests

Condition	Domain	Flag name
$\left\lvert \frac{\mathrm{GHI}}{\mathrm{DNI}\cdot\cos(\mathrm{SZA})+\mathrm{DIF}} - 1 \right\rvert \leq 0.08$	($\mathrm{SZA} \leq 75°$ and $\mathrm{GHI} > 50\ \mathrm{W/m^2}$)	flag3lowSZA
$\left\lvert \frac{\mathrm{GHI}}{\mathrm{DNI}\cdot\cos(\mathrm{SZA})+\mathrm{DIF}} - 1 \right\rvert \leq 0.15$	($\mathrm{SZA} > 75°$ and $\mathrm{GHI} > 50\ \mathrm{W/m^2}$)	flag3highSZA

2.5 BSRN's extremely rare limits test

The three irradiance components are also tested in comparison with extremely rare limits to flag any doubtful data (Long and Dutton 2002). If the condition for each component is not fulfilled for a data point, that point is flagged with the corresponding flag name, as described in Tab. 3.

Tab. 3: Performed extremely rare limits tests

Condition	Domain	Flag name
$-2 \leq \mathrm{GHI} \leq 1.2 \cdot \mathrm{ETN} \cdot \cos^{1.2}(\mathrm{SZA}) + 50$	all data	flagERLGHI
$-2 \leq \mathrm{DIF} \leq 0.75 \cdot \mathrm{ETN} \cdot \cos^{1.2}(\mathrm{SZA}) + 30$	all data	flagERLDIF
$-2 \leq \mathrm{DNI} \leq 0.95 \cdot \mathrm{ETN} \cdot \cos^{0.2}(\mathrm{SZA}) + 10$	all data	flagERLDNI

2.6 BSRN's physically possible limits test

In addition to the extremely rare limits, the physically possible limits of each component are tested as well (Long and Dutton 2002). Considering the high-quality requirement for the benchmark application envisioned here, both tests are considered. If the condition for each component is not fulfilled for one data point, the point is flagged with the corresponding flag name (Tab. 4).

Tab. 4: Performed physically possible limits tests

Condition	Domain	Flag name
$-4 \leq \mathrm{GHI} \leq 1.5 \cdot \mathrm{ETN} \cdot \cos^{1.2}(\mathrm{SZA}) + 100$	all data	flagPPLGHI
$-4 \leq \mathrm{DIF} \leq 0.95 \cdot \mathrm{ETN} \cdot \cos^{1.2}(\mathrm{SZA}) + 50$	all data	flagPPLDIF
$-4 \leq \mathrm{DNI} \leq \mathrm{ETN}$	all data	flagPPLDNI

2.7 Tracker-off test

Since, for most stations, the direct and diffuse components are obtained with a tracker equipped with a pyrheliometer and a pyranometer with shading disc or ball, a tracker failure results in incorrect values for both measurements. Such failures include electromechanical problems within the tracker, loss of power, misalignment or timestamp error, etc. Detecting such problems is critical, but can be difficult, particularly in the case of slight mistracking. This specific test involves comparisons with rough estimates of the coincident clear-sky irradiance

components (GHI_{clear}, DNI_{clear}, DIF_{clear}), which are here obtained as a fixed fraction of the extraterrestrial irradiance at horizontal incidence (Tab. 5). If all conditions described in Tab. 5 are not fulfilled for any data point, it is flagged with the corresponding flag name.

Tab. 5: Performed tracker-off tests

Condition	Definitions	Flag name
$\frac{(GHI_{clear} - GHI_{measured})}{(GHI_{clear} + GHI_{measured})} < 0.2$ $\frac{(DNI_{clear} - DNI_{measured})}{(DNI_{clear} + DNI_{measured})} > 0.95$ $SZA < 85°$	$GHI_{clear} = 0.8 \cdot ETH$ $DIF_{clear} = 0.165 \cdot GHI_{clear}$ $DNI_{clear} = \frac{GHI_{clear} - DIF_{clear}}{\cos(SZA)}$	flagTracker

2.8 Visual inspection with a multi-plot

All test results and selected irradiance data are compiled into a single multi-plot arrangement for easy visualization. Such a plot is made for each year at a single station (see, e.g., Fig. 1 for the Visby station, 2016). For a larger example image of the multi-plot please refer to https://github.com/AssessingSolar/solar_multiplot. More specifically, these plots not only include visualization of the test results discussed above, but also (1) visualization of the deviation of the pyrheliometer's DNI from the DNI calculated from DIF and GHI (i.e., closure error); (2) an overview of the diurnal variation of DNI and GHI as a function of time and solar position; (3) the clear-sky index calculated as the ratio between the measured GHI and the clear-sky GHI from the public-domain McClear v3's database (Lefèvre et al. 2013; Gschwind et al. 2019; Qu et al. 2017); (4) a comparison of the pyranometer GHI measurement to the GHI calculated from DNI and DIF; and (5) comparisons of the pyranometer GHI measurements before and after solar noon to identify possible levelling or timestamp errors; (6) visualization of the data points in *K*-space with the applied limits.

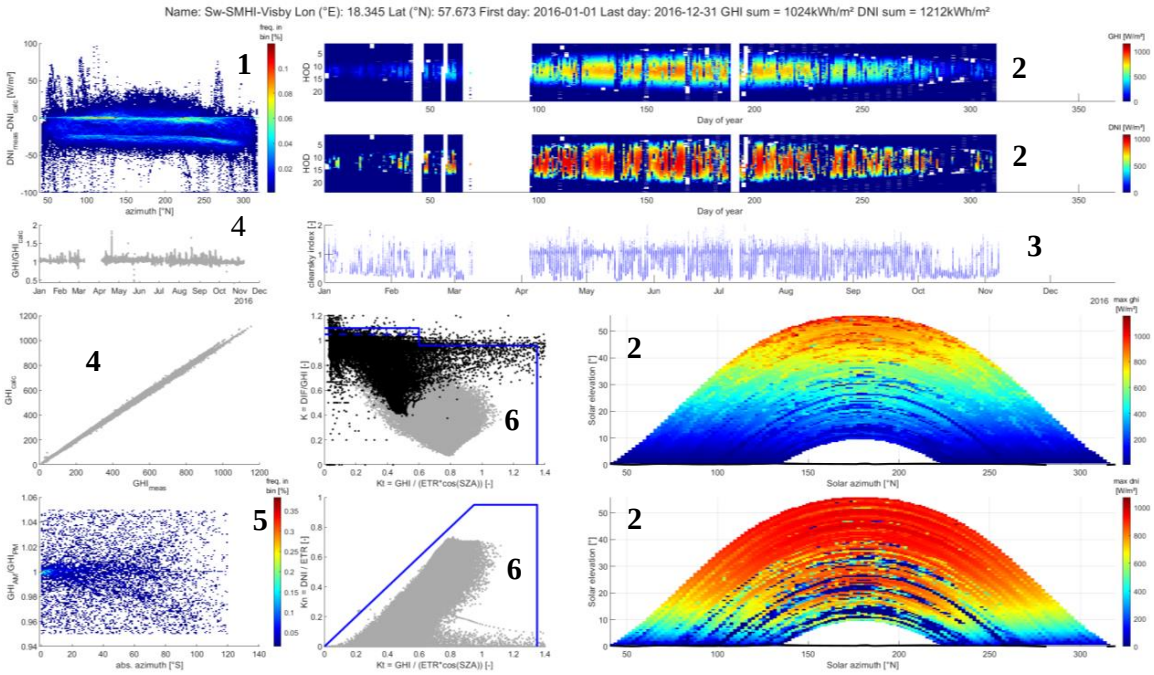

Fig. 1: Visualization of various QC tests used to evaluate the quality of irradiance data at one station (Visby, Sweden, 2016). Numbers in boldface refer to the description in the text.

A multi-plot like the one shown in Fig. 1 is created not only from the raw (pre-QC) data, but also from the data that pass the automatic flagging (as an intermediate result for additional scrutiny), and data that pass the complete QC including manual revision. If suspicious data points are found, further visualizations can be used to confirm whether those points are invalid, in which case an overriding manual flag is set that can be used to exclude such points from processing. For each station and year, the three multi-plots just described offer a complete overview

of the station data and help with the identification of suspicious data or time periods. They serve as a solid starting point for the final manual review.

Plot (1) in Fig. 1 shows the deviation between the measured and calculated DNI with respect to the sun's azimuth angle. In this case, two distinct levels appear over the year. To detect if a specific issue existed at the station (e.g., long periods without cleaning or of tracker issue), one needs further visualization of the data. One example is shown in Fig. 2, which describes the diurnal variation of DNI for each day of a complete year. The day of the year appears on the x-axis, whereas the time of day is shown vertically, using *true solar time* to emphasize the expected symmetry around solar noon. The left plot shows a clearly different deviation pattern in the later part of the year (black rectangle). This is not a station issue per se, but the result of a sensor change. This resulted in a slightly different configuration in terms of levelling and alignment. The right plot of Fig. 2 is for the same station but a different year. It shows a change in deviation caused by sensor soiling, which remained over a long period. While the sensor changes in Fig. 2-left do not lead to an exclusion of the data, the sensor soiling shown in Fig. 2-right can lead to data exclusion. This demonstrates the necessity of manual expert quality control and the general need for station log books, in which cleaning intervals and sensor changes are to be recorded.

The clear-sky index time series (3) is helpful to reveal whether the GHI sensor's calibration is outdated or its sensitivity drifts over time. Ideally, the clear-sky index would be equal to ≈1 under clear-sky conditions. This is rarely the case in the real world, however. (One reason is that GHI_{clear} is only an approximation at any instant.) Nevertheless, cases where the clear-sky index remains constant and well below 1 can be an indication of calibration issue. Similarly, an abrupt or step-like change in that value under clear conditions is typically the signature of a change of calibration factor by a substantial amount. Again, an expert is needed to decide whether a calibration issue is likely at the station. In Fig. 1, the clear-sky index is constant and well below 1 in the later part of the year, but this is the result of cloudy weather conditions rather than a calibration issue. This is apparent when comparing the heat maps of GHI and DNI (plot (2) in Fig. 1), which are placed just above the clear-sky index plot.

If the expert detects issues with individual data points, those are flagged with "flagManual". The results of the individual tests, in the form of a quality flag per test, are properly documented (metadata) and packed into one single file per site and year, which also includes the solar irradiance observations. Finally, all flags are combined into a single usability code, indicating an objective level of quality for each data point.

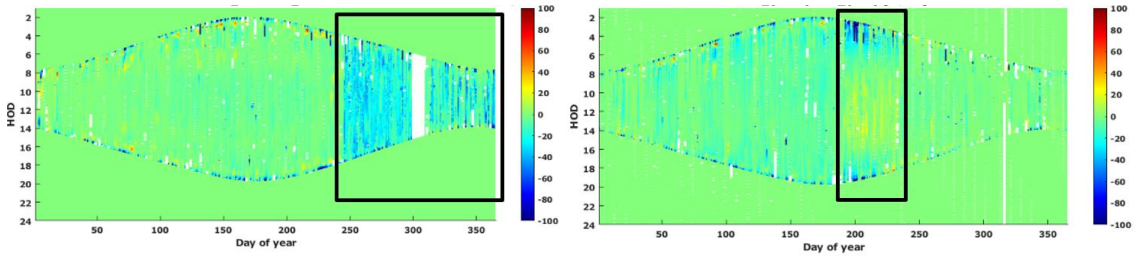

Fig. 2: Two heat maps in W/m² of the difference between DNI measured and DNI calculated on axes of hour of the day (y-axis) and day of the year (x-axis) for different station years. Left: Visby 2015; Right: Visby 2019.

2.9 Usability of a data point

To decide if a data point should be used in any subsequent analysis, not only the result of a single flag per timestamp is needed, but all flags for that timestamp and surrounding timestamps must be considered. For the validation of satellite-based irradiance estimates, and of model-derived radiation data in general, the following procedure is recommended, because it ensures that any data point that passed a test just by chance is excluded from further analysis. Ultimately, a data point is declared "usable" if either it passed all individual QC tests or the test could not be performed while all measured radiation components were available. Moreover, if 30% or more of the timestamps during daytime (i.e., solar elevation >0) from one day are flagged, the entire day is excluded in order to avoid data islands that passed the QC by chance. Intervals between flagged data that are shorter than 60 min are also excluded. For the determination of the length of the interval, only timestamps for sun-up instances are considered. In the case of missing timestamps or missing data in general, the data points are also considered not usable. These applied exclusion rules are quite strict as we exclude for example GHI data for a time stamp although maybe only the DNI measurement was erroneous and as we apply the above described additional rules

to avoid data islands. Less stringent exclusion rules could be applied for other purposes, such as the determination of monthly or yearly sums. These calculations are critically important in solar resource applications, but are often affected by data breaks caused by missing or bad data points. To circumvent this issue, the method outlined in Salazar et al. (2020) is recommended.

3. Results obtained by the nine evaluators

In order to evaluate the influence of the expert decisions and possible variance in the implementation of the method on the QC results, all nine evaluators independently performed their own analysis for the same two stations, Visby (Sweden) and Cairo (Egypt), using five and six years of data, respectively.

3.1 Results for Visby

An overview of the evaluation results for Visby are shown in Fig. 3 as the fraction of flagged sun-up data. The fractions include the flags from automatic tests and manual flagging. The lowest fractions are hence not zero even if an evaluator did not flag any data manually as at least some data points are typically flagged automatically. If at least one test is flagged as non-passed, the timestamp is counted as flagged, i.e., not usable. In terms of total data points being flagged, the overall results show that some evaluators rejected slightly more data than others. In particular, Evaluators 3 and 4—and to a lesser degree, Evaluator 5—were typically much less conservative than the others. The manual flags resulting from the (subjective) visual tests are the cause for these small discrepancies. The overall fraction of flagged data at the Visby station is comparably low.

	flagged_2015	flagged_2016	flagged_2017	flagged_2018	flagged_2019
Eval_1	0.35%	0.61%	0.40%	0.38%	0.28%
Eval_2	0.45%	0.75%	0.52%	0.45%	0.35%
Eval_3	0.16%	0.33%	0.20%	0.20%	0.16%
Eval_4	0.12%	0.38%	0.20%	0.18%	0.14%
Eval_5	0.33%	0.25%	0.30%	0.34%	0.26%
Eval_6	0.51%	0.91%	0.58%	0.44%	0.33%
Eval_7	0.34 %	0.61 %	0.39 %	0.37 %	0.27 %
Eval_8	0.36%	0.60%	0.40%	0.37%	0.28%
Eval_9	0.39 %	0.97 %	1.12 %	0.41 %	0.30 %

Fig. 3: Comparison of QC results from nine evaluators for Visby, showing the number of flagged 1-min data points in percent.

3.2 Results for Cairo

A similar plot as Fig. 3 is shown in Fig. 4 for the Cairo station, which has been installed within the framework of the enerMENA project (Schüler et al. 2016; Hassan et al. 2021). Here, a larger fraction of data points was flagged by all evaluators, and moreover the deviations between different evaluators are higher. Still, Evaluators 3–5 remain the least conservative of them all. The automatic flags accessed by the different evaluators are again the same, so that the deviations are only caused by manual flagging. In the case of year 2016 in particular, Evaluator 1 flagged nearly 5% more data than any other evaluator. The main reason for this has been traced back to a number of temporally short, but periodically occurring shading events caused by objects near the radiometers. After further investigation, the cause of this periodic shading was found to be two guy cables of the windmast at about 5 and 10-m height. They could not be positioned better during the station setup because of space limitations, so that shadows unfortunately appear each morning about half of the year. Considering the short duration of the shading events and the subsequent normal appearance of the DIF, GHI and DNI signals, some evaluators decided to flag longer intervals containing several short shading events, whereas others only flagged individual shading events. It is also possible that some short shading events were not detected, at least by some of the evaluators, even with the automatic tests and the visual inspection. During 2016, short power outages also occurred, explaining why the highest deviations and the highest fractions of flagged data appear in 2016.

The additional exclusion of intervals between flags of less than one hour explains why the effect of these deviations between evaluators are partly removed with the “usability” property of a data point. To avoid missing short intervals with bad data, and to obtain an efficient manual data quality control, flagging longer intervals can be useful. Fewer valid data points then remain, however. Depending on the application of the data, both options can be adequate because rejecting data is a trade-off between the remaining size of the data set and its quality.

	flagged_2015	flagged_2016	flagged_2017	flagged_2018	flagged_2019
Eval_1	3.68%	11.37%	7.38%	3.80%	6.68%
Eval_2	1.69%	6.01%	5.73%	2.73%	5.20%
Eval_3	0.87%	2.35%	2.22%	1.08%	2.04%
Eval_4	0.67%	2.14%	2.10%	1.03%	1.94%
Eval_5	0.99%	3.24%	3.32%	1.51%	2.94%
Eval_6	3.19%	6.92%	5.22%	2.78%	5.20%
Eval_7	2.29%	6.70%	7.01%	3.67%	6.69%
Eval_8	1.63 %	6.04 %	5.63 %	2.72 %	5.22 %
Eval_9	1.44%	4.74%	4.74%	2.62%	5.08%

Fig. 4: Comparison of QC results from nine evaluators for Cairo, showing the number of flagged 1-min data points in percent

As a general rule, it is found that the manual tests operated by an expert can substantially decrease the number of invalid data points in comparison with using automated tests only, and thus decrease the overall uncertainty of a measured dataset. This reduction of the overall uncertainty is the result of excluding the most uncertain data points which might be caused by extraordinary levels of sensor soiling, shading events or sensor malfunction. This operation relies on substantial expertise and takes time, however, and can therefore induce substantial costs in practice.

3.3 Results for the whole database

From a different perspective, the significant deviations found between the rejection rates from different evaluators should also be gauged according to the actual deviations between the usable data for different stations. In the data set of 161 stations used in this study, the whole range from 0 to 100% flagged data points is actually covered in great part, depending also on month and year. To show this more clearly, the fraction of usable data per station and year is analyzed for each continent. The location of the stations, and their total number for each continent, appear in Fig. 5. The database is partly obtained from the Southern African Universities Radiometric Network (Brooks, M.J. et al. 2015), the National Renewable Energy Laboratory (Andreas and Stoffel 1981; Andreas and Wilcox 2012; 2010; Andreas and Stoffel 2006; Vignola and Andreas 2013; Ramos and Andreas 2011), the Baseline Surface Radiation Network (Driemel et al. 2018; Gueymard et al. 2022), and further sources.

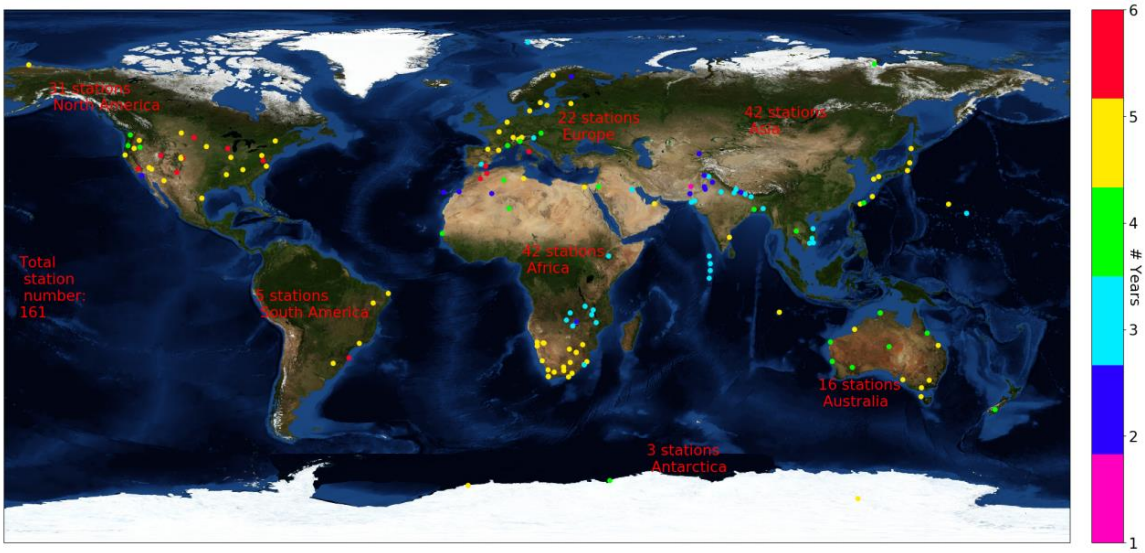

Fig. 5: Map of tested ground stations. The color bar shows the number of years that are used at each station. Source of world map: (Stöckli et al. 2005)**.**

The color of each data point corresponds to the number of calendar years under scrutiny, from one to six. For each station, the maximum period considered in the QC exercise is from 2015 to 2020, inclusive. (Many stations are recent and reported data only between 2015 and 2020, which is why no earlier period was considered for consistency.) The fraction of usable data per year can be seen in Fig. 6, where the average number of usable data

points per year and station is shown as a boxplot for groups of stations within each continent. Only the usable sun-up data points are counted after considering missing data, flags, and the removal of "data islands" (i.e., short periods between flagged data points). The lower and upper quartiles are marked with a box, whereas the full range of the results is marked by whiskers. The star symbol marks the average, and the red line marks the median. The average number of usable data points in each group is very different, and so are the dispersion and the interquartile range. Note that the spread and the interquartile range are also impacted by the number of stations and years within each group. In summary, the covered range of usable data points is relatively large, and varies roughly between 20% and 90%. As could be expected, the usable fraction is consistently low in Antarctica because of the harsh conditions.

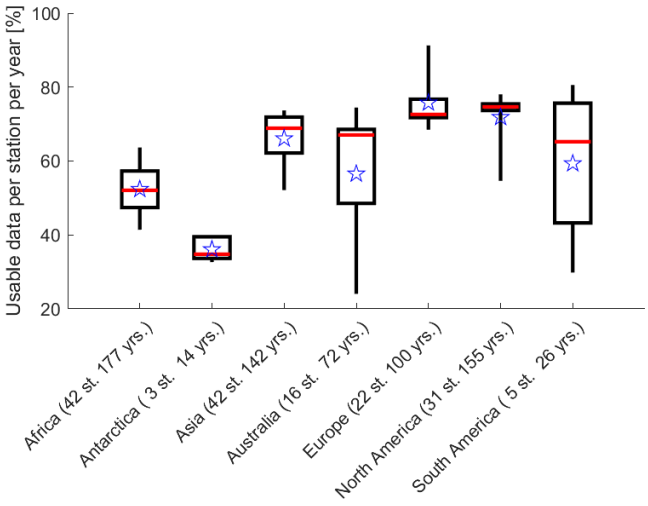

Fig. 6: Box plots of the fraction of usable data points per station and year, grouped by continent. The number of stations per continent and the corresponding total number of calendar years is indicated as well.

Because of time constraints, almost all stations were only evaluated by a single evaluator. In those cases, a detailed comparison similar to that described above for Visby and Cairo is not possible. Nevertheless, the evaluators' results were analyzed on a continental basis in order to at least roughly check for obvious strong deviations that would have been introduced by their manual flagging. Tab. 6 indicates which fraction of stations within each continent was evaluated by each evaluator. Interestingly, Evaluator 1 contributed a number of North American stations that is comparable to that of Evaluators 3–5, but systematically flagged more data than them in the comparison for Visby and Cairo (see Fig. 3 and Fig. 4). The interquartile range and the spread of usable data points for the North American stations are small, as shown in Fig. 6. This does not indicate a too high number of flags by any one of the four main evaluators for North America. Of course, this does not prove that there is no bias in judgment, because the stations tested by Evaluator 1 might have been the most "reliable" stations by chance. In any case, the small spread and interquartile range suggest that the deviation that exists between the results from different evaluators is small compared to the overall variation in data quality.

Tab. 6: Distribution of stations evaluated by each evaluator for each continent.

Evaluator	Africa	Antarctica	Asia	Australia	Europe	N. America	S. America
Eval_1	2%	-	2%	-	-	23%	100%
Eval_2	5%	-	-	-	55%	-	-
Eval_3	-	67%	2%	94%	9%	23%	-
Eval_4	-	-	-	-	-	26%	-
Eval_5	-	-	14%	-	9%	26%	-
Eval_6	19%	-	36%	-	17%	-	-
Eval_7	24%	-	29%	-	5%	-	-
Eval_8	7%	33%	17%	6%	5%	2%	-
Eval_9	43%	-	-	-	-	-	-

4. Conclusion and Outlook

A stringent QC method based on expert consensus was developed based on various tests from the literature, after discussion among a group of international solar radiation experts who participate in IEA PVPS Task 16. The proposed QC method incorporates a suite of up to 15 specific tests, depending on the measurement principle (three independent irradiance components, or only two), and multi-plots describing key variables and covering a whole year of data at each site. The method was successfully applied to 161 radiometric world stations, using up to six years of data per station.

The automated and expert-augmented QC results were compared for two exemplary stations, Visby and Cairo, totaling 10 years of data. The results obtained independently by nine evaluators showed deviations caused by differences in manually set flags, i.e., in their judgment of why a data point should be considered bad or highly suspect. For the Visby station, only few data points were manually flagged, and the evaluators flagged nearly the same data points. In contrast, for the Cairo station, more data points were manually flagged, and the deviation between the different evaluators was much higher. The main reason for this marked deviation was eventually identified as an intermittent shading issue caused by unusual constraints in the station's design. It is expected that such deviations between expert results are small enough in general to not negatively impact demanding applications, such as the radiation data benchmark currently undertaken by the same experts. The deviations could most likely be reduced further if clearer criteria for manual rejection would be defined.

The deviations of the manually flagged data points were also compared to the usable fraction of data produced by the QC procedure for the 161 stations under scrutiny. The deviations between the fraction of data manually flagged by the nine evaluators were found much lower than the station-to-station deviation of the fraction of usable data. This tends to confirm the applicability of the proposed QC method.

An important conclusion is that the visual checks made by experts can substantially improve the quality of a measured dataset and decrease its uncertainty. This suggests that automated tests cannot currently detect all sources of erroneous data. The noted deviation in the manual assignment of quality flags illustrates the inherent subjectivity in discerning whether a data point appears usable or not. In spite of this, not applying any expert-based visual inspection or data control can result in a significant number of overseen erroneous data points, which would be detrimental. Another important conclusion is that further research on even more sophisticated automatic quality control methods would be needed to reduce the effort and costs involved by expert data QC. This could also lead to a lower subjectivity of the quality control.

The QC code is publicly available in order to provide a reference QC tool for researchers and the whole solar industry under this address https://github.com/AssessingSolar/solar_multiplot. For a number of stations used in this study, the quality controlled data including the individual flags is publicly available under this address http://dx.doi.org/10.23646/3491b1a6-e32d-4b34-9dbb-ee0affe49e36
The presented QC procedure constitutes the first step of a benchmark of modeled irradiance data sets that is currently being carried out by experts within IEA PVPS Task 16. The results of that study will be reported subsequently.

5. Acknowledgments

The authors would like to thank NamPower in Namibia, Yuldash Solirov from the Institute of Material Science in Uzbekistan, Frank Vignola from the University of Oregon, David Pozo from the University Jaen, Dietmar Baumgartner from the University of Graz, Julian Gröbner at PMOD, Nicolas Fernay from the University of Lille, Peter Armstrong at the Masdar Institute, Laurent Vuilleumier at MeteoSwiss, Irena Balog at ENEA, Sophie Pelland at CanmetÉNERGIE Varennes, Etienne Guillot at CNRS-PROMES Odeillo, and Majed Al-Rasheedi at the Kuwait Institute for Scientific Research for the provision of data for this study. Furthermore, we gratefully thank the Department of Civil Engineering at the Technical University of Denmark, the Swedish Meteorological and Hydrological Institute, the Australian Government Bureau of Meteorology, the INPE National Institute of Space Research, the CCST Center for Earth System Sciences, along with FINEP Financier of Studies and Projects Ministry of Science and Technology, PETROBRAS Petróleo Brasileiro, the SKYNET organization (in particular Hitoshi Irie and Tamio Takamura (CEReS/Chiba-U.), Chiba University, Tadahiro Hayasaka (Tohoku University), and Chulalongkorn University), as well as the ESMAP program of the World Bank Goup (in particular Joana Zerbin, Clara Ivanescu, Branislav Schnierer, Roman Affolter, GeoSUN Africa, Rachel Fox, and Margot King), the NOAA Global Monitoring Laboratory and the BSRN (in particular for stations 1, 3, 4, 6, 7, 8, 9, 10, 11, 13,

17, 20, 21, 31, 32, 33, 34, 35, 36, 37, 40, 42, 45, 47, 48, 49, 53, 56, 57, 58, 59, 60, 61, 63, 65, 70, 71, 72, 74). We would also like to thank the German Federal Foreign Office for funding and coordinating the enerMENA project., and to express our deep gratitude to the other project partners for their efforts in measuring the meteo-solar data and for agreeing to share their data. These partners are the Cairo University in Egypt, the University of Oujda and Institute Research Solar Energy et Energies Nouvelles (IRESEN) in Morocco, the Research and Technology Centre of Energy (CRTEn) in Tunisia, the University of Jordan in Jordan, and the Centre de Developpement des Energies Renouvelables (CDER) in Algeria.
CSPS and DLR thank the German Ministry for Economic Affairs and Energy for funding their contribution to the study within the SOLREV project (contract number 03EE1010). Adam R. Jensen thanks the Danish Energy Agency for funding his participation (grant number: 64019-0512). Part of this work has been financed by Research Fund for the Italian Electrical System with the Decree of 16 April 2018.

6. References

Andreas, A., Stoffel, T. 1981. NREL Solar Radiation Research Laboratory (SRRL): Baseline Measurement System (BMS); Golden, Colorado (Data). NREL Report No. DA-5500-56488. https://doi.org/10.5439/1052221.

———. 2006. University of Nevada (UNLV): Las Vegas, Nevada (Data). NREL Report No. DA-5500-56509. https://doi.org/10.5439/1052548.

Andreas, A., Wilcox, S. 2010. Observed Atmospheric and Solar Information System (OASIS); Tucson, Arizona (Data). NREL Report No. DA-5500-56494. https://doi.org/10.5439/1052226.

———. 2012. Solar Resource & Meteorological Assessment Project (SOLRMAP): Rotating Shadowband Radiometer (RSR); Los Angeles, California (Data). NREL Report No. DA-5500-56502. https://doi.org/10.5439/1052230.

Brooks, M.J., du Clou, S., van Niekerk,J.L., Gauche, P., Leonard, C., Mouzouris, M.J., Meyer, A.J., van der Westhuizen, N., van Dyk, E.E., Vorster, F. 2015. SAURAN: A New Resource for Solar Radiometric Data in Southern Africa. Journal of Energy in Southern Africa 26: 2–10.

Driemel, A., Augustine, J., Behrens, K., Colle, S., Cox, C., Cuevas-Agulló, E., Denn, F., Duprat, T., Fukuda, M., Grobe, H., Haeffelin, M., Hodges, G., Hyett, N., Ijima, O., Kallis, A., Knap, W., Kustov, V., Long, C., Longenecker, D., Lupi, A., Maturilli, M., Mimouni, M., Ntsangwane, L., Ogihara, H., Olano, X., Olefs, M., Omori, M., Passamani, L., Bueno Pereira, E., Schmithüsen, H., Schumacher, S., Sieger, R., Tamlyn, J., Vogt, R., Vuilleumier, L., Xia, X., Ohmura, A., König-Langlo, G. 2018. Baseline Surface Radiation Network (BSRN): Structure and Data Description (1992–2017). Earth Syst. Sci. Data 10: 1491–1501. https://doi.org/10.5194/essd-10-1491-2018.

Espinar, B., Wald, L., Blanc, P., Hoyer-Klick, C., Schroedter Homscheidt, M., Wanderer, T. 2011. Project ENDORSE - Excerpt of the Report on the Harmonization and Qualification of Meteorological Data:Procedures for Quality Check of Meteorological Data. https://hal-mines-paristech.archives-ouvertes.fr/hal-01493608.

Geuder, N., Wolfertstetter, F. Wilbert, S., Schüler, D., Affolter, R., Kraas, B., Lüpfert, E., Espinar, B. 2015. Screening and Flagging of Solar Irradiation and Ancillary Meteorological Data. Energy Procedia 69 (May): 1989–98. https://doi.org/10.1016/J.EGYPRO.2015.03.205.

Gschwind, B., Wald, L., Blanc, P., Lefèvre, M., Schroedter-Homscheidt, M., Arola, A. 2019. Improving the McClear Model Estimating the Downwelling Solar Radiation at Ground Level in Cloud-Free Conditions - McClear-V3. Meteorologische Zeitschrift 28 (2): 147–63. https://doi.org/10.1127/metz/2019/0946.

Gueymard, C.A. 2017. Cloud and Albedo Enhancement Impacts on Solar Irradiance Using High-Frequency Measurements from Thermopile and Photodiode Radiometers. Part 1: Impacts on Global Horizontal Irradiance. Solar Energy 153 (September): 755–65. https://doi.org/10.1016/J.SOLENER.2017.05.004.

Gueymard, C. A., Bright, J. M., Sun, X., Augustine, J., Baika, S., Brunier, L., Colle, S., Cuevas-Agulló, E., Dehara, K., Denn, F. M., Duprat, T., Haeffelin, M., Kallis, A., Knap, W., Kustov, V., Lupi, A., Maturilli, M., Ntsangwane, L., Ogihara, H., Olano, X., Omori, M., Pereira, E. B., Ramanathan, K., Schmithüsen, H., Tully, M., Vogt, R., Vuilleumier, L. 2022. BSRN Data Set for IEA-PVPS Task-16 Activity 1.4 Quality

Control. PANGAEA. https://doi.org/10.1594/PANGAEA.939988.

Hassan, M. A., Khalil, A., Abubakr, M. 2021. Selection Methodology of Representative Meteorological Days for Assessment of Renewable Energy Systems. Renewable Energy 177: 34–51.

Hoyer-Klick, C., Beyer, H.G., Dumortier, D., Schroedter-Homscheidt, M., Wald, L., Martinoli, M., Schilings, C., Gschwind, B., Menard, L., Gaboardi, E., Ramirez-Santigosa, L., Polo, J., Cebecauer, T., Huld, T., Suri, M., Blas, M. de, Lorenz, E., Pfatischer, R., Remund, J., Ineichen, P., Tsvetkov, A., Hofierka, J. 2008. Management and Exploitation of Solar Resource Knowledge. In EUROSUN 2008, 1st International Conference on Solar Heating, Cooling and Buildings, Lisbon, Portugal.

Hoyer-Klick, C., Beyer, H., Dumortier, D., Wald, L., Schillings, C., Gschwind, B., Menard, L., Gaboardi, E., Polo, J., Cebecauer, T., Suri, M., Blas, M De., Lorenz, E., Kurz, C., Remund, J., Ineichen, P., Tsvetkov, A 2009. MESoR – MANAGEMENT AND EXPLOITATION OF SOLAR RESOURCE KNOWLEDGE. In SolarPACES 2009, Berlin : Germany (2009).

Journée, M., Bertrand, C. 2011. Quality Control of Solar Radiation Data within the RMIB Solar Measurements Network. Solar Energy 85 (1): 72–86. https://doi.org/10.1016/j.solener.2010.10.021.

Lefèvre, M., Oumbe, A., Blanc, P., Espinar, B, Gschwind, B., Qu, Z., Wald, L., Schroedter-Homscheidt, M., Hoyer-Klick, C., Arola, A., Benedetti, A., Kaiser, J. W., Morcrette, J. J. 2013. McClear: A New Model Estimating Downwelling Solar Radiation at Ground Level in Clear-Sky Conditions. Atmospheric Measurement Techniques 6 (9): 2403–18. https://doi.org/10.5194/amt-6-2403-2013.

Long, C. N., Dutton, E.G. 2002. BSRN Global Network Recommended QCtests, V2.0. Baseline Surface Radiation Network. https://bsrn.awi.de/fileadmin/user_upload/bsrn.awi.de/Publications/BSRN_recommended_QC_tests_V2.pdf.

Long, C. N., Shi, Y. 2008. An Automated Quality Assessment and Control Algorithm for Surface Radiation Measurements. The Open Atmospheric Science Journal 2: 23–37.

Maxwell, E, Wilcox, S., Rymes, M. 1993. Users Manual for SERI QC Software, Assessing the Quality of Solar Radiation Data. Solar Energy Research Institute, Golden, CO.

Qu, Z., Oumbe, A., Blanc, P., Espinar, B., Gesell, G., Gschwind, B., Klüser, L., Lefèvre, M., Saboret, L., Schroedter-Homscheidt, M., Wald, L. 2017. Fast Radiative Transfer Parameterisation for Assessing the Surface Solar Irradiance: The Heliosat-4 Method. Meteorologische Zeitschrift 26 (1): 33–57. https://doi.org/10.1127/metz/2016/0781.

Ramos, J., Andreas, A. 2011. University of Texas Panamerican (UTPA): Solar Radiation Lab (SRL); Edinburg, Texas (Data). NREL Report No. DA-5500-56514. https://doi.org/10.5439/1052555.

Salazar, G., Gueymard, C., Bezerra Galdino, J., de Castro Vilela, O., Fraidenraicha, N. 2020. Solar Irradiance Time Series Derived from High-Quality Measurements, Satellite-Based Models, and Reanalyses at a near-Equatorial Site in Brazil. Renewable and Sustainable Energy Reviews 117 (109478).

Schüler, D., Wilbert, S., Geuder, N., Affolter, R., Wolfertstetter, F., Prahl, C., Röger, M., Schroedter-Homscheidt, M., Abdellatif, G., Guizani, A. Allah, Balghouthi, M., Khalil, A., Mezrhab, A., Al-Salaymeh, A., Yassaa, N., Chellali, F., Draou, D., Blanc, P., Dubranna, J., Sabry, O. M.K. 2016. The EnerMENA Meteorological Network - Solar Radiation Measurements in the MENA Region. AIP Conference Proceedings 1734 (May 2016). https://doi.org/10.1063/1.4949240.

Sengupta, M., Habte, A., Wilbert, S., Gueymard, C., Remund, J. 2021. Best Practices Handbook for the Collection and Use of Solar Resource Data for Solar Energy Applications: Third Edition. National Renewable Energy Laboratory. https://www.nrel.gov/docs/fy21osti/77635.pdf.

Stöckli, R, Vermote, E., Saleous, N., Simmon, R., Herring, D. 2005. The Blue Marble Next Generation - A True Color Earth Dataset Including Seasonal Dynamics from MODIS. NASA Earth Observatory.

Vignola, F., Andreas, A. 2013. University of Oregon: GPS-Based Precipitable Water Vapor (Data). NREL Report No. DA-5500-64452.