Contents lists available at ScienceDirect

Solar Energy

journal homepage: www.elsevier.com/locate/solener

Data Article

Structure of a comprehensive solar radiation dataset

Josh Peterson *, Frank Vignola

Solar Radiation Monitoring Lab, University of Oregon, Eugene, OR, United States

ARTICLE INFO

Keywords
Solar radiation data
Resource assessment
Uncertainty

ABSTRACT

A new, comprehensive, file format has been developed for solar and other meteorological data from the University of Oregon's Solar Radiation Monitoring (UO SRML) network in the Pacific Northwest. In an effort to provide essential information that would enable users to assess the quality and uncertainty in the data, a more comprehensive file format was developed that contains significantly more information about the station and the various instruments used to make the measurements. The file format utilizes month blocks and starts with a header containing detailed information about the site location, instruments used, calibration values utilized, and uncertainties in the calibration values. The second region of the file contains daily metadata for the instruments and useful information about the extraterrestrial irradiance and average nighttime offsets. After the metadata come the short-term data values and associated flags that help describe the status of the data. In addition, a variety of time stamps are used to facilitate the use of the data. The format also contains room for comments about the data, intended to help users see what was done to the data during the analysis process. Data in this comprehensive format is available on the UO SRML website at.

(http://solardat.uoregon.edu/SelectArchivalUpdatedFormat.html).

1. Introduction

Searching the internet one can find a wide variety of sources for solar radiation data from several satellite derived solar radiation databases, to various ground-based measurements including: Baseline Surface Radiation Network, the Surface Radiation Budget network (SURFRAD), and the University of Oregon Solar Radiation Monitoring Network (UO SRML) to name a few. These solar radiation data bases have a myriad of uses from validating solar models to predicting and characterizing the performance of solar energy systems. The quality and accuracy of the data from various networks varies and many questions arise when using the data that are available. Now that large scale solar electric systems are being developed and deployed, the uncertainty and quality of the data in databases are becoming significantly more important.

Any measurement or data point has an uncertainty associated with the value and just saying that the irradiance has a certain value, for example 1000 W/m^2, doesn't fully characterize the value. Not only are the type of measurement, the time, and location important but the uncertainty in the measurement is needed to specify the measurement. Systematic efforts are underway to provide a uniform method to specify the uncertainties in measurements, see the Guide to Expected Uncertainty in Measurements (G.U.M) found at (Bureau International des Poids et Mesures (BIPM) et al., 1995). In order to use the G.U.M. methodology, information about the measurements are required. Previous efforts and databases have strived to provide the data values, but it is also important to supply auxiliary information with the data to reliably estimate the uncertainty in the databases. As the need for a more comprehensive database has been become apparent, the UO SRML has started to enhance its database with information needed to better assess the data. This paper discusses what is being put into the more comprehensive database and the rational for the information included in the database.

Irradiance data gathered by the University of Oregon Solar Radiation Monitoring Laboratory (SRML) has been available on the internet for many years. By having the data available online, users do not have to contact the lab directly for data unless they have specific questions. Originally the data format was designed to be compatible with the Research Cooperator Format designed by the Solar Energy Research Instituted (SERI) for network data sharing (SERI, 1988). At the time of its development, disk space was limited and files had to be compact. Therefore, the files consisted of ASCII tab-separated integers. Associated with the data files are documents that explained the file structure and what is contained in the files. An overview of the history of the SRML is given in the appendix where more information on some of these early

* Corresponding author.
E-mail address: jpeters4@uoregon.edu (J. Peterson).

https://doi.org/10.1016/j.solener.2020.08.092
Received 22 May 2020; Received in revised form 29 August 2020; Accepted 29 August 2020
Available online 4 October 2020

decisions is given. Today, such a file structure is archaic and difficult to use. In addition, solar data are now used by developers and financial institutions to evaluate operation and fiscal performance of solar facilities. More detailed information about the dataset is required to provide confidence in the analysis that result from using the data. This information is essential when building a "bankable" dataset (Vignola et al., 2012).

This article describes the enhanced data file format currently used by the UO SRML. It also provides guidance that can be used to create more comprehensive solar data files that help clarify the quality and uncertainty in the dataset. More specifically, the files contain information on the instruments, calibrations used, and the uncertainty associated with calibrations and the data.

Many data users are only interested in irradiance values. However, if the data are to be used for validating models or satellite derived irradiance values, estimating system performance, or forecasting the solar resource, more detailed information is necessary to ascertain the confidence level of the input and output values. In an effort to provide essential information that would enable users to assess the quality and uncertainty in the data, the comprehensive file format was developed that contains significantly more information about the station and the various instruments used to make the measurements. This information is now contained in the data file itself and it is not necessary to seek other files to find this information. The new file format contains information on the specific instruments used to make the measurements including their model and serial number as well as calibration values used to translate voltage measurements into irradiance values and the uncertainty in the calibration values. Basic information about the monitoring station are included in the header and auxiliary information such as the solar zenith and azimuthal angles are embedded with the short interval data to facilitate the use of the data.

2. Data file structure - overview

The files are separated into month blocks to maintain a manageable file size. The files contain metadata in terms of daily total or average values followed by the short time interval data. A schematic diagram of the new format is shown in Fig. 1. This article will discuss each of the areas shown in the schematic in the following order. Region 1 contains general information about the station. Region 2 contains information about the various instruments that are used. Regions 3, 4, 5 contain daily total information about each instrument in a daily summary table. Regions 6, 7, 8 contain the short time interval data for each instrument. Region 9 provides room for comments about the data in the file.

The daily total and short time interval regions are subdivided into three parts (left, center, right) using the following metric: The left columns contain non-measured quantiles such as: date, time, solar zenith angle, azimuthal angle, sunrise time, sunset time, extraterrestrial radiation, etc. The center columns contain processed and calculated irradiance information. The right columns contain the original measured irradiance quantities and other meteorological information such as temperature and air pressure.

Regions 3, 4, 5 are metadata calculated at the end of a completed day. The short time interval information in regions 6, 7, 8 are the measured or calculated data and are generated as data are input into the file. The short interval data are used to generate the metadata. Comments can be added in region 9 if deemed appropriate when the data in the file is analyzed. The contents of the file are discussed in the order they appear in Fig. 1.

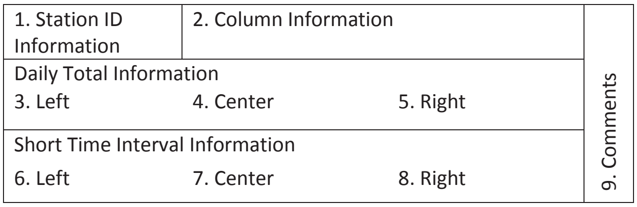
1. Station ID Information
2. Column Information
Daily Total Information
3. Left
4. Center
5. Right
Short Time Interval Information
6. Left
7. Center
8. Right
9. Comments

Fig. 1. File structure overview. The different regions of the file are labeled 1–9.

2.1. *About the data in the file*

Before going into details on how the data are stored, it is important to understand the data going into the file. The data in the files come from data loggers attached to radiometers, meteorological instruments, and other sensors. The raw data is gathered and a program inputs the data into the comprehensive file format. Header information is input and represents that status of the instrumentation at the beginning of the month. The raw data are then analyzed and faulty data are identified and marked as questionable, bad, or adjusted, if this can be done in a reasonable manner. Examples of faulty data include cleaning of the instrument, soiling, ice or snow on the instrument, or misalignment of the pyrheliometer. For data considered "bad", irradiance values from other collocated instruments are used to replace the faulty data if the auxiliary data are available.

Once the data are analyzed and edited, the irradiance information is processed to remove some of the systematic biases. For Rotating Shadowband Radiometers (RSR)s, specific algorithms are used to remove the systematic biases (Vignola, 2006). For broadband thermopile pyranometers which measure the irradiance with one thermopile attached to a black disc and another to a thermal sink, the black sensor radiates to the dome of the pyranometer and hence to the night sky. This creates a thermal loss and reduces the measured irradiance. The nighttime thermal offset is an indication of the magnitude of the problem, although daytime thermal offsets can be twice as large as the nighttime values (Vignola et al., 2009). Modern pyranometers have significantly small thermal offsets, but they still exist. To partially adjust for this bias, the average nighttime values are subtracted from the daytime values removing some of this bias in the data. This subtraction of the average nighttime values is also applied to other radiometers for consistency. This process is not applied to meteorological data and other data such as sensors that monitor photovoltaic systems.

The comprehensive format keeps the original edited values and the processed values. This enables users to evaluate other models to remove the biases or to use our best estimate of the irradiance values.

3. Description of file structures by regions

The nine areas of the file structure shown in Fig. 1 are discussed in detail. Sample information is given to illustrate information in each region.

3.1. *File structure region 1 – Station ID information*

The upper left corner of each file contains useful information about

Table 1
A sample of the station information for the data contained in Region 1 of the file structure.

Labels	Station values
Station ID Number:	94255
Station Name:	EUO
Station Location:	Eugene_Oregon_USA
Latitude:	44.046761
Longitude (+East):	−123.074243
Altitude (m):	150
Time Zone (+East):	−8
Time Interval (Minutes):	1
Year//Month	2016//12

the file. An example for the Eugene, Oregon station is shown below in Table 1.

- The station ID number was originally a WBan number (NWS Website) and were obtained from the National Weather Service for the stations. Once photovoltaic monitoring stations were added to the network, this practice was abandoned and numbers in a similar format were added as needed. This number is given in the upper left corner of the of the original research cooperator format.
- A shorthand station name was designated for each station. This three-letter code is a short hand notation for each station with the first two letters indicating the city location and the final letter representing the state (O = Oregon, W = Washington, I = Idaho, U = Utah, Y = Wyoming, M = Montana)
- Station Location is the City, State, and country name of the station. The three names are separated by an underscore "_". The use of spaces is discouraged to facilitate computer reading of the information.
- Latitude, Longitude, and altitude of the station. The latitude and longitude are expressed as decimal degrees and are given to an accuracy of the ±200 m. By convention, longitudes west of the prime meridian are negative. Locations east of the prime meridian are positive. Similarly, latitude north of the equator are positive and south of the equator are negative. The altitude of the station is given in meters above sea level.
- The time zone of the station. The time zone is useful for calculating the sun's position in the sky. Similar to longitude, time zones west of the prime meridian are negative. For Eugene, Oregon, the time zone is −8 meaning that the sun will rise 8 h after it rises at the prime meridian. Times are expressed in standard time. Daylight savings time is never used.
- The time interval, given in minutes, is the step size between each data point. The data are usually summed or averaged over the time interval. For one-minute data, the time stamp of 11:31 would contain data gathered from 11:30:01 through 11:31:00. Some instruments such as a rotating shadowband radiometer (RSR) or spectroradiometers are sampled only once per minute. The time interval indicates that the information was gathered in the 11:30 to 11:31 time interval. Early measurements contained hourly averaged values or a time interval of 60 min. Many high quality stations today have a time interval of 1 min. The time format is structured to allow time intervals shorter than one-minute in the future.
- The year and month of the file block are separated by double forward slash marks "//". This technique prevents some programs, such as Excel, from auto reformatting dates and times into their preferred format. By using the double forward slash, the information will not be recognized as a date and the format of the file will be preserved. This technique is also done using double colons "::" to separate the hours from the minutes when giving a time. For a similar reason, quotation marks "" are not used because they are interpreted in several ways by different programs. If dates and times are preferred without the "//" or "::", then a bulk remove or replace can be used.

3.2. File structure region 2 – Column header information

The header rows in the column information region contain information about each column. There are 10 rows of information. There are roughly twice as many columns as there are instruments (with some exceptions and additions). In the upcoming section, first the rows will be discussed. Then variations to the different columns will be discussed. Table 2 is a subsample of some common data headers.

Each instrument has a data file column and a quality control flag column. The header rows for these two types of columns are different.

- **Type of measurement:** The type of measurement that is made in this column. In the above example, GHI corresponds to Global Horizontal Irradiance, DNI corresponds to Direct Normal Irradiance, and DHI corresponds to Diffuse Horizontal Irradiance. The Direct Horizontal Irradiance is denoted as DrHI, not to be confused with the Diffuse Horizontal Irradiance (DHI). The notation for other options uses common abbreviations and is intended to be self-descriptive to the user. The columns labeled as xxx_Flag contain information about the data in the previous column. This information can be a flag describing the quality of the data or an uncertainty in the data value. A more detailed discussion is contained in Section 3.7.
- **Element:** A numeric value relating to the type of measurement based on definitions of the research cooperator format. Now that storage space is less of a premium, a more readable description is preferred. The element numbers are included here as a link between our original file format and the new comprehensive format. Some of UO SRML data files are forty years old and have not been updated to the comprehensive file format. A complete description of the various element numbers are available at http://solardat.uoregon.edu/DataElementNumbers.html.
- **Instrument Serial Number:** To provide traceability, the model name and the serial number of each instrument is recorded. If an instrument is changed during the month, the day and time of the change should be noted in the comment section of the header. Some columns contain calculated data and are identified as such instead of listing the serial numbers of the instruments used. If instruments are changed during a month period, then the change and replacement instrument is noted in the comment column. Except for temporary replacements the general practice is to keep the same instrument in place year after year. This practice helps in evaluating relative long-term trends. Sometimes instruments do have to be replaced and this is noted in the files. This was a major problem with in the old files that just had the element number listed. No information in the data file was specified when or if an instrument was changed.
- **Shorthand name:** When discussing instruments at various stations, it is easier to remember and say P23 than identify the instrument by the serial number, 23973F3 for the example in Table 1. This is

Table 2
Sample of the data contained in the column headers region 2.

Type of Measurement:	GHI	GHI_Flag	GHI_Calc	GHI_Calc_Flag	GHI_original	GHI_original _Flag
Element:	1000	–	1001	–	1000_original	–
Instrument Serial Number:	CMP22(110265)	–	Computed from DHI and DNI	–	CMP22(110265)	–
Instrument Shorthand Name:	CMP22	–	NA	–	CMP22	
Responsivity:	8.9182	$\frac{microV}{Wm^{-2}}$	NA	$\frac{microV}{Wm^{-2}}$	8.9182	$\frac{microV}{Wm^{-2}}$
Estimated Uncertainty (U95%):	2.049	–	3.339	–	2.049	–
Sample Method:	Avg	–	–	–	Avg	–
Units:	$W{\cdot}m^{-2}$	–	$W{\cdot}m^{-2}$	–	$W{\cdot}m^{-2}$	–
Column Notes:	AdjustedColumn	–	CalculatedColumn	–	MeasuredColumn	–

pyranometer P23, that is the 23rd instrument that was used in the network. While the serial number is important for keeping track of the instrument, especially during calibrations, having a shorthand name is useful. The name is assigned to each instrument upon purchase. The shorthand name is not related to the serial number and can be independent from any manufacturer information.

- **Responsivity:** The responsivity is the calibration constant that was used to convert the raw measured voltage (or millivolts) to irradiance values. Responsivity is typically defined at a single or range of solar zenith angles. As many responsivities change over time, it is extremely helpful to know what responsivity was used when the data of interest is put to use.

The formula relating voltage to irradiance is given by Equation (1).

$$\text{Irradiance} = \frac{\text{Voltage}}{\text{Responsivity}} \tag{1}$$

The voltage of each measurement is not recorded, only the corresponding irradiance and responsivity are recorded. The SRML characterizes each instrument at various angles of incidence and this information is available upon request. Future plans call for posting the calibration records for each instrument on the UO SRML website and these files are usually updated annually. In the comment section there is room to reference the calibration records.

- **Estimated uncertainty:** An estimated uncertainty (at the 95% level of confidence) is the value reported in that column. Responsivity values are computed at an angle of incidence of 45° using a methodology based on pyranometer calibrations made using the Broadband Outdoor Radiometer Calibration methods (BORCAL) prior to the year 2015 as discussed by Wilcox et al., 2002. The estimated uncertainty of the measurement is determined using the Guide to the Expression of Uncertainty in Measurement (GUM) methodology (Bureau International des Poids et Measures (BIPM) et al., 1995). The uncertainty includes many factors including the calibration uncertainty, deviation from true cosine response, effect of temperature, etc.

The instruments in these files have different characteristics, biases and uncertainties. The uncertainties quoted are an attempt to identify the quality of the data and help select the best instrument to use. However these uncertainties apply only to a narrow range of conditions and uncertainties outside the defined ranges of the uncertainties can significantly affect the measurements. For example, during the winter, the solar zenith angles (SZA) can be large and outside the range of SZA used to determine the quoted uncertainties. In general the instruments with the least uncertainties are listed first and auxiliary instruments are labeled with the type of measurement followed by the _Aux

- **Sample Method:** The method used to generate the data is listed. Irradiance data is typically measured once ever second or two and averaged over the time interval. Some older data sets were obtained using data loggers that produced integrated values. The phrase "averaged" means that the data was integrated or sampled every one to three seconds over the time period of interest. Information about the exact sample rate is not given in the file. Certain sensors (such as a Rotation Shadowband Radiometer (RSR)) only make a measurement once per time interval. For sensors such as this, the measurement method would be labeled as "instantaneous".
- **Units of each measurement:** Typical units for irradiance are W·m^{-2} (W/m^2). In the data files, the carrot symbol ^ is used to describe a number raised to a power. Typical units for Temperature are Celsius (C).
- **Rows 9 – 10:** These two rows allow for notes about each column. Notes may include information about RSR sensors e.g., when a sensor was changed. These columns are not as strictly defined and are a place for the user/editor to make notes about the various columns.

This completes a general overview of the typical information of the header rows. The discussion was generic and only the most common types of measurements were used as examples. There are four different types of data: processed data, calculated data, meteorological data, and unprocessed measured data. Some data are processed to remove some of the systematic biases or to calculate other irradiance information. The method used to process and adjust the data are discussed in Section 3.6. The calculation method of the measured data are discussed in Section 3.7. Each measurement is represented with a pair of columns. The first column contains the measurement value and the second column corresponds to a quality control flag. Flags with the "Process data" are discussed at the end of Section 3.7.

- **Unprocessed-measured data:** Each irradiance measurement is presented in the form that it comes from the data logger after it has been modified by the appropriate responsivity. The raw data is analyzed for problems and the problems are addressed or the data are flagged questionable or bad as appropriate (see Section 3.7 for information on the flags used in these files). If available information from auxiliary instruments may be substituted for the faulty data. Raw data is often differentiated from processed data using the notation "_original" following the type of measurement name, for example GHI_original. This implies that the column still has the nighttime offset included. Bad data has been flagged, removed, or replaced by edited data. The associated flag column identifies any changes to the raw data.
- **Processed Data:** This corresponds to irradiance data that has been evaluated and possibly adjusted to remove some systematic biases. For example, adjustments for radiative losses are determined by evaluating the nighttime offsets that are subtracted during the day and adjustments to RSR measurements to account for systematic deviations. The adjustment algorithm will be discussed in Section 3.7. Data labels are found in the first row of each column. Examples for commonly encountered adjusted data labels include GHI, DNI, DHI, GTI_Tilt_Azm, and GHI_Auxiliary. The processed data represent the best available data in the file.
- **Computed Data:** These are data that have been computed using one or more processed data sets. Data sets that have been computed are given the notation "_Calc" for example GHI_Calc. A commonly computed column is the computation of the GHI from DNI and DHI using the formula

$$GHI\_Calc = DNICos[SZA] + DHI \tag{2}$$

Where SZA corresponds to the solar zenith angle. The formula used to calculate the SZA is discussed in Section 3.6. GHI_Calc is the calculated global horizontal irradiance, DNI is the processed direct normal irradiance, and DHI is the processed diffuse horizontal irradiance.

Another common example is the computation of the direct horizontal irradiance (DrHI) from a direct normal irradiance using the formula.

$$DrHI\_Calc = DNI{*}Cos[SZA] \tag{3}$$

The uncertainty of calculated columns is computed by combining the uncertainty of the various components using the GUM model (BIPM et al., 1995).

- **Meteorological Data:** Meteorological measurements include such information as air temperature and atmospheric pressure. While meteorological data are evaluated for problems, they are not processed or adjusted like irradiance data.
- **Photovoltaic Data:** Several data files contain data from photovoltaic (PV) arrays. Unless calibration problems with sensors are found, the photovoltaic data comes directly from the data logger. Shading,

snow cover, soiling, malfunctioning photovoltaic equipment, are not typically not flagged as bad. This data are to provide actual PV system performance information and effects such as snow cover or soiling can be studied using this data.

3.3. File structure region 3. Daily total information Sunrise/sunset/solar noon time, daily total ETR, ETRn radiation

The daily total information are contained in Rows 12 – 42 for each instrument in the file. This daily metadata serves as an overview of the month's weather and irradiance conditions.

The first seven columns of the daily total portion of the file include the day of month, day of year, sunrise and sunset times, solar noon times, daily total extraterrestrial radiation on a horizontal surface (ETR), and daily total extraterrestrial radiation on a normal surface (ETRn). In Fig. 1 this is designated as Region 3.

The sunrise, sunset, and solar noon times are given in columns 3–5 and are written in the following format (hh::mm:ss). The double colon is used to separate the hours from the seconds in an effort to prevent spreadsheet programs from auto formatting the time information. The sunset and sunrise times are good to ±30 s and do not account for obstructions on the horizon at the site. Sunrise and sunset occur when the apparent disk of the sun is completely below the horizon (SZA = 90.267°). Solar noon is defined as when azimuthal angle of the sun is at AZM = 180°. The SRML network operates exclusively in the Northern Hemisphere where this azimuthal condition is always true.

The daily total extraterrestrial radiation (Column 6) is a measure of the total energy incident on one square meter horizontal surface outside the atmosphere in one day. The ETR is measured in kW h/m^2, with 1 kW h/m^2 = 3,600,000 J/m^2. The ETR is computed using the following formula.

$$\mathrm{ETR} = \frac{\mathrm{time\ \ interval}}{60{*}1000}\sum_{\mathrm{i}}\mathrm{IRR_i} \tag{4}$$

where $\mathrm{IRR_i}$ is the individual extraterrestrial irradiance values (ETR) reported throughout the day. The time interval is the time interval of the data set given in minutes. The radiation incident on a normal surface (ETRn) is computed using a similar formula and is given in Column 7. Time intervals that encompass the sunrise and sunset times are scaled accordingly.

3.4. File structure region 4. Daily total or average information

The total energy for the processed and computed data sets are computed for each day. The total energy for a day is computed using Equation (5) where the $\mathrm{IRR_s}$ is the individual irradiance values reported throughout the day for a given instrument.

$$\mathrm{Daily\ \ Total\ \ Energy} = \frac{\mathrm{time\ \ interval}}{60{*}1000}\sum_{\mathrm{s}}\mathrm{IRR}_s \tag{5}$$

Missing and bad data points are interpolated using a linear fit. If more than one hour of data are missing or flagged bad, then the daily total for that day is not computed and that cell is left blank. A dash can also be used to indicate missing data.

An uncertainty estimate of the daily values is also given for each day. The uncertainty estimate uses the uncertainty in the instrument's responsivity. Data points that are edited or questionable (Flags 22 through 82) are given twice the uncertainty. The uncertainty is given at the 95th level of confidence. The units of the uncertainty are in kW h/m^2. At this point the uncertainty estimate does not include uncertainties associated with variations in the responsivity of the instrument due to changes in the angle of incidence, temperature, or spectral response of the instrument. These changes may be significant and the uncertainty should be used with caution. Systematic uncertainties may average out over the day and this factor also is not included in the uncertainty estimates.

3.5. File structure region 5. Daily total information nighttime offset and min/max meteorological data

Irradiance measurements from certain sensors exhibit a nighttime value that can be associated with radiation from the sensor to the night sky referred to as the nighttime offset. The nighttime offset is subtracted from the measured value to partially account for the thermal radiation. (Vignola et al., 2009). This is discussed in greater detail in Section 9.

The nighttime offset is computed using the following method. The data from each instrument is investigated on a daily basis. Only good data points are used to compute the nighttime value. Astronomical night is defined as when the sun has a solar zenith angle greater than 108°. Only data points that have a SZA greater than 108° are used in the calculation of the nighttime offset. If there are not any good data points for a particular 24 h nighttime period, the average nighttime value from the entire month is used. If there still are not any good nighttime values, a reasonable nighttime offset is supplied from the past history of the instrument. Along with the average nighttime offset, the standard deviation (1 sigma) of the nighttime offset is calculated for each night. The average nighttime offset and standard deviation of the nighttime values are both given in W/m^2.

The minimum and maximum meteorological data are calculated for each 24 h period. This offers the user a brief snapshot of the conditions during the day. PV output data are summed to give the energy in kWh per day.

3.6. File structure region 6. Short time interval information Date/Time, SZA/AZM, ETR/ETRn

The short time interval data set contains the data gathered from the station. This time interval is the time interval that is output by the data logger. Older files had a time interval of one hour. Currently most of the monitoring stations have time intervals of one minute.

The short time interval portion of the data file is separated into three sub-regions. The left most region contains date and time information, solar position information, and extraterrestrial irradiance information.

- **Date and Time (Columns 1 – 3):** The date and time of each row are written in three different date/time formats to facilitate use by users with different requirements. The first column is the day of the year with a decimal point representing the fraction of a year using the formula.

$$\mathrm{year.fraction\ \ of\ \ year} = \mathrm{year} + \frac{(\mathrm{day\ \ of\ \ year.fraction\ \ of\ \ day} - 1)}{\mathrm{days\ \ in\ \ year}} \tag{6}$$

For example: 2017, January 1st at 6 AM would be 2017.00068493.

The days in the year include the leap day in the calculation when appropriate to correctly identify the fraction of the year the day represents.

The second column is the day of the year with the decimal point representing the fraction of a day using the formula.

$$\mathrm{day\ \ of\ \ year.fraction\ \ of\ \ day} = \mathrm{day\ \ of\ \ year} + \frac{(\mathrm{minute\ \ of\ \ day})}{1440} \tag{7}$$

The day of the year starts at one, not zero by convention. For example: 2017, January 1st at 6 AM would be 1.25,000. The year is not included in this column.

The third column is the traditional view of dates and times, in order from largest to smallest, year-month-day–hour:minute:second (YYYY-MM-DD–hh:mm:ss). The double dash marks "—" separate the date and the time. This is done to maintain the date and time format that are often altered when files are imported into spreadsheets.

As an example: 2017, January 1st at 6 AM would be 2017-01-01–06:00:00.

- **Solar zenith angle and solar azimuthal angle (Columns 4 – 5):** The solar zenith angle (SZA) and solar azimuthal angle (AZM) are calculated using the SOLPOS algorithm available from the NREL website (SOLPOS Website). The solar zenith angle is computed using refraction through the atmosphere. The calculation is done for the middle of the time interval. Unlike the SOLPOS code the SZA is also given when the sun is below the horizon.
- **Extraterrestrial irradiance and Extraterrestrial normal irradiance (Columns 6 – 7):** The Extraterrestrial irradiance (ETR) and Extraterrestrial normal irradiance (ETRn) are calculated using the SOLPOS algorithm https://www.nrel.gov/grid/solar-resource/solpos.html. The units of ETR and ETRn are in W/m^2. The ETRn is first calculated using Equation (8).

$$\begin{aligned} ETRn = 1361 * (1.000110+ \\ 0.034221 * Cos[DA] + 0.001280 * Sin[DA]+ \\ 0.000719 * Cos[2DA] + 0.000077 * Sin[2DA]) \end{aligned} \tag{8}$$

where DA is the day angle in degrees given by the Equation (9). Equation (8) was modified to include the most recent estimate of the average total solar irradiance (TSI) value outside the atmosphere of 1361 W/m^2 instead of 1367 W/m^2 that was used in the SOLPOS reference. See Vignola et al., 2020, Kopp and Lean, 2011; and Kopp, 2016 for further discussion of the TSI.

$$\mathrm{DA} = (\mathrm{day\ \ of\ \ year} - 1)*\frac{360}{\mathrm{days\ \ in\ \ year}} \tag{9}$$

In a leap year there are 366 days in the year, instead of 365.

The ETR is computed from the ETRn using Equation (10).

$$\mathrm{ETR} = \mathrm{ETRn}*\mathrm{Cos}(\mathrm{SZA}) \tag{10}$$

The ETR and ETRn are set to zero when the entire disk of the sun is below the horizon (SZA greater than 90.267°). The angular radius of the sun is 0.267°. When the sun is entirely below the horizon diffuse irradiance can still contribute to GHI and DHI so while ETR is zero, the GHI value can be positive. During the time intervals of sunrise and sunset, when the sun crosses the SZA = 90.267° boundary, the ETR and ETRn are decreased by a scale factor dependent on the fraction of time the sun is visible during the time interval.

3.7. File structure region 7. Short time interval information processed and calculated data

For some measurements, some irradiance data adjustments are made to help eliminate systematic effects. One of the most common effects of thermopile-based radiometers is caused by radiation to the sky (thermal offsets). Under clear sky conditions the thermal offset can be twice the night time thermal offset while under cloudy conditions, the thermal offset is about equal to the nighttime thermal offset (Vignola et al., 2009). The measured irradiance data and the nighttime offset from each instrument are used to adjust irradiance data using Equation (11).

$$IRR = IRR\_original - NO \tag{11}$$

where IRR is the adjusted irradiance, IRR_original is the measured irradiance signal, and NO is the average nighttime offset of the instrument from midnight to maritime sunrise and maritime sunset to midnight. The data label for the adjusted irradiance data removes the tag "_original" because the average nighttime offset has been subtracted from the value.

If the irradiance is determined from a rotating shadowband, further adjustments are applied. These adjustments remove some of the systematic effects associated with deviations from true cosine response, temperature dependence, and sensitivity to the spectral distribution of incident radiation as discussed by (Vignola, 2006).

Calculated columns can be determined using the processed data. The calculated GHI, DrHI, and DHI are obtained using Equations 12–14 respectively.

$$GHI\_calc = DNI*Cos(SZA) + DHI \tag{12}$$

$$DrHI\_calc = DNI*Cos(SZA) \tag{13}$$

$$DHI\_calc = GHI - DNI*Cos(SZA) \tag{14}$$

Each measurement in a data file has a quality control flag which guides the user as to the quality of each measurement. The data from each station is manually inspected for problems. If a problem is found the data are flagged appropriately. The flag column is listed to the right of the data column and the column header has the phrase "_Flag" appended to the irradiance label.

A table of the quality control flags is listed in Table 3. The UO SRML uses simple quality control flags given in Table 3. However, other quality control flags such as those produced by SERI QC program (User's Manual, 1993) can be used. The meteorological and measured irradiance data have flags that end in a one (11, 21, 31, 71, 81). Processed and calculated irradiance measurements have a flag that ends in a two (12, 22, 32, 72, 82). The word "processed" is used to mean data that have the systematic biases removed. The difference between the measured data and the processed data is that the processed data has adjustments made to it such as those given by equation (11). To be clear, both measured and processed data sets can be manually edited to correct for errors.

Users that only want to use the most accurate data that has not been edited should select data points with flags 11 (good meteorological data), 12 (good irradiance data), or 72 (good calculated irradiance data). As a disclaimer, there are occasional undetected problems that have been missed in the data analysis procedure. The user should perform their own quality control check during their analysis.

- **Best data** is processed or raw data with which no problems have been identified. The automatic detection of outlier data is developing and this is part of IEA Task 16. Currently the UO SRML network data are visually scanned to spot problems and assisted by information in log sheets. In addition the analysis process involves comparison of output from various instruments. While scanning by eye is more time consuming, patterns are often more easily spotted visually than use of computer programs that may not have the particular scenario included. For example, the melting of snow off the dome of a pyranometer has a characteristic pattern that is easy to spot and one can identify when the snow has melted off. A computer aided analysis is being developed and it is particularly useful in spotting difference in irradiance during partially cloudy periods where the irradiance is changing rapidly.

Table 3
Short time interval quality control flags.

Quality of measurement	Meteorological data (Air Temperature, Pressure, etc.)	Irradiance data (Measured)	Irradiance data (Processed)	Irradiance data (Calculated)
Best data	11	11	12	72
Substituted data	21	21	22	22
Interpolated data from this instrument	31	31	32	32
Questionable data	81	81	82	82
Obstructions (GHI and DNI)	NA	91	92	92
Bad data	99	99	99	99

- **Substituted Data:** The goal of the SRML is to provide the highest quality data set with an attempt to have as complete a set of data possible. With this in mind, problems in the data that are identified and can be fixed in a rational manner are changed. For example, if there is more than one instrument making the same type of measurement and a problem is identified in one instrument, the data from the other instrument can be substituted for the problem data. The substituted data comes from the processed data column, such that both the destination and substituted data have thermal offsets already taken into account. Values in the original, unprocessed, column are also substituted. The substituted data is modified according to Equation (15), where the nighttime offset of the bad data column is unapplied to the substituted data.

$$IRR\_original = substituted\_data + NO_{bad\_data\_column} \tag{15}$$

- **Interpolated data from this instrument:** If a data set has only a short break, between good data points, the bad data may be replaced by interpolated data from this instrument using a linear fit. Typically, this is just several minutes, but extended fits of up to an hour can occur under some sky conditions. Extended fits require knowledge of very stable sky conditions, typically through the use of another sensor. For example, if a DNI data set shows a very clear period, and the GHI data set has irregularities, the GHI data set can be adjusted using a linear fit. Fitting across routine instrument maintenance and cleaning is common, when the cleaning period can be easily identified. For fifteen minute or longer data intervals, identifying cleaning dips is very difficult to discern and the data can be marked as questionable. With pyrheliometers on manually adjusted trackers that have been somewhat out of alignment, it is possible to estimate the percent decrease in DNI caused by the misalignment when the alignment is corrected in the morning. This type of adjustment is also flagged as interpolated data. Data is typically interpolated when we are confident that the interpolated data can be brought within 5% the expected value. Typically, this can be done during clear or completely overcast periods.
- **Questionable data:** When analyzing data and one is uncertain about the accuracy of the data and inserting interpolated or auxiliary data for the problem data cannot be done with confidence, the data are flagged questionable. For example, if the pyrheliometer is out of alignment on one day, the instrument may or may not be out of alignment on the previous day. If there is uncertainty, then the data are flagged questionable. Questionable data have not been altered except through the automated process procedures to remove nighttime offsets. Users that want a complete data set, should consider using the questionable data points. Users that want a clean data set should consider not using the questionable data points.
- **Obstruction:** A flag has been implemented in the comprehensive format that designates data points where the sun is behind a permanent obstruction. Permanent obstructions include: trees, telephone poles, telephone wires, buildings, mountains etc. Permanent obstructions do not include: insects, people, cleaning events, clouds, temporary problems etc. Permanent obstructions are identified by comparing the dips in the data during clear periods to sun path charts. A permanent obstruction will show a consistent dip for several days indicating that the sun is blocked at the same time each day. The motion of the sun from day to day is taken into account for this process. Only GHI, DNI, DrHI data sets are given the permanent obstructions flag. The data for a permanent obstruction is not altered, simply flagged as such.
- **Bad:** Bad data points are given the flag 99 and may have a data value of NA. Bad data should not be used.
- **Calculated:** Calculated data points are given the flag 72 to distinguish them from measured or adjusted data. If the data used to generate the calculated value is flagged questionable or adjusted, then the flag for the calculated values uses the flag from the most problematic data. This is typically the data with the largest estimated uncertainty.

3.8. File structure region 8. Short time interval information measured irradiance and other meteorological data

Unprocessed or raw irradiance and other meteorological data from the monitoring station are displayed in the right most columns of the data file. The raw irradiance columns are denoted with the notation "_original", meaning that any the nighttime offset has not been subtracted. Other adjustments have not been applied, including adjustments for non-Lambertian cosine response or adjustments for dependence on spectral distribution of the incident irradiance. These data are measurements from the data logger. Meteorological data are included in this portion of the data file as well. Meteorological measurements do not have the "_original" label but can be considered as such.

3.9. File structure region 9. Comments

The farthest right column of the entire data set is devoted to comments. Comments about the data file are given in the header rows. Comments about individual data points are given in the data set at the appropriate place, for example when an instrument was changed. The ability to make comments concerning the data set is extremely useful because it offers the user the ability to know any special influences or factors that exist for any particular data point(s). Some of the more common comments are snow or ice on the instrument, misalignment, or the instrument was being cleaned. These comments can be taken directly from log sheets and inserted into the data set when the files are certified.

4. Summary

With the plethora of data becoming available today, it is essential to give users a basis for comparing results and relating data. Too often one is presented with a time series of numbers without crucial information that provides understanding of the biases and uncertainties in the data. While this information may be available in documents describing the dataset, it is often hard to piece together.

The comprehensive format used by the University of Oregon Solar Radiation Monitoring Laboratory has been explained in detail to serve as a template for others who want to present data in files that can be easily accessed by spreadsheet based or other data processing programs. This facilitates the evaluation of the relationships between components and developing and/or testing models. More importantly, the format provides space for information vital for the assessment of the quality and uncertainty of the data in the file. No dataset is perfect and providing the estimated uncertainties enables users to judge on the reliability of the results obtained from the use of the data. This gives the user a level of confidence in the results and if handled properly can make the dataset "bankable". Calibrations, uncertainties, and adjustments to the dataset are essential for users to judge the dataset and results. They need to be included in any dataset so that the dataset becomes more than a set of numbers without substance behind them. The structure of these data files is flexible enough so that those generating the dataset can insert their own processes as long as the processes are acknowledged to the user.

The comprehensive data format provides users with more detailed information on how the data are obtained and processed. Inexperienced users will benefit from the increased description of the various components of the data set as well as the daily summary at the top of each data file. Experienced users will benefit from the more detailed information on the instrumentation and the uncertainties associated with the data calibration values. This data set is intended to be easier to use with various formats for date and time, solar zenith and azimuthal angles. The SRML has begun reformatting the existing data set into the more comprehensive format. The current data in the comprehensive format are being made available on the website after the end of the month. The original formatted files are being updated to the new comprehensive format as time allows. These files that are converted to the comprehensive files are put on the website as the transformation has been completed.

Declaration of Competing Interest

I (Josh Peterson) hold two positions in the solar energy field. The first is for the University of Oregon, which is where I am publishing this paper from. The second is an hourly employee of Groundwork Renewables (GR). GR is a metrology company. My work at GR involves soiling analysis, quality control processing, and code writing. This publication has very little relevance to the work that I do at Groundwork Renewables. I will not profit at GR through this publication.

Acknowledgements

The UO Solar Radiation Monitoring Laboratory would like to thank the Bonneville Power Administration, Energy Trust of Oregon, the U.S. Department of Energy's National Renewable Energy Laboratory under Contract No. DE-AC36-08GO28308 with Alliance for Sustainable Energy, LLC, the Manager and Operator of the National Renewable Energy Laboratory. Funding provided by U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Solar Energy Technologies Office., and Portland General Electric for their support that makes this work possible.

Appendix A. Overview of the University of Oregon solar radiation monitoring laboratory solar monitoring network

The University of Oregon Solar Radiation Monitoring Laboratory (UO SRML) started collecting high quality GHI data at Eugene, Oregon in 1975. In 1977, in partnership with the Atmospheric Sciences department at Oregon State University, a grant was received to start training students in solar radiation measurements and The UO SRML added four additional stations to a Pacific Northwest Solar Radiation Network. The original equipment for the new stations consisted of Schenk pyranometers and strip chart recorders. As funding for equipment became available, an Eppley NIP and manual tracker was purchased for Eugene in December 1977. With additional funding from the grant and the local utility, stations were started in several other locations in the region with first class Eppley pyranometers and pyrheliometers. A digital data logger was developed by the science shop at the university and integrated data was recorded at 5-minute intervals on cassette tape. Strip charts and digital printers backed up the data in case problems developed when recording or reading the tapes. Data output was reported in hourly intervals. For more information about the early years at the UO SRML, see http://solardat.uoregon.edu/PacNWSolarRadiationDataBook.html.

The file format that was developed was an ASCII tab-separated integers file format. Only numeric characters were allowed. The column headers were identified using the research cooperator format developed at SERI (SERI, 1988). The format was functional but it is non-intuitive for new users, which is why the new comprehensive format was developed.

For many years the instruments were calibrated at the National Renewable Energy Laboratory (NREL), called the Solar Energy Research Institute at the time. Calibrations of the pyrheliometers were consistent with factory calibrations, but the pyranometers calibrated at NREL, at Eugene using a shade/unshade methodology, and at Eppley labs did not agree. It turns out that the thermal offset of the Eppley PSP pyranometers at NREL was significantly higher than at Eugene, causing a one to two percent difference in calibrations. In addition, the UO SRML originally calibrated at solar noon and the NREL calibration was at a solar zenith angle of 45°. Over the years, research in the field identified and standardized the calibrations. In 2006, the UO SRML went through, using clear sky data and calibration data to standardize the calibration record (Riihimatki et al., 2006). Having long-term calibration records of the various instruments used in the field allows for better determination of the calibration changes over time and to provide calibration values that are consistent with a long-term trend in the calibration values. Using the trend in calibration values, the estimated calibration value is estimated for the middle of the coming year and applied to all data gathered for that year. That way, if the calibration value changed by 1% over a year, then the calibration values would be within ±0.5% of the value chosen. The change in calibration value is mainly dependent on the exposer of the instrument to ultraviolet radiation varies from instrument to instrument.

The process of translating to the new comprehensive format allows for the use of a standardized calibration methodology. In addition, the data currently only available in hourly intervals will be made available at 5-minute intervals if the information is available. The data in the comprehensive format is made available on the UO SRML website as it is reanalyzed and reformatted.

Currently the SRML collects data from 15 sites throughout Washington and Oregon. The data in the comprehensive format from these sites are made publicly available approximately one month after it is collected. The high-quality sites include: Eugene Oregon, Hermiston Oregon, Burns Oregon, and Seattle Washington that have an automatic tracker and measure all three irradiance components (GHI, DHI, DNI). Most stations in the network upload one-minute data to the webserver every minute in the original format. In the near future, we intend on uploading comprehensive format files to webserver on a minute basis.

The Eugene station acts as a reference research station and contains measurements made with a variety of instruments used in other stations in the network. This allows for comparisons with different instruments and is useful for identifying uncertainties and biases in the measurements.

As interest in shorter time interval data increases, 1-minute data are now being gathered at most stations. This changeover started around 2018, but was initiated earlier for several stations. Along with the irradiance data, several of the stations also include measurements from photovoltaic systems. The number and location of these stations have varied over the years although some stations like Ashland, Oregon have had measurements since 2000. Data from the UO SRML network are available at http://solardat.uoregon.edu/SelectArchivalUpdatedFormat.html or http://solardat.uoregon.edu/SelectArchival.html. For information on the structure of the website, click on the site button on the tool bar. This will go to http://solardat.uoregon.edu/SiteMap.html.

In 2019, the website was accessed by approximately 140,000 distinct users from 160 countries and transfers about 500 gigabytes of data (normally around 100 gigabytes) and supplied 96,000 archived data files.

References

BIPM, IEC, IFCC, ISO, IUPAC, IUPAP and OIML, 1995. Evaluation of Measurement Data - Guide to the Expression of Uncertainty in Measurement. Geneva: ISO TAG 4.

Kopp, G., Lean, J.L., 2011. A new, lower value of total solar irradiance: Evidence and climate significance. Geophys. Res. Lett., 38, L01706, doi: 10.1029/2010GL045777.

Kopp, G., 2016. Magnitudes and timescales of total solar irradiance variability. J. Space Weather Space Clim. 6 (27), A30. https://doi.org/10.1051/swsc/2016025.

Riihimatki, L., Lohmann, L.S., Meyers, R., Perez, R., Vignola, F., 2006. Long-Term Variability of Global and Beam Irradiance in the Pacific Northwest Proc. of the 35th ASES Annual Conference, Denver, CO.

SERI Standard Broadband Format A Solar and Meteorological Data Archival Format, 1988. SERI/SP-320-3305 DE88001145 (https://www.nrel.gov/grid/solar-resource/assets/data/3305.pdf).

Vignola, F., 2006. Removing Systematic Errors from Rotating Shadowband Pyranometer Data, Proceedings of the 35th ASES Annual Conference, Denver, CO.

Vignola, F., Long, C.N., Reda, I., 2009. Testing a model of IR radiative losses. Proceedings of the SPIE Conference, San Diego, CA.

Vignola, F., Gover, C., Lemon, N., McMahan, A., 2012. Building a bankable solar radiation data. Sol. Energy 86 (8), 2218–2229.

Vignola, F., Michalsky, J., Stoffel, T., 2020. Solar and Infrared Radiation Measurements. Francis Taylor.

Wilcox, S., Andreas, A., Reda, I., Myers, D., 2002. Improved Methods for Broadband Outdoor Radiometer Calibration (BORCAL). Proceedings of the ARM Science Team Meeting, St. Petersburg, Florida, April 2002.

NWS Website: https://www.ncdc.noaa.gov/homr/reports/platforms (May 2020).

SOLPOS Website: https://www.nrel.gov/grid/solar-resource/solpos.html (May 2020).

User's Manual for SERI QC Software Assessing the Quality of Solar Radiation Data 1993. NREL TP-463-5608 DE93018210. https://urldefense.com/v3/__https://journals.ametsoc.org/bams/article/79/10/2115/56250/Baseline-Surface-Radiation-Network-BSRN-WCRP-New__;!!C5qS4YX3!WQRwpe9JZQ5B7Zt_gy9HnGS39MeqiaQgNJcfyycapK8WaLLMYZIENQZlU6z2CzeSuEY$ (July 2020). http://solardat.uoregon.edu/SelectArchivalUpdatedFormat.html (August 2020) http://solardat.uoregon.edu/SelectArchival.html. (August 2020).