Structure of a comprehensive solar radiation dataset

Josh Peterson and Frank Vignola

University of Oregon, Eugene (United States)

Abstract

A new, comprehensive, file format has been developed for solar and other meteorological data from the University of Oregon's Solar Radiation Monitoring network in the Pacific Northwest. The new 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 that would help users see what was done to the data during the analysis period. The goal is to make these more comprehensive data files available on the UO SRML website (http://solardata.uoregon.edu).

Keywords: *solar radiation, resource assessment, uncertainty*

1. Introduction

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 developed to be compatible with the Research Cooperator Format originally designed for the CONFRRM network for data sharing. 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. Today, such a file structure is archaic and difficult to use. In addition, solar data is now used by developers and financial institutions to evaluate operation and fiscal performance of solar facilities. More detailed information is required to provide confidence in the analysis that result from using the database. One could say that this information is necessary to make the dataset "bankable".

In an effort to facilitate the use of the data, a new file format was developed that contains significantly more information about the station and the various measurements. This information is 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 use to translate voltage measurements into irradiance values and the uncertainty in the calibration values. Additional information such as the solar zenith and azimuthal angles are provided along with basic information about the monitoring station.

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 Figure 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 metrological information such as temperature and air pressure.
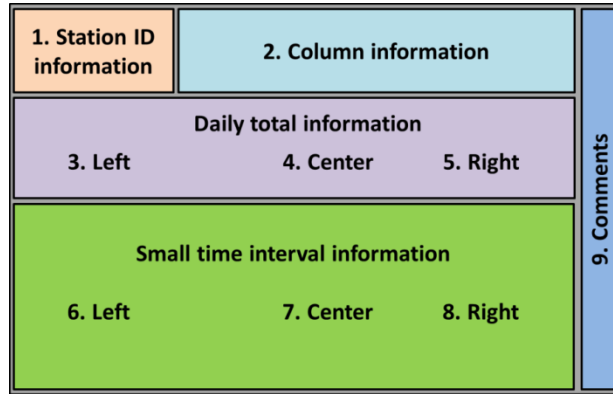

Fig. 1: Schematic diagram of the new file structure. The different regions of the file are labeled 1 - 9.

Note that regions 6, 7, 8 are the actual data and are generated first when the files are created. Regions 3, 4, 5 are metadata generated at the end of completed day. In this paper, the topics are discussed in the order they appear in Figure 1 and not in the order in which they are generated.

3. File structure region 1.
Station ID information

The upper left corner of each file contains useful information about the file. An example for the Burns station is shown below in Table 1.

Tab. 1: A sample data set of the data contained in Region 1 of the file structure.

Labels	Station values
Station ID Number:	94170
Station Name:	BUO
Station Location:	Burns_Oregon_USA
Latitude:	43.51920
Longitude (+ East):	-119.02162
Altitude (m):	1260
Time Zone (+ East):	-8
Time Interval (Minutes):	1
Year//Month	2016//12

- The station ID number was originally a WBan number [NOAA 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 previous file structure

- 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, Y = Wyoming, M = Montana)

- Station Location is the City, State, and country name of the station. The three names are separated by an underscore "_".

- Latitude, Longitude, and altitude of the station. The latitude and longitude are given to an accuracy of the ±200 meters. The longitude of the station is given as a negative number because the stations are all west of the prime meridian. 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. The time zone is a negative number.

- 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, although some instruments, such as a rotating shadowband pyranometer (RSP) are sampled only once per minute. Early measurements had a time interval of 60 minutes. Many stations today have a time interval of 1 minute.

- 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 formatting dates and times into their predetermine 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.

4. 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.

Tab. 2: Sample of the data contained in the column headers region 2

Row Labels	Instrument 1	Instrument 1 Flag	Instrument 2	Instrument 2 Flag	Instrument 3	Instrument 3 Flag
Type of Measurement:	GHI	GHI_Flag	GHI_Calc	GHI_Calc Flag	GHI_withNO	GHI_withNO Flag
Element:	1000	-	1001	-	1000_withNO	-
Instrument Serial Number:	PSP (23973F3)	-	Computed from DHI and DNI	-	PSP (23973F3)	-
Instrument Shorthand Name:	P23	-	NA	-	P23	-
Responsivity:	8.6844	$\frac{microV}{W/m^2}$	NA	-	8.6844	$\frac{microV}{W/m^2}$
Estimated Uncertainty (U95%):	3.587	-	6.003	-	3.587	-
Sample Method:	Avg	-	-	-	Avg	-
Units:	W/m^2	-	W/m^2	-	W/m^2	-
Column Notes:	AdjustedColumn	-	CalculatedColumn	-	Measuredcolumn	-

- **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 DfHI Diffuse Horizontal Irradiance. The notation for other options will be discussed in latter sections. The columns labeled as _Flag will be discussed in latter sections as well.

- **Element:** A numeric value of each type of measurement. When the data from the SRML was first being published there were logistical challenges that required to use the rather cryptic notation. For a complete description of the various element numbers see: http://solardat.uoregon.edu/DataElementNumbers.html. Some of the common element numbers are shown above. To enable users of the previous format to use the new format, the element numbers used in the previous file format are carried over to the new format.

- **Instrument Serial Number:** To provide traceability, the serial number of each instrument is recorded. Some columns contain calculated data and will be identified as such instead of using a serial number.

- **Shorthand name:** This is an internal notation used by the SRML to reference the various sensors and is used for internal discussions. The name is assigned to each instrument upon purchase. The shorthand name is not related to the serial number.

- **Responsivity:** The responsivity is the calibration constant that was used to convert the raw measured voltage (or millivolts) to irradiance values. The formula relating voltage to irradiance is given by Equation 1.

$$Irradiance = \frac{Voltage}{Responsivity}$$ (eq. 1)

The voltage of each measurement is not recorded, only the corresponding irradiance and responsivity are recorded. The calibration records for each instrument will be posted on the UO SRML website and these files are usually updated annually.

- **Estimated uncertainty:** The estimated uncertainty (at the 95% level of confidence) in the measured value reported in that column. Responsivity values are computed at an angle of incidence of 45°. A discussion on instrument calibration is similar to the methodology used for the pyranometer calibration made using the Broadband Outdoor Radiometer Calibration methods (BORCAL) prior to the year 2015 as discussed by Wilcox et al. 2002. The SRML characterizes each instrument at various angles of incidence and hopes to make this information available on its website in the future.

- **Sample Method:** The method that was used to gather data. 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. Information about the sample rate is not given in the file. Certain sensors (such as an RSP) 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. Typical units for Temperature are Celsius (C). Note the carrot symbol ^ is used to describe a number raised to a power.

- **Rows 9 – 10:** These two rows allow for notes about each column. Notes may include information about RSP sensors, 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 as they see fit.

The previous discussion examined the different rows of a data set. The discussion was generic and only the most common types of measurements were used as examples. In the upcoming paragraphs, the different columns that may be encountered will be discussed. There are four different types of data: processed data, calculated data, measured data, and metrological data. The method used to process and adjust the data will be discussed in Section 9. The calculation method of the metrological and measured data will be discussed in Section 10. Each measurement is represented with a pair of columns. The first column contains the measurement value and the second column corresponds a quality control flag. Process flags will be discussed in at the end of Section 10.

- **Processed Data:** This corresponds to irradiance data that has been evaluated and possibly adjusted. For example, adjustments for radiative losses are determined by evaluating the nighttime offsets that are subtracted during the day and adjustments to RSP measurements to account for systematic deviations. The adjustment algorithm will be discussed in Section 9. Several examples for commonly encountered adjusted data labels include GHI, DNI, DfHI, GHI_Tilt, GHI_Auxiliary These data labels are found in the first row of each column.

- **Computed Data:** This is data that has 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 DfHI using the formula

  GHI_Calc = DNI Cos[SZA] + DfHI   (eq. 2)

Where SZA corresponds to the solar zenith angle. GHI_Calc is the calculated global horizontal irradiance, DNI is the processed direct normal irradiance, and DfHI 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]   (eq.3)

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

- **Measured data:** Each irradiance measurement presented in its unprocessed form. The raw data has not been processed or otherwise adjusted. Raw data is often differentiated from processed data using the notation "_withNO" following the type of measurement name, for example GHI_withNO. 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.

- **Metrological data:** Meteorological measurements such as air temperature and atmospheric pressure. Meteorological data is not adjusted.

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

The daily total information contained in Rows 12 – 42 contain summary information about each instrument for each day of 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 region include the day of month, day of year, sunrise and sunset times, solar noon times, daily total extraterrestrial energy on a horizontal surface (ETR), and daily total extraterrestrial energy on a normal surface (ETRn).

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

$$\text{Energy} = \frac{\text{time interval}}{60*1000}\sum_i \text{IRR}_i \qquad\qquad\qquad\qquad (\text{eq.4})$$

where 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 Energy incident on a normal surface (ETRn) is computed using a similar formula and is given in Column 7. Time intervals which span the sunrise and sunset are scaled accordingly.

6. File structure region 4. Daily total information
Daily total energy

The total energy for the processed and computed data sets are computed for each day. Equation 4 is used to compute the total energy on a surface. In Equation 4, IRR$_i$ is the individual irradiance values reported throughout the day for that surface and instrument.

Missing and bad data points are interpolated using a linear fit. If there is more than one hour of data is missing or flagged bad, then the daily total for that day is not computed and that cell is left blank.

An uncertainty estimate of the total energy 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 21 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.

7. File structure region 5. Daily total information
Nighttime offset and min/max metrological 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. (Reference). 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 hour nighttime period, the average nighttime value from the entire month is calculated. 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 metrological data is calculated for each 24 hour period. This offers the user a brief snapshot of the conditions during this time period.

8. 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 many of the monitoring stations have time intervals of one minute.

The short time interval portion of the data file is separated into three 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. The first column is the day of the year with a decimal point representing the fraction of a year using the formula.

$$\text{year.fractionofyear} = \text{year} + \frac{(\text{dayofyear.fractionofday} - 1)}{\text{days in year}} \qquad \text{(eq.5)}$$

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

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

$$\text{dayofyear.fractionofday} = \text{dayofyear} + \frac{(\text{minuteofday} - 1)}{1440} \qquad \text{(eq. 6)}$$

For example: 2017, January 1st at 6 AM would be 1.25000. 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). Note 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. The solar zenith angle is computed using refraction through the atmosphere. The calculation is done for the middle of 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. The units of ETR and ETRn are in W/m^2. The ETRn is first calculated using Equation 7.

$$\text{ETRn} = 1367 * (1.000110 + 0.034221 * \text{Cos[DA]} + 0.001280 * \text{Sin[DA]} + \\ 0.000719 * \text{Cos[2 DA]} + 0.000077 * \text{Sin[2 DA]}) \qquad \text{(eq. 7)}$$

where DA is the day angle in degrees given by the formula.

$$\text{DA} = (\text{day of year} - 1) * \frac{360}{\text{days in year}} \qquad \text{(eq. 8)}$$

The ETR is computed from the ETRn using Equation 9.

$$\text{ETR} = \text{ETRn} * \text{Cos(SZA)} \qquad \text{(eq. 9)}$$

In Equations 7 and 9, the solar constant is defined as 1367 W/m^2 instead of the current value of 1361 W/m^2. Efforts are underway to correct this error. The ETR and ETRn are set to zero when the entire disk of the sun is below the horizon (SZA > 90.267°). The angular radius of the sun is 0.267°. 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.

9. 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 night time thermal offset (Vignola, 2009). The measured irradiance data and the nighttime offset from each instrument are used to compute the adjust irradiance data using Equation 10.

$$\text{IRR} = \text{IRR\_withNO} - \text{NO} \qquad \text{(eq. 10)}$$

where IRR is the adjusted irradiance, IRR_withNO is the measured irradiance signal, and NO is the nighttime offset of the instrument for each 24 hour period. The adjusted irradiance data label does not include the tag "_withNO".

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 DfHI are obtained using Equations 11 - 13 respectively.

$$\text{GHI\_calc} = \text{DNI} * \text{Cos(SZA)} + \text{DfHI} \qquad \text{(eq. 11)}$$

$$\text{DrHI\_calc} = \text{DNI} * \text{Cos(SZA)} \qquad \text{(eq. 12)}$$

$$\text{DfHI\_calc} = \text{GHI} - \text{DNI*Cos(SZA)} \qquad \text{(eq. 13)}$$

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 is

flagged appropriately. The flag column is listed to the right of the data column and is denoted with the label "_Flag".

A table of the quality control flags is listed below (Table 3). The metrological and measured irradiance data have flags that end in a one. Processed and calculated irradiance measurements have a flag that ends in a two. Users that only want to use the most accurate data that has not been edited should select data points with flags 11, 12, or 72. 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.

Tab. 3: Short time interval quality control flags

Quality of measurement	Other metrological data (Temp, 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
Bad data	99	99	99	99

- **Best data** is processed or raw data with which no problems have been identified.

- **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 are changed. For example, sometimes there are more than one instrument making the same type of measurement. If one instrument has problem the other instruments data can be substituted for the problem data. The nighttime offsets and other adjustments are accounted for in this process to produce a seamless data set.

- **Interpolated data from this instrument:** If a data set has only a short break, between good data points (less than 1 hour), the bad data is replaced by interpolated data from this instrument using a linear fit. This commonly occurs during routine instrument maintenance.

- **Questionable data:** When analyzing data and one is uncertain about the accuracy of the data, then 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 is flagged questionable. Questionable data has not been altered except through the automated process procedures.

- **Bad:** Bad data points are given the flag 99 and may have a data value of NA. Bad data should not be used.

- **Calculated:** Good calculated data points are given the flag 72 to distinguish them from measured or adjusted data. The calculated flag mimics the flags of the parents if there are non-good flags.

10. 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 "_withNO", meaning the nighttime offset has not been subtracted. Other adjustments have not been applied, including adjustments for non-lambersian cosine response or adjustments for dependence on spectral distribution of the incident irradiance. These data are measurements from the data logger. Metrological data are included in this region of the data file as well.

11. 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 comment the data set is extremely important because it offers the user the ability to know what influences or factors exist at any moment.

12. Conclusion

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. The data set is intended to be easier to use with various formats of the date and time, solar zenith and azimuthal angles. The SRML has begun reformatting the existing data set to the new format. The new format will be available to the public after a significant portion of the entire data set has been reformatted.

13. Acknowledgements

14. 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.

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

Vignola, F., Charles N. Long, Ibrahim Reda. (2009). Testing a model of IR radiative losses. Proceedings of the SPIE Conference, San Diego, CA.

Wilcox, S., Andreas, A., Reda, I., and 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 (October 2017)

SOLPOS Website: http://rredc.nrel.gov/solar/codesandalgorithms/solpos/solpos.c (October 2017)