A New National Database for Solar Resource Assessment Compiled from Ground-Based Observations

James W Hall, PhD
JHtech/SolarDataWarehouse, San Antonio, FL, USA

Abstract

A new solar radiation resource with complete coverage of the United States has been created from ground-based observations of global irradiance from 2005 to the present. Daily and hourly data have been collected from over 7000 professionally maintained weather stations. These sites belong to nearly 100 different federal, state and university networks. The average distance between sites in the continental US is about 37 km - somewhat farther in the great plains and closer in more populated areas. They have been combined into a "network of networks" so the accuracy of each individual site can be judged by comparing to nearby neighbors, often from other networks. Advanced filtering and sensor fusion techniques common in artificial intelligence are used to identify sensors that are not performing correctly and filter errors from the data stream. Validation tests have shown that the daily and hourly measurements of global irradiance in this ground-based resource have roughly half the uncertainty, half the observation error and half the bias error compared to the satellite based observvations in the National Solar Radiation Database (NSRDB-PSM). Accuracy, bias and uncertainty at the monthly and annual levels are comparable. This new solar resource can be combined with existing data, or used as a stand-alone database, to enhance site planning, production forecasts and on-going monitoring of solar projects.

Keywords: solar radiation resource, solar radiation database, ground-based solar measurements, National Solar Radiation Database, NSRDB

1. Introduction

The solar power sector relies on accurate assessment of resource throughout the life cycle of a solar power plant. This reliance begins with site selection, continues through design and acceptance testing, and persists as efficient operation is validated to the end of the plant's service life. Historically, the industry has relied on satellite-based solar resource since it is readily available, has global coverage and long-term measurements. However, due to the known limitations of satellite data, most significant projects also require additional ground-based, site-specific measurements of solar resource to tune the satellite data and lower the uncertainty in energy production estimates. This merging of satellite data with ground observations is often referred to as creating a bankable solar dataset.

Solar irradiance measurements from thousands of ground-based sites are publicly available. However, solar irradiance is among the most difficult of meteorological measurements, especially in a production setting. Some have questioned the direct use of these public sites for solar resource due to concerns about sensor maintenance, sparse distribution and the lack of validation methodologies.

SolarDataWarehouse has recently completed a new bankable, long-term solar resource based on ground measurements with complete coverage of the United States – the US Ground-Based Solar Irradiance Database (GSID). This paper explains how artificial intelligence (AI) filters and sensor fusion were applied to lower the uncertainty and quality control the observations from professional ground sites. Detailed validation of this new resource followed the same procedure used to validate National Solar Radiation Database Physical Solar Model (NSRDB-PSM version 3). In this validation, the ground-based observations in the GSID had roughly half the bias error, half the observation error, and half the uncertainty at hourly and daily levels compared to the satellite-based data in the NSRDB-PSM for the years 2005-2018. The GSID is therefore a ground-based

Selection and/or peer review under responsibility of Scientific Committee
doi:10.18086/solar.2020.01.07 Available at http://proceedings.ises.org

solar resource that can be used as a stand-alone dataset, or in conjunction with other resource data, for the planning and monitoring of solar power projects.

2. A Network of Networks

There are many different professional networks in the US that measure solar radiation. These networks have been established by the federal government, state governments and universities for a specific purpose such as climate research, resource management and agriculture. Each entity establishes their own criteria for site selection, maintenance and data quality control. Examples of these networks include:

- US Climate Reference (USCRN) – A network of over 130 sites across the US providing high-quality, long-term data for climate research.
- The University of Oregon's Solar Radiation Monitoring Laboratory (SRML) – a network of 37 sites designed to provide high-quality solar resource data for the Pacific Northwest.
- California Irrigation Management Information System (CIMIS) – a network of over 200 stations across California to help manage water resources more efficiently.

Taken separately, each network fulfils specific, local purposes for which it is operated. However, no single network provides nationwide solar coverage, and generally the sites are located too far apart to cross-validate observations with other sites in the network. The Ground-Based Solar Irradiance Database (GSID) is a compilation of data from nearly 100 different overlapping networks, professionally operated and maintained with over 7000 individual sites (Figure 1). Merging these networks into one provides complete spatial coverage across the US. The average distance between sites is 37 km, with somewhat greater spacing in the Great Plains and tighter spacing in more populated areas.

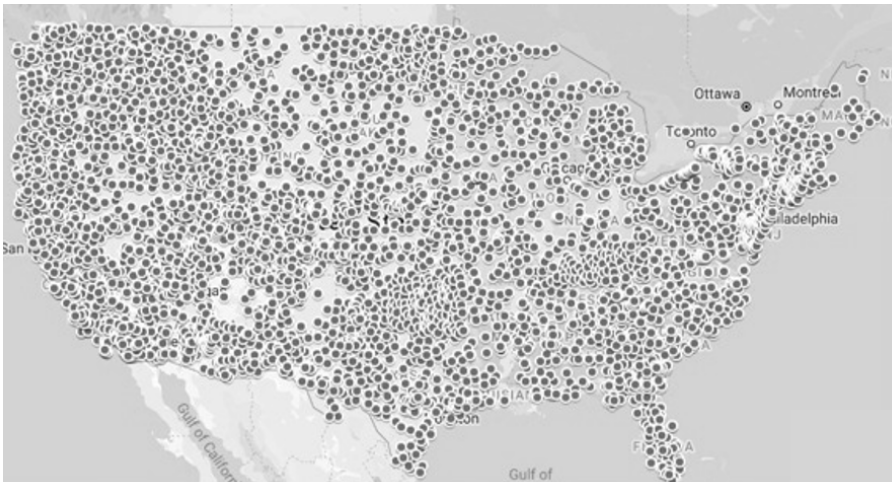

Fig. 1: The US-GSID is a network of networks, combining solar observations from over 7000 sites and nearly 100 networks

The many individual sites contained in the GSID database provide good coverage in key solar areas such as California, with 1350 sites at an average spacing of 18 km. The coastal and mountainous regions of California often contain microclimates that have proved challenging for satellite models. Researchers have demonstrated that significant improvement in solar resource for California can be obtained by utilizing ground-based data rather than relying solely on satellite observations (Anders Nottrott and Jan Kleissl, 2010, 2014).

Ingesting and merging the data from nearly 100 different networks into a single database presents challenges. Networks are constantly adding, decommissioning, or on occasion re-locating sites. Each network is built and operated for specific professional tasks - public data access is never the primary purpose and is generally limited to one site at a time. The networks measure different parameters, have different sampling rates, time

may be expressed differently and measurements expressed in different units. Web access and website formats change regularly, so constant modifications to the ingest software are required.

3. Reducing Error Via Artificial Intelligence and Sensor Fusion

In addition to complete spatial coverage, this dense network of professional ground sites provides an extensive amount of solar resource data. As with any large set of measurements, there are always errors. Rather than rejecting the data as a viable resource because of these errors, one can learn to identify specific data patterns associated with the errors and use modern processing techniques to filter them. Artificial Intelligence (AI) is particularly well suited to this task.

A common technique in artificial intelligence is sensor fusion: combining data from multiple sensors to compensate or correct for the deficiencies of individual sensors. It relies on multiple sensor sites or multiple types of sensors viewing the same event. Such an approach is data driven rather than model driven. Intelligent algorithms process the data and fuse it into a data stream with lower uncertainty than the original sensors might have. The algorithms also give an indication of sensors that are not preforming correctly so they can be removed from the data stream until repaired.

For a basic example: the calibration of a solar irradiance sensor is conventionally done via side-by-side comparison to a high-quality reference. This works well for an initial calibration, but is not feasible for on-going performance monitoring. A data-driven method to monitor sensor performance is to compare a sensor's output to that of neighboring sensors viewing the same event. Clear sky days are an instance when all neighboring sensors should be seeing the same event.

There are several physics-based models for estimating clear sky radiation from solar position and atmospheric conditions, however, accurate knowledge of atmospheric conditions may not be available at the desired location and date. One data-driven alternative is plotting a large number of ground-based solar observations for a specific location over a period of time. To illustrate, the daily GHI observations from four ground sites near Kennewick, WA are in Figure 2. From such plots, one can estimate the average clear sky global irradiance at that location for any day of the year. Similarly, time histories of hourly data can be used to estimate clear sky global irradiance for any hour at a specific location and date. This type of analysis can also be used to quantify a bias error in any of individual sensors.

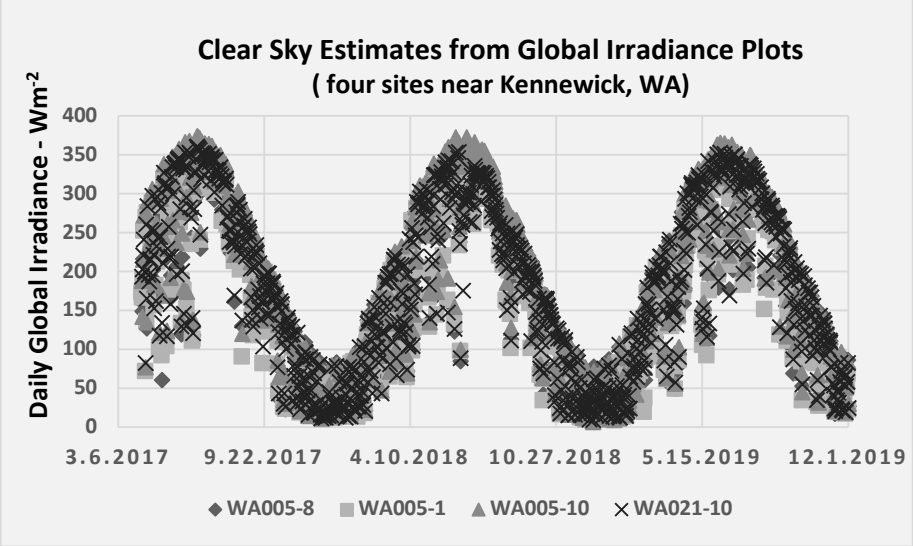

Fig. 2: Estimating clear sky global irradiance from observations near Kennewick, WA

Once the historical clear sky irradiance is known, clear sky events can be identified in real-time. On clear sky days, a deviation by one sensor from the mean of the neighboring sites can indicate a malfunction, and the amount of uncertainty for the malfunctioning sensor can be quantified using traditional statistical methods.

Figure 3 shows a simple example of a malfunctioning sensor near Sacramento, CA on October 2, 2013. Daily GHI observations of 17 ground sites from five different solar networks are shown. The circle indicates a 40 km radius. The site in the center of the circle is in disagreement with the surrounding sites, and is removed from the data stream until future observations indicate it has been repaired.

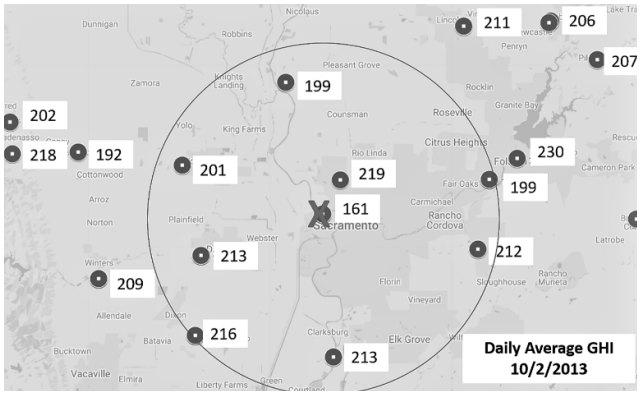

Fig. 3: A malfunctioning sensor near Sacramento was identified by comparing nearby sites on a clear sky day.

Analysis of hourly irradiance on clear sky days can identify improper siting or sensor misalignment (Figure 4). A repeated pattern of hourly observations below the historical envelope occurring at the same time of day can indicate shading of the sensor. A skewed curve, with recurring deviations below the historical envelope in the morning and above the envelope in the afternoon is a pattern typical of sensor misalignment. Artificial Intelligence (AI) algorithms are well suited to finding these patterns in large datasets.

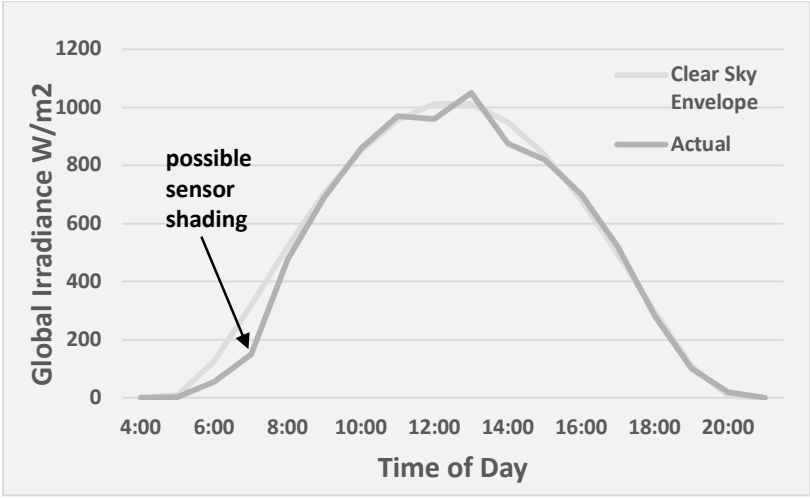

Fig. 4: Identifying siting or sensor installation problems from hourly data

Sensor fusion can perform complex quality control of the sensor data. For example, an algorithm for sensor soiling can learn the historical patterns in the time histories of neighboring sensors. Fusing precipitation data with solar observations enables the algorithm to identify sensor soiling from the pattern changes following a major precipitation event.

Each of these examples of finding and filtering out sensor errors requires pattern recognition. For small datasets, this can be performed by a skilled practitioner. However very large datasets require the type of automated pattern recognition offered by AI tools.

4. Validation Methodology

The validation of the GSID duplicated as closely as possible the methodology used to validate the NSRDB-PSM (Sengupta, et al., 2015; Habte, et al., 2017; Habte, et al., 2018). High-quality data from seven sites in the Surface Budget Radiation (SURFRAD) network (Figure 5) were downloaded from the NOAA wesbite and used as the reference. Data from all SURFRAD sites were intentionally excluded from the GSID so they could be used for validation purposes. For this comparison, data from the GSID ground site closest to each SURFRAD site were used. NSRDB-PSM satellite-based data (version 3.0.1, 2005-2018) was downoladed from the NREL's NSRDB Data Viewer (see https://maps.nrel.gov/nsrdb-viewer) by entering the latitude and longitude of each SURFRAD site into the viewer. The hourly data was aggregated into daily, monthly and annual values. The only material difference in the validation methodology was that NREL excluded observations where solar zenith angles were less than 80 degrees in the original NSRDB-PSM validation. This validation procedure excluded observations where global irradiance at the reference site was less than 50 W/m^2. Both methods were intended to exclude errors that can occur under low-light conditions near sunrise and sunset.

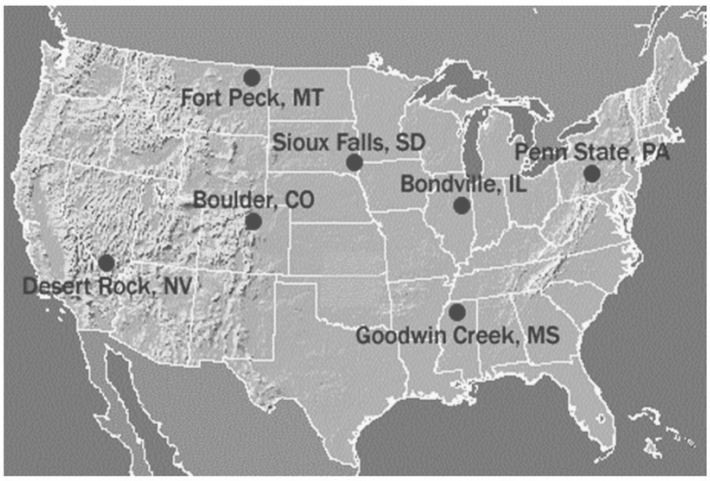

Fig. 5: SURFAD sites used to validate the National Solar Radiation Database (image courtesy of Surface Radiation Budget Network)

5. Validation Results

The primary validation statistics were Mean Bias Percent Error (an indicator of long-term bias in the observations), Mean Absolute Percent Error (an indicator of the average error in the observations) and the Overall Uncertainty at 95% confidence (a confidence interval for the observations). These statistics were calculated using the following equations:

Percent Error: $\%E = 100\ x\ \frac{(observed - true)}{true}$ (eq. 1)

Mean Bias Percent Error: $MB\%E = \frac{1}{n}\sum_1^n(\%E)$ (eq. 2)

Mean Absolute Percent Error: $MA\%E = \frac{1}{n}\sum_1^n|\%E|$ (eq. 3)

Root Mean Square Percent Error: $RMS\%E = \sqrt{\frac{1}{n}\sum_1^n(\%E)^2}$ (eq. 4)

Overall Uncertainty at 95% Confidence: $U_{95} = \pm 2\sqrt{\left(\frac{Uref}{2}\right)^2 + \left(\frac{MB\%E}{2}\right)^2 + \left(\frac{RMS\%E}{2}\right)^2}$ (eq. 5)

In these equations, n represents the number of observations and U_{ref} is the estimated uncertainty in the SURFRAD reference data. This validation used 5% for U_{ref}, as did the NRSDB validation.

The validation statistics for all seven sites combined are shown in Figures 6-8. The Appendix contains a table of the statistics for the individual comparison sites.

Mean Bias Errors for the NSRDB-PSM and GSID are shown in Figure 6. It can be seen that the GSID had significantly lower observation bias at the hourly, daily, monthly and annual levels.

Mean Absolute Errors for the NSRDB-PSM and GSID are shown in Figure 7. The GSID had almost half the observation error at the hourly and daily levels. Monthly and annual observation errors were similar for both datasets.

Overall Uncertainties at 95% Confidence for the NSRDB-PSM and GSID are shown in Figure 8. The NSRDB observations had twice the uncertainty of the GSID observations at the hourly and daily levels. Monthly and annual confidence intervals were similar for both datasets.

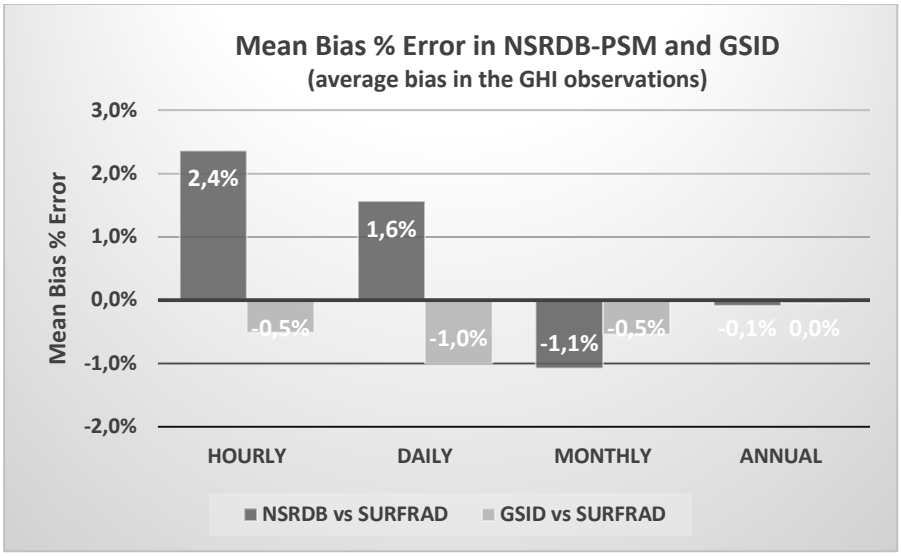

Fig. 6: Global irradiance: Mean bias errors (%) for the NSRDB-PSM and the GSID database

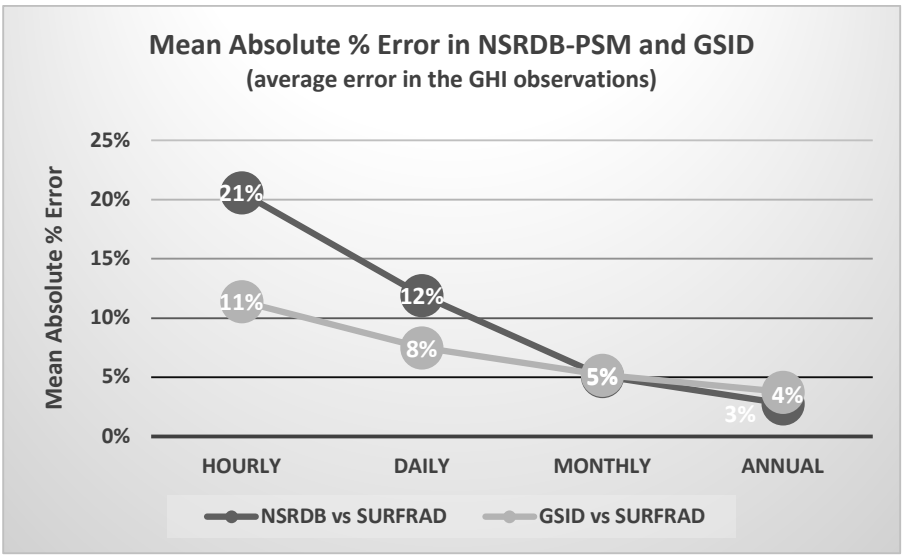

Fig. 7: Global irradiance: Mean absolute errors (%) for the NSRDB-PSM and the GSID database

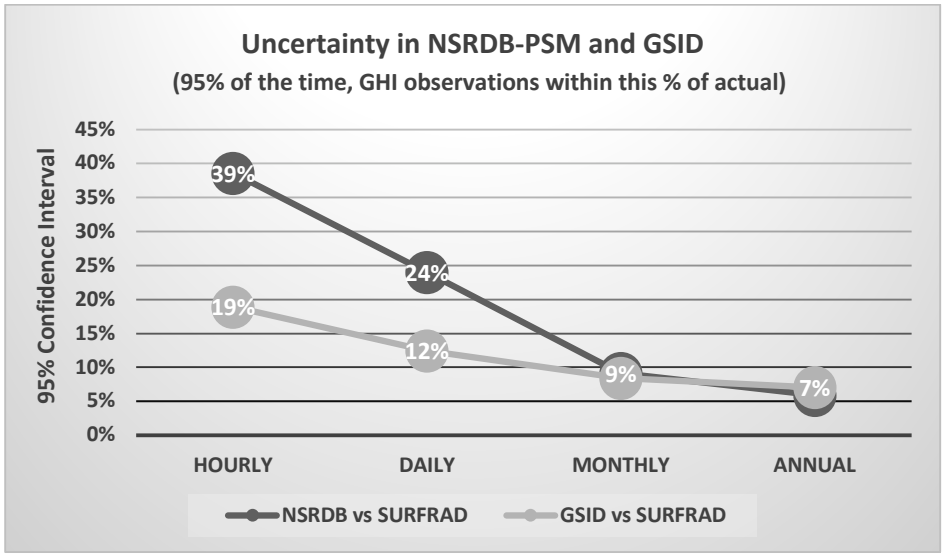

Fig. 8: Global irradiance: Overall uncertainty (%) at 95% confidence for the NSRDB-PSM and the GSID database

These statistics confirm that the GSID database has lower bias and is significantly more accurate than the NSRDB-PSM at the hourly and daily level. For applications requiring only monthly or annual resource data, either dataset would be acceptable.

One derivation of the daily GSID data, available without cost, is a gridded database (0.1-degree increments) of the continental US, Alaska and Hawaii for the years 2005-2018 (see www.SolarDataWarehouse.net). The full GSID database contains additional daily and hourly resource for measured global irradiance, estimates of beam irradiance, diffuse irradiance, and other meteorological measurements at over 7000 locations for the years 2005-present. The GSID also includes a suite of tools that can provide further data analysis, quality control, filtering, gridding, etc. that can be tailored to the specific needs of the end user.

6. Summary

A new solar resource with complete US coverage from 2005-present has been generated from ground-based observations: the GSID. The dataset has been quality controlled using AI technology and verified using the

same validation procedure as NREL's National Solar Radiation Database. The validation statistics confirm that the GSID database has lower bias and is significantly more accurate than the NSRDB-PSM at the hourly and daily level. For monthly or annual resource data, the GSID and NSRDB-PSM are comparable. The GSID can be used stand-alone, or combined with other solar resource for a variety of projects including site selection, power output simulations and on-going monitoring of solar plants.

7. References

Anders Nottrott and Jan Kleissl (2010). Validation of the NSRDB-SUNY Global Horizontal Irradiance in California. Solar Energy Volume 84, Issue 10, October 2010, Pages 1816-1827

Anders Nottrott and Jan Kleissl (2014). Validation of the National Solar Radiation Database in California. https://ww2.energy.ca.gov/2014publications/CEC-500-2014-017/CEC-500-2014-017.pdf, accessed June 2020

Aron Habte, Manajit Sengupta, Anthony Lopez (2017). Evaluation of the National Solar Radiation Database (NSRDB Version 2): 1998–2015. https://www.nrel.gov/docs/fy17osti/67722.pdf accessed June 2020

Aron Habte, Manajit Sengupta, Anthony Lopez, Yu Xie and Galen Maclaurin (2018). Assessment of the National Solar Radiation Database (NSRDB 1998-2016). https://www.nrel.gov/docs/fy18osti/71607.pdf accessed June 2020

Sengupta, M.; Weekley, A.; Habte, A.; Lopez, A.; Molling, C.; Heidinger, A. (2015). Validation of the National Solar Radiation Database (NSRDB) (2005–2012). https://www.nrel.gov/docs/fy15osti/64981.pdf accessed June 2020

Appendix: Statistical Results

Table. 1: Statistical Results at Varying Time Averages for the National Solar Radiation Database and the US Ground-Based Solar Irradiance Database (relative to USCRN ground reference sites)

			NSRDB vs SURFRAD			GSID vs SURFRAD		
		observations	MB%E	MA%E	RMS%E	MB%E	MA%E	RMS%E
Bondville	Hourly	52,112	4.1	20.3	37.9	0.4	6.9	11.0
	Daily	5,076	4.7	12.7	27.3	0.2	6.3	10.2
	Monthly	168	0.3	4.1	6.5	1.6	4.9	7.9
	Annually	14	1.0	2.0	2.2	2.1	3.9	5.9
Desert Rock	Hourly	56,715	-1.5	11.1	24.9	-0.4	4.4	6.7
	Daily	5,060	-1.9	4.9	9.4	-0.4	3.3	4.9
	Monthly	168	-2.3	2.5	3.1	-0.3	3.1	4.5
	Annually	14	-1.9	1.9	2.0	-0.3	2.7	4.0
Boulder	Hourly	54,641	-1.6	22.9	39.9	1.6	31.6	48.9
	Daily	5,079	-4.0	10.9	18.6	-1.0	16.0	25.0
	Monthly	168	-5.0	5.4	7.4	-0.2	6.8	8.6
	Annually	14	-3.8	3.8	4.3	0.7	4.4	4.9
Ft Peck	Hourly	53,254	-3.2	22.6	39.1	-0.2	5.8	8.2
	Daily	5,088	-5.4	14.6	25.0	-0.4	4.9	6.9
	Monthly	168	-6.4	8.8	14.7	0.8	3.7	4.5
	Annually	14	-3.0	3.6	4.5	2.3	2.9	3.2
Goodwin	Hourly	52,785	6.9	19.0	38.7	-4.0	14.9	24.7

	Daily	5,025	7.2	10.6	25.9	-3.5	11.0	13.9
	Monthly	168	3.2	3.9	5.6	-3.3	9.3	10.9
	Annually	14	3.2	3.2	3.5	-3.8	6.9	8.6
Penn St	**Hourly**	51,313	9.2	26.4	45.9	-1.8	9.4	18.1
	Daily	5,071	8.7	15.3	30.5	-2.9	7.3	11.1
	Monthly	168	4.4	5.1	6.5	-3.6	5.7	7.3
	Annually	14	3.6	3.6	4.0	-2.4	3.6	5.7
Sioux Falls	**Hourly**	52,909	2.7	21.8	41.3	0.9	6.7	10.3
	Daily	5,080	1.7	**14.4**	**27.6**	**1.0**	**3.8**	**7.8**
	Monthly	168	-1.8	5.9	9.5	1.2	2.9	3.8
	Annually	14	0.3	1.4	1.7	0.9	2.1	2.5
Averages	**Hourly**		2.4	20.6	38.2	-0.5	11.4	18.2
	Daily		1.6	11.9	23.5	-1.0	7.5	11.4
	Monthly		-1.1	5.1	7.6	-0.5	5.2	6.8
	Annually		-0.1	2.8	3.2	0.0	3.8	5.0