Modeling Solar Irradiance and Solar PV Power Output to Create a Resource Assessment Using Linear Multiple Multivariate Regression

CHRISTOPHER T. M. CLACK[a]

Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, Colorado

(Manuscript received 25 May 2016, in final form 24 September 2016)

ABSTRACT

The increased use of solar photovoltaic (PV) cells as energy sources on electric grids has created the need for more accessible solar irradiance and power production estimates for use in power modeling software. In the present paper, a novel technique for creating solar irradiance estimates is introduced. A solar PV resource dataset created by combining numerical weather prediction assimilation model variables, satellite data, and high-resolution ground-based measurements is also presented. The dataset contains ≈152 000 geographic locations each with ≈26 000 hourly time steps. The solar irradiance outputs are global horizontal irradiance (GHI), direct normal irradiance (DNI), and diffuse horizontal irradiance (DIF). The technique is developed over the United States by training a linear multiple multivariate regression scheme at 10 locations. The technique is then applied to independent locations over the whole geographic domain. The irradiance estimates are input into a solar PV power modeling algorithm to compute solar PV power estimates for every 13-km grid cell. The dataset is analyzed to predict the capacity factors for solar resource sites around the United States for 2006–08. Statistics are shown to validate the skill of the scheme at geographic sites independent of the training set. In addition, it is shown that more high-quality, geographically dispersed, observation sites increase the skill of the scheme.

1. Introduction

Over the last decade the use of solar photovoltaics (PV) has expanded dramatically. The deployment of solar PV has societal benefits, such as no pollution from electric power production, very little water use, abundance as a resource, silent operation, long lifetime, and little maintenance. However, the application of solar PV to electric grids has downsides, most notably the variability of power output, which can add strain to the system. The variable nature of solar PV could hamper further deployment or diminish the carbon mitigation potential because of the need for more reserves on the electric grid to compensate for fluctuations in the power output. For a more detailed overview of solar PV, see Dominguez-Ramos et al. (2010), Lueken et al. (2012), Mills and Wiser (2010), Parida et al. (2011), and Solanki (2009).

When estimating the solar PV power output, the following two-step procedure is generally carried out. First, meteorological data are supplied and the solar irradiance is estimated, and then the solar irradiance is input into a power modeling algorithm with information about the solar PV cell and temperature (Deshmukh and Deshmukh 2008; Huang et al. 2013; Zhou et al. 2007). The solar irradiance can be estimated for a past time (hindcasting), the present time (analysis), or for a future time (forecasting). Once the solar irradiance is found, the techniques for calculating the power output are essentially the same. The technique developed in this paper takes historical data and performs the algorithms as if it were the present time to create an analysis.

If the input solar irradiance for the PV power modeling is inaccurate, then the power output will be incorrect regardless of the precision of the power algorithm. There has been intensive research into the accuracy of the solar irradiance measurements (e.g., Geuder et al. 2003; Myers 2005) and improving the prediction of solar irradiance (e.g., Kratzenberg et al. 2008; Paulescu et al. 2013; Wong

Denotes Open Access content.

[a] Current affiliation: Vibrant Clean Energy, LLC, Erie, Colorado.

Corresponding author e-mail: Christopher Clack, chrisclack84@gmail.com

DOI: 10.1175/JAMC-D-16-0175.1

and Chow 2001). The prediction of solar irradiance usually falls into two categories. First, short-term prediction using an array of novel techniques, for example, neural networks (Wang et al. 2011). Second, and more commonly, using satellite data as a proxy, the solar irradiance is computed (Hammer et al. 1999; Houborg et al. 2007; Vignola et al. 2007). The aforementioned methods also use basic numerical weather prediction (NWP) model outputs or ground data. The present paper relies upon NWP assimilation data of hydrometeors complemented with satellite data. The solar irradiance (shortwave and longwave radiation fields) from the NWP assimilation model is not used because at time zero there is not a model output for it with the model being used. Moreover, some NWP assimilation models do not currently give direct-normal (the amount of radiation per unit area received by a plane perpendicular to the rays that come from the sun in a straight line) or diffuse (the amount of radiation per unit area that does not arrive in a direct path from the sun) radiation output fields.

Recently, there have been several studies on numerical weather prediction and solar energy (Mathiesen et al. 2013; Mathiesen and Kleissl 2011; Perez et al. 2013). In addition, there has been extensive effort at the National Renewable Energy Laboratory (NREL) to produce the national solar radiation database (http://rredc.nrel.gov/solar/old_data/nsrdb/) and there are commercial products available that provide resource mapping for the United States (from, e.g., Vaisala, Clean Power Research, or GeoModel Solar). All of these products are estimates, are not produced in concert with other weather-driven renewables, and are subject to improvement. The improvements could be higher spatial resolution, higher temporal resolution, and reductions in biases or RMSE. Nevertheless, the production of these products shows the growing need within the United States for datasets of solar irradiance and power. In theory, all these products can have the procedure to be outlined in the present paper applied to them (to further enhance the accuracy of the results). The model developed in the present paper finds estimates for the entire United States at a spatial discretization of 13 km and temporal resolution of 1 h for 3 yr. The scale of the model and its inputs is a first and is a demonstration that will be applied to much larger datasets in the near future. It is also the first to combine satellite and NWP assimilation data, along with ground-based observations, for solar irradiance estimates using multiple multivariate linear regression over such a wide spatial and temporal range with high resolution.

To produce accurate solar irradiance estimates, the use of excellent quality solar measurements is fundamental. The United States has many such high-quality measurement networks. Two of them are used in the present paper: the Surface Radiation Budget Network (SURFRAD; http://www.esrl.noaa.gov/gmd/grad/surfrad/) and the Integrated Surface Irradiance Study Network (ISIS; http://www.esrl.noaa.gov/gmd/grad/isis/). For more information on these two networks, see Augustine et al. (2005), Hicks et al. (1996), and Wang et al. (2012). The present paper uses all seven of the SURFRAD sites and five of the ISIS sites for the majority of the solar irradiance measurements. The locations of the SURFRAD sites are Bondville, Illinois; Table Mountain, Colorado; Desert Rock, Nevada; Goodwin Creek, Mississippi; Fort Peck, Montana; The Penn State University, College Park, Pennsylvania (PSU); and Sioux Falls, South Dakota. The locations of the ISIS sites are Albuquerque, New Mexico; Madison, Wisconsin; Salt Lake City, Utah; Sterling, Virginia; and Hanford, California (HNX). There are three sites from the ISIS network that were not active during the study dates (2006–08) and, therefore, are not included (Seattle, Washington; Bismarck, North Dakota; and Tallahassee, Florida). The locations of the measurement sites are shown in Fig. 1.

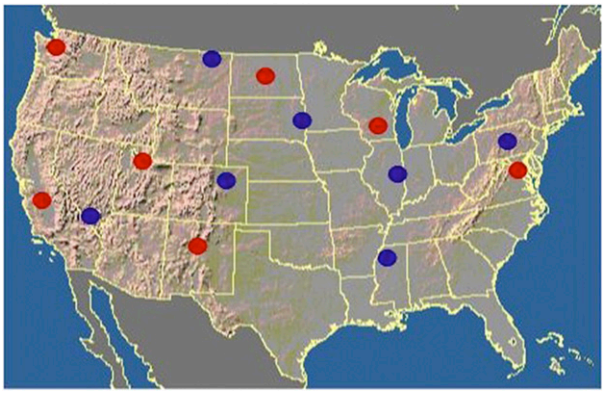

FIG. 1. Geographic locations of the SURFRAD (blue) and ISIS (red) network sites. Images are courtesy of NOAA's Global Monitoring Division.

To investigate the validity of the scheme employed, seven other publicly available solar irradiance measurement sites are leveraged to compare the solar irradiance estimates and the observations at these independent sites. Two sites, Elizabeth, North Carolina, and Golden, Colorado, were acquired from the Measurement and Instrumentation Data Center (MIDC) run by the NREL (http://www.nrel.gov/midc/) and the remaining five sites (Burns, Oregon; Silver Lake, Oregon; Herminston, Oregon; Moab, Utah; and Dillon, Montana) were obtained from the University of Oregon Solar Radiation Monitoring Laboratory (http://solardat.uoregon.edu/SolarData.html). Additionally, one ISIS (Hanford, California) and one SURFRAD (The Pennsylvania State University, College Park, Pennsylvania)

location were reserved exclusively to serve as further validators. In total, three years of data (2006–08) at 10 training and 9 validation sites were concatenated for the proposed method.

The primary goal of the present paper is to provide a novel technique for computing solar irradiance and solar PV power estimates that can be applied to any weather model. The secondary goal is to produce a high-quality demonstration resource mapping dataset of solar irradiance and solar PV power over the United States at high resolution (13 km, hourly). The paper is organized as follows. Section 2 explains the basic methods of the technique, its mathematical underpinning, and the data processing; section 3 contains the procedure carried out for the solar irradiance estimates, along with the statistics associated with its implementation; section 4 explains the power modeling algorithm using the solar irradiance as inputs; and finally, in section 5, the conclusions and future work are discussed.

2. Data and methods

The method used in the present paper for solar irradiance estimates is linear multiple multivariate regression (Pearson 1908; Stanton 2001). The first task is to collect all the data that are needed: NWP assimilation model variables on an hourly basis, GOES-East satellite data for the continental United States, and ground-based measurements of global horizontal irradiance (GHI), direct-normal irradiance (DNI), and diffuse horizontal irradiance (DIF). The GHI is the total amount of irradiance falling on a horizontal unit area. The DNI is defined as the amount of irradiance falling on a unit area that is perpendicular to the rays propagating in a straight line from the sun. The DIF is the amount of irradiance falling on a horizontal unit area that is not directly from the sun. The satellite measurements are at 15-min temporal resolution for the years 2006–08. There is a percentage of time when there were not any satellite data available because of full disk images, maintenance, and other malfunctions, which resulted in a dataset with 87.99% of the hours having all of the wavelengths required.

The numerical weather prediction assimilation model used is the 13-km Rapid Update Cycle (RUC; information online at http://ruc.noaa.gov/). The satellite data are obtained from the Geostationary Operational Environmental Satellite-East (GOES-East; http://www.ssec.wisc.edu/datacenter/archive.html). All of the data are publicly available. The RUC was used because having a dual dataset with wind and solar PV power that are on a synchronous temporal scale and spatial grid was desired. Moreover, the technique (or model) is devised to be as accessible as possible so that as many users as possible can utilize it with different models and geographic areas.

The author at the time of writing was only able to handle the data from the GOES-East satellite. It would have been beneficial to have a combination of the GOES-East and -West satellite datasets. The parallax effect created by only having the GOES-East data is minimized by NOAA algorithms for use in NWP models and, thus, is assumed to be negligible on the regression results. It is understood, however, that there is still an effect. The regression would be more successful with blended satellite data. Five channels of satellite data are utilized: four in the infrared spectrum [3.8–4.0 μm, 6.5–7.0 μm (water vapor), 10.2–11.2 μm, and 11.5–12.5 μm] and one in the visible spectrum (0.55–0.75 μm). The data are simply the unsigned bit count values on a scale of 0–255. The count values B can be converted to temperature T using the following formulas:

$$T = \frac{1}{2}(660 - B) \quad 0 \le B \le 176 \quad \text{and}$$
$$T = 418 - B \quad 176 < B \le 255. \tag{1}$$

The temperature in Eq. (1) has units of kelvins. The count values are used instead of the temperature because they stretch out the highest temperatures (0.5 K per count) and map directly (one to one) to the lowest temperatures (1 K per count). The geographic resolution of the satellite data is 4 km, except for the visible, which is 1 km. Since the spatial resolution of the RUC is at 13 km and the temporal resolution is 60 min, interpolations were performed to bring the satellite data to the RUC discretization. The satellite data are regridded to the RUC resolution for three reasons. First, coarser resolution is computationally easier for the demonstration dataset. Second, the required dataset is designed to be coincident with a wind dataset from Clack et al. (2016) on the 13-km grid that utilizes the same model physics. Third, interpolating from a finer resolution to a coarser one will smooth the data, whereas the reverse will be an extrapolation of data and is subject to more errors. The spatial regridding is performed using weighted data points from nearby cells and a cubic spline fit from 4 km (and 1 km) to the 13-km grid. The temporal interpolation was only used if the top of the hour (hh00) was not available (when the NWP assimilation model data are output) because of maintenance of the satellite or full disk scans. A linear interpolation was applied for successive 15-min intervals around the top of the hour up to a maximum of 45 min on each side of that hour. If there were no data for the whole period of (hh − 1)15–hh45, no interpolation is applied and no satellite data are reported. In total, a dataset was

created that contained all five channels on 23 145 h of the possible 26 304 h between 2006 and 2008. Because of missing satellite data, multiple regressions were performed to increase the accuracy of the solar irradiance estimates in the absence of some of the satellite channels.

The RUC is cycled hourly for the whole 3-yr period of 2006–08. The RUC assimilates thousands of measurements across the contiguous United States. The 3D data assimilation matrix was downloaded for each hour for the three years. For the purposes of the solar irradiance modeling, the following variables were extracted from the data: water vapor, cloud water, rain, cloud ice, snow, graupel, and temperature at 2 m. All the variables, except temperature, are the total throughout the vertical column within the model. The variables were chosen because of their known direct impact on solar irradiance attenuation. After all of the data were extracted, there were 25 663 h remaining of the 26 304 possible (97.6%).

In addition to the satellite and NWP assimilation data, the solar irradiance falling onto the top of the atmosphere is computed for each hour. The irradiance at the top of the atmosphere takes into account the eccentricity of Earth's orbit. The average extraterrestrial irradiance (I_0), about which the irradiance fluctuates, is 1360.8 W m^{-2} (Kopp and Lean 2011; Vignola et al. 2012). The equation for the extraterrestrial irradiance outside Earth's atmosphere (normal to the photosphere of the sun) is

$$\mathrm{DNI}_0 = I_0\left(\frac{R_{\mathrm{av}}}{R}\right)^2, \tag{2}$$

where R_{av} is the mean sun–Earth distance and R is the actual sun–Earth distance at a specific instant. An approximation for $(R_{\mathrm{av}}/R)^2$ was used:

$$\left(\frac{R_{\mathrm{av}}}{R}\right)^2 \approx 1.000\,110 + 0.034\,221\cos(\delta) + 0.001\,280\sin(\delta) + 0.000\,719\cos(2\delta) + 0.000\,077\sin(2\delta). \tag{3}$$

Here, $\delta = 2\pi d/365.242$ radians, and d is the day of the year (Spencer 1971). The error associated with the Fourier approximation is very small (0.0001%). Another parameter that was computed for the dataset was the solar zenith angle (sza). The solar zenith angle is defined as

$$\cos(\mathrm{sza}) = \sin(\mathrm{lat})\sin(\mathrm{dec}) + \cos(\mathrm{lat})\cos(\mathrm{dec})\cos(\mathrm{ha}), \tag{4}$$

where dec is the declination angle, ha is the hour angle, and lat is the latitude in radians. The declination angle can be approximated by (Spencer 1971)

$$\mathrm{dec} = \varepsilon\sin\left\{\delta + \frac{\pi}{180}[279.93 + 1.915\sin(\delta) - 0.0795\cos(\delta) + 0.02\sin(2\delta) - 0.001\,62\cos(2\delta)]\right\}, \tag{5}$$

where ε is Earth's axial tilt or obliquity of the ecliptic in radians (0.409 173°). The hour angle is simply computed as

$$\mathrm{ha} = \pi\left(1 - \frac{\mathrm{hr}}{12}\right) - \mathrm{lon}, \tag{6}$$

with hr being the hour of the day in UTC and lon the longitude in radians. Equation (6) applies when lon < 0 (as is the case for the contiguous United States); when lon ≥ 0, then ha $= \pi(\mathrm{hr}/12 - 1) + \mathrm{lon}$.

The ground-based observations of solar irradiance are taken from publicly available sites across the contiguous United States. Both the SURFRAD and ISIS sites have a measurement frequency of 3 min. Averages of the solar irradiance measurements were taken over time to compensate for the fact that the SURFRAD and ISIS sites are point measurements and the NWP assimilation model variables are over a gridded area. The averages are taken from 6 min before the top of the hour to 6 min after the top of the hour (five measurements). The averaging time was chosen to balance the need for accurate measurements along with the need for a reliable average value to use in the regression. It is designed to be short enough that the clouds do not have enough time (on average) to advect fully across the RUC cell, but long enough to remove scattered cloud in a small percentage of the box that happens to be over the measurement site at a single time. The chosen time scales gave the best overall performance, which is defined as the lowest bias and RMSE values for the training set comparisons. Solar irradiance measurement averages that were produced from *all* of the data points were used. All of the times of the measurements were shifted to coordinated universal time (UTC) to make sure all data at different locations match with the NWP and satellite data. Only time steps that had both measurements of DNI and DIF were included. The DNI is measured at all sites with a Normal Incidence Pyrheliometer, while the DIF is measured with an Eppley 8–48 "black and white" pyranometer. The irradiance measurements are spectrally integrated between 280 and 3000 nm. The SURFRAD and ISIS sites do measure GHI; however, the measurements are less accurate than calculating the GHI from the DNI and DIF measurements, known as the component-sum technique (Michalsky et al. 2003):

$$\mathrm{GHI} = \mathrm{DNI}\cos(\mathrm{sza}) + \mathrm{DIF}. \quad (7)$$

The instrument errors were taken to be ±1% of the observed value (see documentation online at http://www.esrl.noaa.gov/gmd/grad/instruments.html). The instrument errors are in a simplistic form for computational expedience; however, it is recognized that for a more accurate regression, the errors should be taken for each instrument at each site. The SURFRAD and ISIS sites were chosen because of their high quality, regular servicing, and calibration.

Once the NWP assimilation data, ground measurements, and satellite data are collated, the linear multiple multivariate regression can be performed. The regression can be represented mathematically as

$$Y_{n\times p} = X_{n\times(r+1)}\beta_{(r+1)\times p} + \varepsilon_{n\times p}, \quad (8)$$

where $Y_{n\times p}$ are the endogenous variables or regressands, $X_{n\times(r+1)}$ are the exogenous variables or regressors, $\beta_{(r+1)\times p}$ are the effects or regression coefficients, and $\varepsilon_{n\times p}$ are the disturbance or error terms. In Eq. (8), n is the number of observations, p is the number of different properties modeled, and $(r + 1)$ is the number of independent inputs. For our specific cases, Y are the ground-based measurements of GHI, DNI, and DIF; X are the NWP assimilation model variables and satellite data; ε is the residuals from the model versus data; and β are the regression coefficients to be applied to all other locations when the training set has been regressed against. It is assumed that the expected value of the error term is zero; that is, $E(\varepsilon_i) = 0$. It was also assumed that the errors are independent between species or irradiance; that is, $\mathrm{cov}(\varepsilon_i, \varepsilon_k) = \sigma_{i,k}I$, $i, k = 1, 2, \ldots, p$. The irradiance species are dependent; however, assuming they are not does not significantly change the results of the regression in comparison with performing them separately (the RMSE and bias are the same to two decimal places). Computationally, the linear multiple multivariate regression is more efficient. The solution of the linear multiple multivariate regression can be found to be

$$\hat{\beta} = (X'X)^{-1}X'Y, \quad (9)$$

with $\hat{\beta}$ being the estimators of the regression. Equation (9) is derived by minimizing Eq. (8). The minimization finds the smallest sum of deviations from *all* the independent variables. The estimators are placed into

$$I_q = \sum_{j=1}^{r+1} \hat{\beta}_j x_j \quad (10)$$

to model the irradiance at all locations over the domain being studied. The estimated GHI, DNI, or DIF at a single instant (hour) is I_q. There are numerous software packages that find the solution of Eq. (8). The IDL Advanced Statistics package was used to perform the regressions. The algorithm takes advantage of single-value decomposition to ensure that the matrix inversion is accurate. When the regressions are carried out, the analysis of variance (ANOVA) can be performed to determine the performance of the technique. Once the values for each $\hat{\beta}_j$ are found, those values can be applied throughout the contiguous United States.

3. Solar irradiance estimates

As established in section 2, satellite data, numerical weather model assimilation data, and ground-based measurement data that have been interpolated to exactly the same gridded space over the contiguous United States with a temporal resolution of an hour were obtained. The ground-based measurements are at 10 different sites for the regression and 9 independent sites for validation purposes. Once the quality control and nighttime removal had taken place, 32 different regressions were performed for each of the irradiance species. The large number of regressions was required to account for times when some (or all) of the satellite data were unavailable. To get the most comprehensive dataset possible required carrying out the regression with data being denied to replicate the missing data. Training the regressions in this manner allows for all eventualities when applying the technique to sites outside the training cells. In addition, a further regression with just the satellite data (not assimilation data) was computed to compare our new technique with the simple technique of regressing only against satellite data and the extraterrestrial irradiance. For the sake of brevity, the results of every single regression are not shown, but rather the results from the three main regressions are presented: those that include *all* the data, those that include only the *satellite* data, and those that include only the *assimilation* data. Further to this, comparisons between the *overall* output from the procedure (which uses the appropriate regressions when necessary) and the measurements at the training and validation sites are performed.

The regressors x_j for Eq. (8) are as follows: x_0 is a constant; x_1 is the total solar irradiance at the top of the atmosphere, corrected for the variability of the distance of Earth from the sun, multiplied by the cosine of the zenith angle; x_2 is the water vapor; x_3 is cloud water; x_4 is rain; x_5 is cloud ice; x_6 is snow; x_7 is graupel; x_8 is 2-m temperature, x_9 is a 4-μm satellite; x_{10} is an 11-μm satellite; x_{11} is a 13-μm satellite; x_{12} is a visible satellite; and x_{13} is the satellite water vapor. Thus, x_0 and x_1 are calculated, x_2–x_8 are the RUC assimilation model hydrometeors, and x_9–x_{13} are the satellite measurements.

TABLE 1. RUC assimilation model GHI regression coefficients. The $\hat{\beta}_j$ are the coefficients that multiply the regressors x_j (written out in the text) that linearly combine to provide the irradiance estimates. The regression with both the assimilation and satellite data is GHI A, the satellite only regression is GHI B, and the assimilation only regression is GHI C.

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
GHI A	716	0.694	−21.2	−211	91.9
GHI B	11.3	0.696	—	—	—
GHI C	−727	0.659	−43.5	−447	251
	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$
GHI A	−74.8	−71.2	617	−2.06	0.866
GHI B	—	—	—	—	2.37
GHI C	−402	−192	1520	2.69	—
	$\hat{\beta}_{10}$	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{13}$	
GHI A	−3.71	1.83	−1.36	0.528	—
GHI B	−4.59	2.21	−1.23	0.378	—
GHI C	—	—	—	—	—

The linear multiple multivariate regression was performed over the period 2006–08 to improve the accuracy of the procedure. The total number of *training* data points is 81 434 for each of the irradiance species, which is very dense. However, it was found that each addition of an extra site improved the regressions' performance in terms of mean biased error (MBE), RMSE, and coefficient of variation (CV), and thus the regression has not been saturated or overfitted. Additional sites would be most beneficial from areas of poorly sampled climates, that is, remote locations relative to the existing training set of locations.

Increasing the number of training data points will increase the value of the regular definition of the multiple linear correlation coefficient (the dimensional extension of R^2, so the symbol is retained); thus, when analyzing the statistics, only the adjusted version is computed, $\overline{R}^2$, which takes into account the additional data points by (Theil 1961)

$$\overline{R}^2 = 1 - (1 - R^2)\frac{\rho - 1}{\rho - \eta - 1} = R^2 - (1 - R^2)\frac{\eta}{\rho - \eta - 1}, \tag{11}$$

where η is the number of regressors and ρ is the sample size.

The linear multiple multivariate regression coefficients are shown in Tables 1–3. The A denotes the regression that includes all the data, B designates the regression that includes only the satellite data, and C represents the regression that only includes the assimilation data. To reiterate, when the coefficients are applied to locations outside the training domain, the model utilizes the best of the 32 multivariate regressions based upon the data available for that time step. As the linear multiple multivariate regression can result in negative values, a nonnegative filter is applied and sets negative values to zero. The tabulated form of the regression coefficients allows us to compare which terms significantly change when the regression is altered. For example, it can be seen that $\hat{\beta}_1$ is almost completely unchanged between the three regressions in Table 1, which is to be expected as the coefficient relates how the solar irradiance at the top of the atmosphere multiplied by the cosine of the zenith angle affects the irradiance.

TABLE 2. As in Table 1, but for DNI.

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
DNI A	1170	0.447	−75.0	−293	375
DNI B	334	0.411	—	—	—
DNI C	−2510	0.280	−131	−746	751
	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$
DNI A	−8.76	−94.2	719	−1.95	5.58
DNI B	—	—	—	—	7.74
DNI C	−807	−342	2180	10.2	—
	$\hat{\beta}_{10}$	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{13}$	
DNI A	−11.5	3.46	−1.63	0.703	—
DNI B	−13.5	5.08	−1.45	0.500	—
DNI C	—	—	—	—	—

TABLE 3. As in Table 1, but for DIF.

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
DIF A	−179	0.159	17.0	−36.8	−49.4
DIF B	−4.43	0.157	—	—	—
DIF C	750	0.185	34.7	132	−205
	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$
DIF A	−30.2	−36.7	−1.62	0.341	−2.10
DIF B	—	—	—	—	−2.86
DIF C	261	47.3	−477	−2.76	—
	$\hat{\beta}_{10}$	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{13}$	
DIF A	3.64	−1.06	0.273	−0.0412	—
DIF B	4.27	−1.36	0.0816	−0.0436	—
DIF C	—	—	—	—	—

The same coefficient is only slightly altered for the DNI and DIF regressions as well, as shown in Tables 2 and 3. The satellite coefficients are not changed dramatically between regressions A and B (order of magnitudes are typically the same), but their values are slightly altered. This result is to be expected because the assimilation data were included to provide information about the optical thickness (water content) of the clouds that the satellites measure. For the majority of the time, this results in an important correction but does not necessitate a large alteration in the satellite coefficients. The final use of Tables 1–3 is to facilitate the procedure to be leveraged without the need to repeat the training of the regression for other users. The users would need satellite and/or RUC assimilation information at their location to produce an estimate of the resource at their site for a time period not encapsulated in the dataset produced by the present paper.

To analyze the performance of the linear multiple multivariate regressions, various statistics are calculated because a single statistic on its own may improve when the performance could be considered to be diminished depending upon the eventual use of the data. The most important statistics are displayed in Table 4 for the training set only and the values are for the hourly data. Within the training set, there are 10 different sites, and the accuracy of the regression varies from site to site, but the salient features are captured in the displayed combined statistics (because it is a requirement that the dataset be as accurate as possible over as many sites as possible). In Table 4 it becomes clear that the regression is best at estimating the global horizontal irradiance (in terms of all metrics shown). The range of GHI MBE is 2%–4% for all of the regressions, which is similar to results found by others that consider much smaller geographic areas (Vignola et al. 2007). The adjusted multiple linear correlation coefficient is in the high 90% range, which, along with the RMSE and CV of 20%–25%, shows great accuracy in predicting the GHI at the training sites overall.

It can be seen in Table 4 that the regressions get progressively worse as data are removed from them. The negative bias gets larger between A and C, $\overline{R}^2$ decreases, and both RMSE and CV increase. The regression with only satellite data (B) is better than the assimilation data only (C), and both are worse than when satellite and assimilation data (A) are used in concert. The improvement can be attributed to the removal of errors and biases with the combination of the two data types. The remaining unexplained variance and error is likely

TABLE 4. Statistics of the regressions over all of the training sites. Regression A has both the assimilation and satellite data, regression B is has satellite-only data, and regression C is the assimilation-only scheme.

Irradiance		Mean (W m^{-2})	MBE (%)	$\overline{R}^2$(%)	RMSE (%)	CV (%)
GHI	A	442.00	−2.82	94.17	20.67	20.48
	B	442.00	−3.33	92.96	22.63	22.39
	C	442.00	−4.26	91.08	25.60	25.25
DNI	A	512.37	−12.41	77.75	41.82	39.94
	B	512.37	−15.33	71.80	47.92	45.40
	C	512.37	−22.16	54.29	57.46	53.01
DIF	A	148.66	−4.19	82.87	42.42	42.21
	B	148.66	−4.63	80.83	44.56	44.32
	C	148.66	−6.90	69.20	55.40	54.97

TABLE 5. As in Table 4, but only over two initial validation sites.

Irradiance		Mean (W m^{-2})	MBE (%)	$\overline{R}^2$(%)	RMSE (%)	CV (%)
GHI	A	458.13	2.41	89.37	19.57	19.42
	B	458.13	2.67	88.16	20.67	20.50
	C	458.13	1.08	83.91	24.03	24.01
DNI	A	468.03	2.35	65.91	39.51	39.44
	B	468.03	0.21	58.98	43.27	43.27
	C	468.03	−9.80	41.86	52.93	52.01
DIF	A	164.60	−9.26	66.26	40.33	39.25
	B	164.60	−10.32	63.43	42.08	40.80
	C	164.60	−10.60	48.26	49.92	48.78

to be due to measurement errors, aerosols, and the averaging of single point data over a gridded space. It is worth noting that the spatial resolution of the irradiance estimates is 13 km, yet they are able to reproduce accurate estimations by other models that are at higher resolution (Vignola and Perez 2004). The direct normal irradiance estimates are the worst in terms of MBE and $\overline{R}^2$. The large negative bias is associated with the spatial resolution of the satellite and assimilation data versus the single-point measurements of DNI. The measurement site can have small clouds (and aerosols) pass by that specific site, but not be registered in the estimate. Another source of error is that the regression uses vertical column values. Thus, when the irradiance ray is impinging at an angle, it may be attenuated by the atmosphere in neighboring cells.

The statistics shown so far are for the training set. One SURFRAD and one ISIS site were retained to perform an "initial" validation of the procedure at two independent sites excluded from the training set. In Table 5, the same statistics as in Table 4 are shown, but for the two initial validation sites. Again, these are for the hourly values. Table 5 shows that in general terms the validation sites perform as would be expected. That is, there are no significant changes in RMSE, CV, or $\overline{R}^2$. However, there are some differences that are worth discussing. The sign of the biases of the GHI and DNI are reversed and the $\overline{R}^2$ is lower than was previously found, which suggests that the procedure is less accurate at sites independent from the training set, which is to be expected.

To take a different look at the accuracy, analysis of the residuals of the estimated irradiance minus the ground-based measurement was carried out. The probability density functions (PDFs) of the residual divided by the measurement (relative error) were computed and are plotted in Fig. 2. In the images the black lines are for regression A, with both the assimilation and satellite data; the red lines are for regression B, the satellite-only version; and the blue lines are for regression C, which includes the assimilation only. The top panel in Fig. 2 is the histogram for the training sites and the bottom panel is for the validation sites. It is clear from the panels that the training sites histograms are sharper, and the negative bias can be seen (left of the zero line), which is also listed in Table 4. The left-hand tail of the PDF for both the training and validation panels in Fig. 2 falls off faster

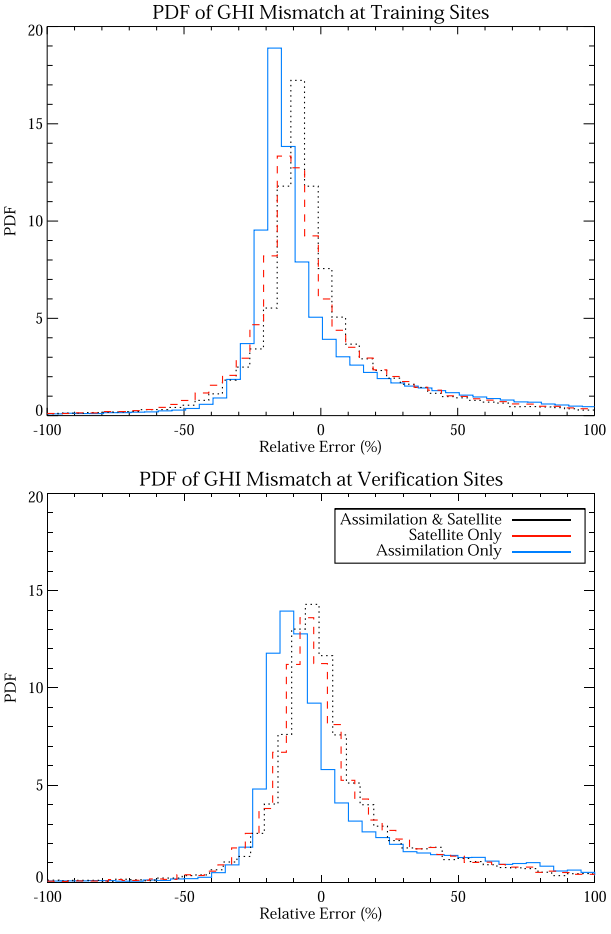

FIG. 2. Histograms of the difference between the estimated GHI and the measured GHI at the (top) training and (bottom) verification sites. The black dotted lines denote regression A with both the assimilation and satellite data; the red dashed lines are for the satellite-only regression; and the blue solid line is for the assimilation-only regression C. The relative error is the difference divided by the measurement.

than the right-hand tail. It becomes apparent that the black histogram (NWP and satellite data) is more centered about the zero line, and the narrowest. The worst performance is displayed by the blue lines (NWP data only). The two different plots show the same general characteristics, indicating that the technique is working well at sites that are independent to the training sites.

It is instructive to see the training and validation computations versus the measurements for comparison. In Fig. 3, the GHI, DNI, and DIF differences are shown (estimated minus measurements) versus the measurements for the three regression types. The panels in Fig. 3 show the median values of the differences with solid lines. The 25% and 75% percentiles are shown as the horizontal bars. Additionally, the vertical bars continue to the 10% and 90% percentiles. For comparison, guidelines are added to the panels that show 25% (dotted), 50% (dotted–dashed), and 100% (dashed) relative errors. The vertical lines are separated for image clarity, but are computed at the same points. Further, Fig. 4 displays the differences (estimated minus measurements) versus the zenith angle. The same percentiles as in Fig. 3 are shown.

The top panel in Fig. 3 shows the GHI differences versus the measurements. There are three colors in Fig. 3, which represent the three regression types being displayed in the present paper. The black is for regression scheme A, the red is for scheme B, and blue is for scheme C. All three are plotted in the same figure to illustrate that they all have the same overall features with regard to bias and slope; however, there is increasing accuracy and decreasing scatter from scheme C to A. This provides some verification that the additional data improve the performance of the model. It shows that, in general, the estimated GHI is close to the measured result with a slight positive bias (on average) at low irradiance and a slight negative bias (on average) at high irradiance. Note that the median of the error peaks at 150 W m^{-2}. The range of errors is largest between 200 and 400 W m^{-2}, which could be attributed to scattered cloud within the gridded domain over the observation site and, possibly, clouds that are not in the grid cell, but rather in neighboring cells that are affecting the measurements, whereas the regression has no knowledge of these clouds. It could also be attributed to the parallax effect of only using a single-satellite data stream. After 400 W m^{-2}, the median errors become negative. It can be seen that the median errors remain within 25% of the observations, with the exception of very low values of irradiance. The distribution of errors is narrower (sharper) for the combined regressions when compared with the other two. The middle panel in Fig. 3 displays the DNI differences and the bottom panel shows the DIF differences. The DNI differences have much larger

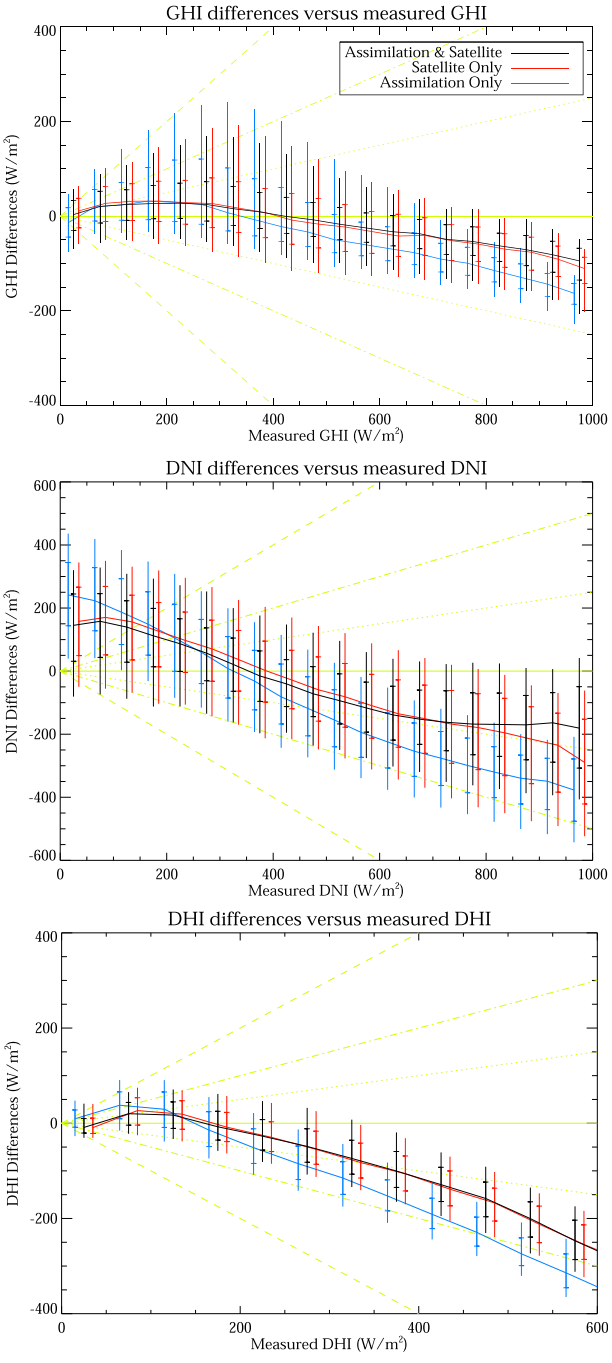

FIG. 3. The difference between the estimated irradiance and the measurement vs the measured irradiance: (top) GHI, (middle) DNI, and (bottom) DIF. The black, red, and blue are for regression scheme A, B, and C, respectively (like all other figures). The light green line designates the zero line.

slopes than the GHI and the variance of the error is also larger (as shown in Tables 4 and 5). The larger slope, from a positive bias to a negative bias with increasing irradiance, is predominantly due to the point-to-grid averaging, the parallax effect of a single-satellite data

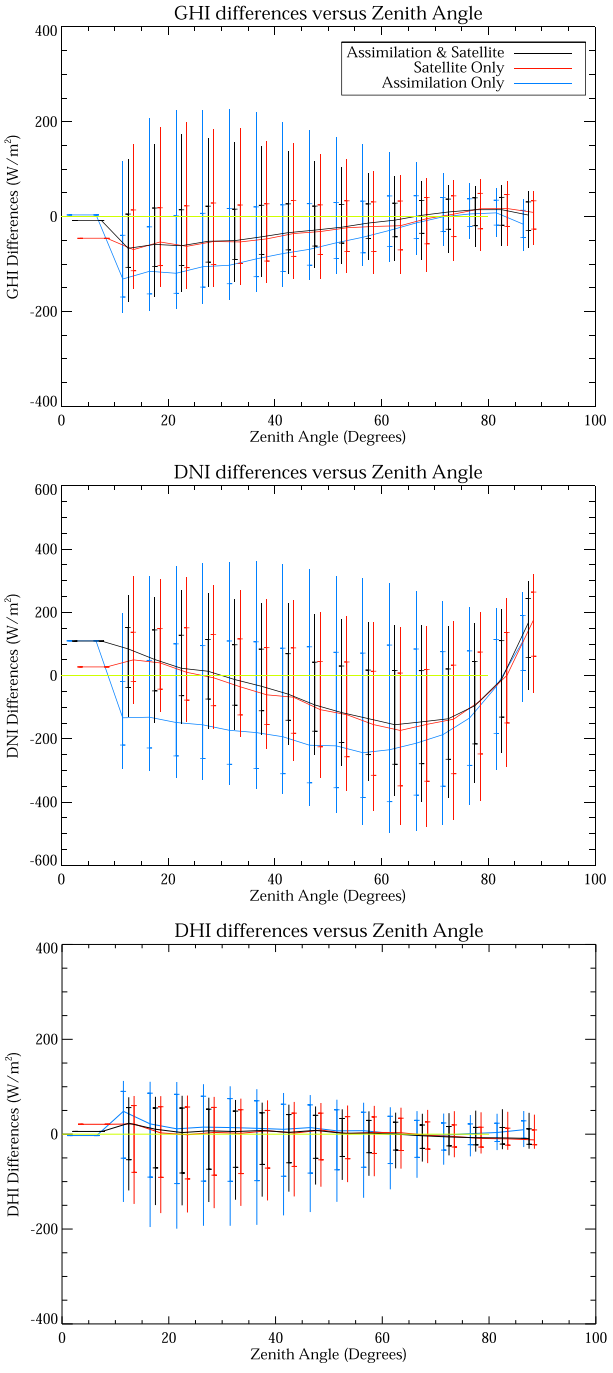

FIG. 4. As in Fig. 3, but vs the zenith angle.

stream, and nonmodeled aerosols. The more extreme values occur in the wintertime. The slope is typical when this type of computation is carried out (Vignola and Perez 2004; Vignola et al. 2007). The regression including all the variables is more accurate than the other regressions, particularly at high DNI values. The DIF differences also show a slope after about 200 $\mathrm{W\,m^{-2}}$ toward a negative bias and could be explained by the same effects as the DNI biases.

The information gained by displaying Fig. 4 is the dependency of the errors on the measurement zenith angle. It is obvious that GHI and DIF have no statistical dependency on the zenith angle for any of the regressions, whereas the DNI seems to have an increasingly negative bias from 20° to 70° and then becomes a positive bias by 85°. The dependency occurs in all three of the regression types, but the least effected is the regression with *both* satellite and assimilation data. It is thought that the dip is caused by interference of the beam by clouds, aerosols, and atmospheric disturbances in neighboring grid cells (nearby locations) that are not in the regression. The effect is over a large range of zenith angle values due to (a smaller effect of) high-level clouds, and then as the sun progresses through the sky, the DNI is blocked by a lower, and usually thicker, atmosphere in the surrounding cells. The same phenomenon is seen in Vignola and Perez (2004) and Vignola et al. (2007); however, because of their smaller dataset, they found it not to be statistically significant. Here, it is shown that it is a real effect, not just anomalous outliers. One way to correct this would be to perform the regression not in terms of the vertical column, as is done in the present paper, but rather in terms of the path integral of the DNI beam (along the zenith angle); however, this is a substantially harder problem to solve, which the author plans to address in future work. It should be noted that some of the effect may be attributed to the parallax angle created by using only the GOES-East satellite dataset, because it is reduced in the assimilation-only regression.

In creating the previous statistics, residuals, and histograms, only the training sites and the two verification sites have been analyzed. The following part of the present section will analyze the results from the seven independent sites provided by NREL and the University of Oregon when the full model has been applied to them. The model applies the best regression (of the 32) with the data available for each hour and geographic location. The analysis of these results will give a fuller description of how the regression model is working at sites completely separate from the training set (both in terms of location, but also for the agency responsible for the sites). The seven sites have different frequencies of measurement than the SURFRAD and ISIS sites, typically being 5 min. When necessary, the averaging of the measurements was altered to give accurate top-of-the-hour results. For the 5-min output frequency, averaging was carried out from 15 min before to 15 min after (seven measurements). The alteration of the averaging does have an impact on the metrics of the performance of the solar irradiance model. Figures 5–7 display time series of the measured and estimated solar irradiance. The top panels are for the

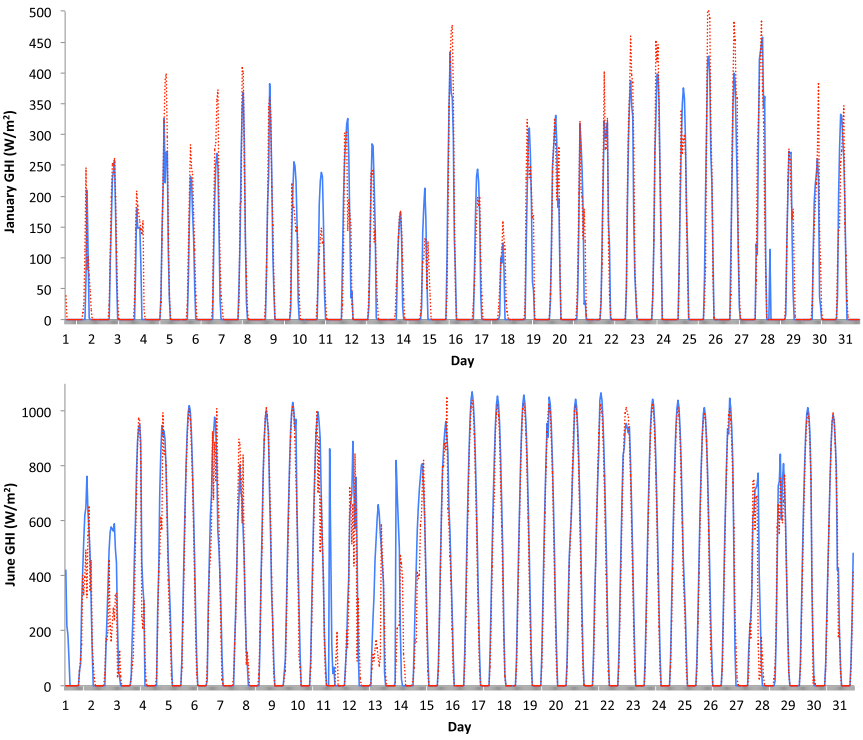

FIG. 5. Time series of measured (dashed red) and estimated (solid blue) GHI for Burns: (top) the 31 days from 1 Jan 2006 and (bottom) the 31 days following 1 Jun 2006. The panels show high correlation between the estimated and the measured results.

31 days from 1 January 2006 and the bottom panels are for the 31 days following 1 June 2006. The dashed red lines are for the measured irradiance and the solid blue lines denote the estimated irradiance. The GHI from Burns is shown in Fig. 5, the DNI from Hermiston in Fig. 6, and the DIF from Elizabeth in Fig. 7. The time series are displayed to give an absolute comparison between the estimation behavior and the actual measurements.

It can be seen in Figs. 5–7 that the estimated irradiance performs better in summer than in winter, on average. Another salient feature that all three irradiance species have in common is that the estimates appear to be slightly smoother than the measurements, but retain the general shape throughout the 31-day period (which continues over the entire 3-yr period evaluated). The GHI time series in Fig. 5 show a very close match between the model and the observations through time. The estimates are generally slightly below the measurements, as was seen in the MBE. The features of variability are captured in the GHI estimate, albeit the results are smoothed. The June time period is more accurate than the January time period, which is important, because the purpose of the irradiance dataset is to supply a solar PV model for power output, and summertime is more sensitive to errors (as the electric load is highest and so is the cost of electricity).

In Fig. 6, it can be seen that the DNI is much harder to estimate. The estimated DNI is almost always lower than the observed in winter and higher in summer. The smoothness of the estimation versus the measurements is most apparent in these panels, simply because the DNI is much more prone to variability than GHI and DIF. The estimated DNI is accurate with the overall trend for a specific day; for example, day 11 in the bottom panel of Fig. 6 shows the estimation including the extreme reduction in the DNI after clear skies, and then the rapid increase after the sky clears again before the end of the day (although the increase was at an earlier time). Finally, Fig. 7 shows how the DIF estimate can be very accurate for some time periods (e.g., day 15 onward). It can be seen in Fig. 7 that there are high values of DIF in the measurements from days 1 to 6. In trying to explain this trend, the author found that the Elizabeth site had a poor quality of data for the time period being evaluated. The problem was only discovered after the analysis was carried out, and it is shown in the results to illustrate that there are two sources of error for a regression model such as the current proposed one: measurement error and model error. The data log for the Elizabeth site can be found online (http://rredc.nrel.gov/solar/new_data/confrrm/ec/).

Figure 8 displays the MBE (top panel) and RMSE (bottom panel) results for the seven independent

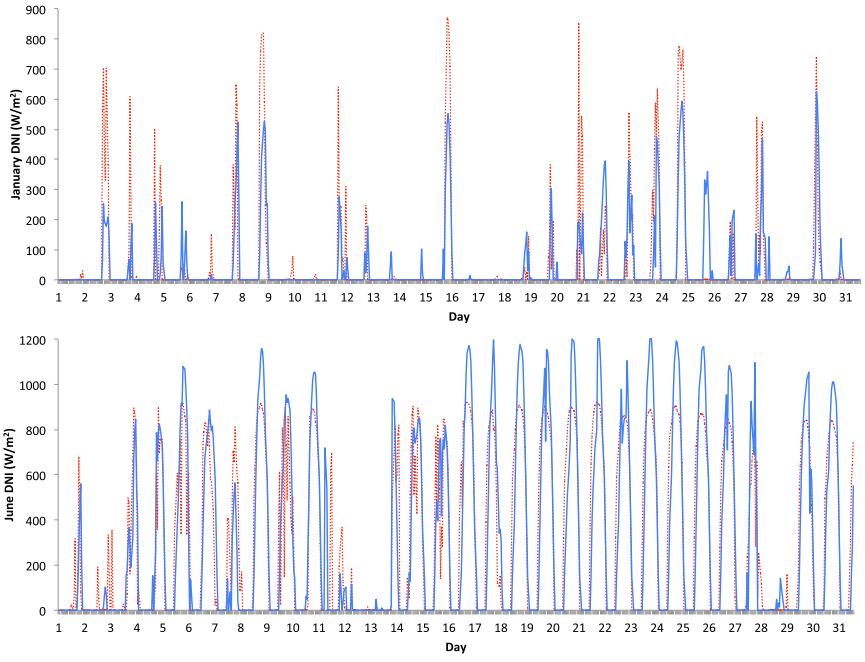

FIG. 6. As in Fig. 5, but for DNI at Hermiston.

verification sites and the two initial verification sites from SURFRAD and ISIS. The metrics are for the complete solar irradiance model. It can be seen that each site has a different value, illustrating the different levels of performance at each geographic location. The GHI estimates perform, on average, as well as they did for the training sites. DNI and DIF are slightly worse in terms of MBE and RMSE than they were at the training sites.

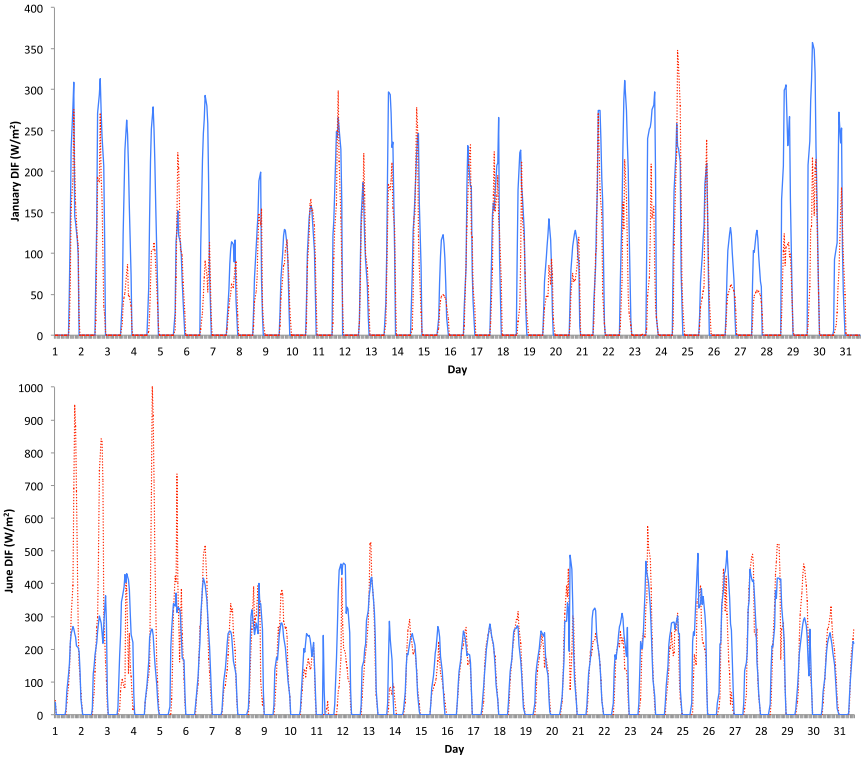

FIG. 7. As in Fig. 5, but for DIF at Elizabeth.

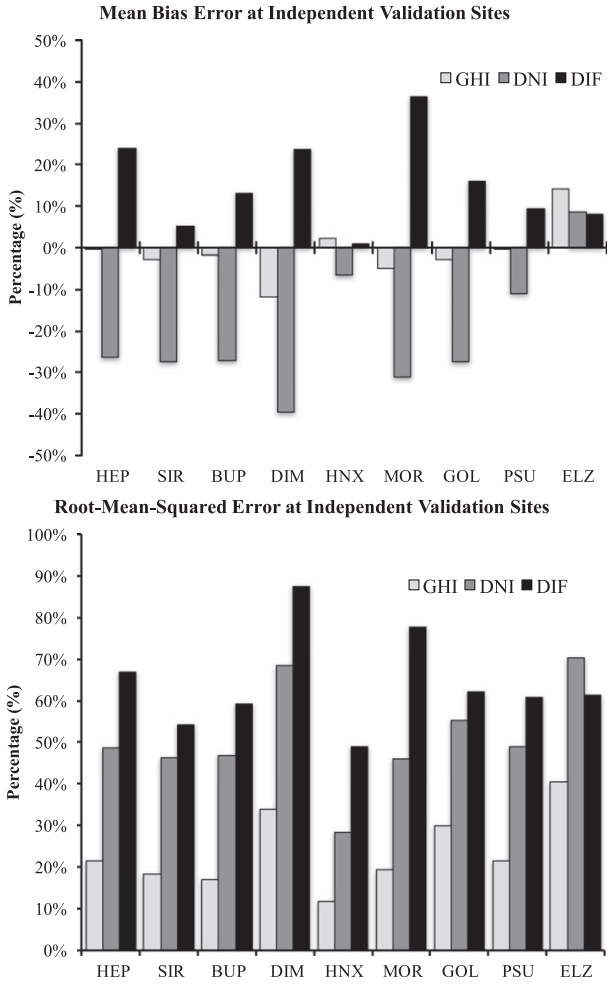

FIG. 8. (top) MBE and (bottom) RMSE for the seven independent verification sites and the initial verification sites. Light gray is for GHI, dark gray is for DNI, and black is for DIF.

Overall, there is a reduction in the accuracy of the regression technique away from the training sites, which is to be expected. Some of the reduction seen, relative to the training sites, is due to full datasets being analyzed, as can be observed by reading the value for the ISIS (HNX) and SURFRAD (PSU) sites and comparing with the initial verification in Table 5, again highlighting the importance of being able to obtain all of the possible measurements. The most important feature from Fig. 8 is that the regression technique created here performs with the same order of accuracy as other available techniques (e.g., Vignola et al. 2012), with the added benefit of being created specifically to be temporally aligned with other datasets on the same spatial grid so that they can be applied to electric power modeling seamlessly. The technique was verified against the State University of New York (SUNY) dataset provided by NREL (http://maps.nrel.gov/prospector) for time periods that overlapped

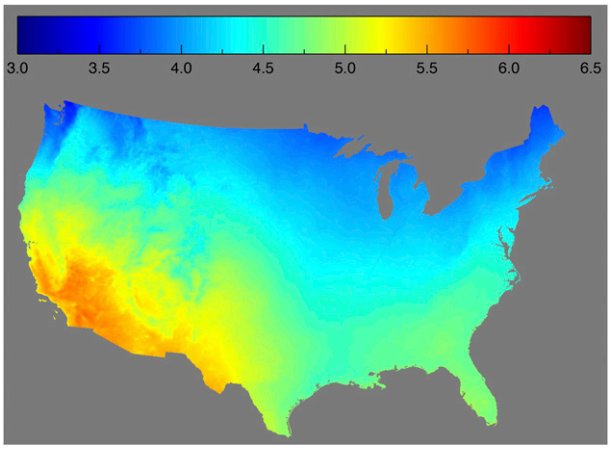

FIG. 9. The average estimated GHI ($kW\,h\,m^{-2}\,day^{-1}$) for the contiguous United States over the 3-yr period of 2006–08. The Southwest has the greatest solar resource while the Northwest and East have the least. All boundaries have been removed to display the detail of the data.

with the one investigated here across a sample of the seven independent sites. It was found that the present regression technique is superior in terms of MBE and RMSE. For example, at the Burns site the current technique has an MBE of −1.64% for GHI, while the SUNY dataset over the same period has an MBE of −2.00%. Similar statistical differences were found with the other irradiance species and at different sites. The differences are not very large, and a review of the SUNY dataset statistics can be found in, for example, Nottrott and Kleissl (2010) and Djebbar et al. (2012). More comparisons need to be made at more sites to establish if indeed the current technique is consistently more accurate.

The linear multivariate multiple regression method has provided estimates of the solar irradiance over the contiguous United States. The dataset is composed of ≈152 000 geographic cells that each contain ≈26 000 hourly data points. Figures 9–11 show the 3-yr averages of GHI, DNI, and DIF over the contiguous United States ($kW\,h\,m^{-2}\,day^{-1}$). To convert from kilowatt hours per meter squared per day to average watts per meter squared, one multiplies the value by 41.695; thus, the range from Fig. 9 is 125–271 $W\,m^{-2}$. Figure 9 shows that the southwest is the best resource site in terms of GHI, which is very important for solar PV. All three maps show that the extreme Northwest and Northeast are very poor sources in terms of irradiance. The maps are consistent with other datasets, but cover a longer time period and wider geographic area with no blending of different datasets. Figure 10 is interesting because DNI, which is very important for concentrated solar power (CSP), is shown to be best as a resource at locations in the extreme Southwest. The map in Fig. 11 shows how clear

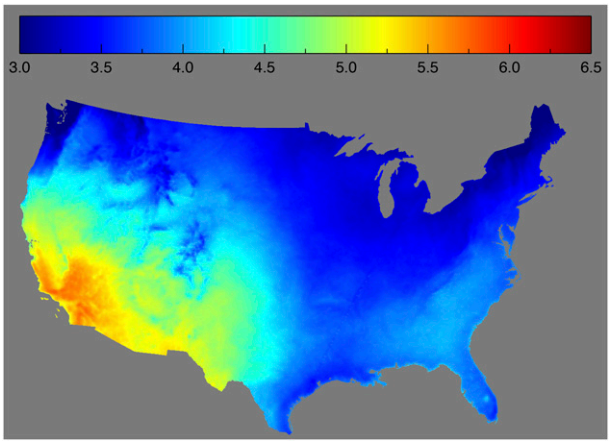

FIG. 10. As in Fig. 9, but for the DNI. The Southwest is the best solar resource area whereas the rest of the United States is much poorer.

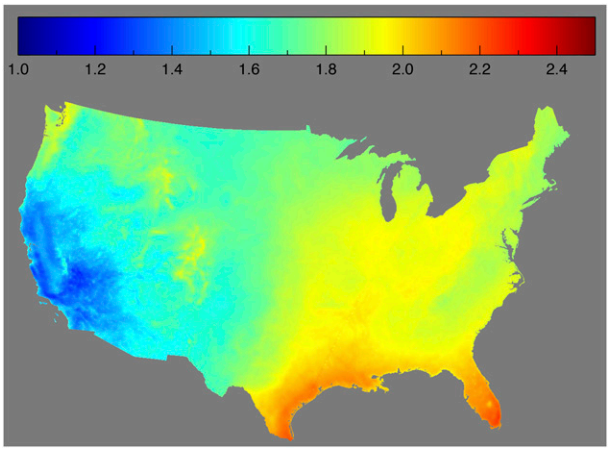

FIG. 11. As in Fig. 9, but for the DIF (the range is different than in Figs. 9 and 10). The Gulf Coast has the most DIF resource, the Southwest has the least DIF, and in general the East has more DIF than the West.

the skies are over the desert Southwest, and how the Gulf Coast region is dominated by large amounts of DIF versus DNI, which means it would be suitable for solar PV (as GHI is a relatively good resource there), but not as suitable for CSP. Note that the scale has changed in Fig. 11. Figures 9–11 illustrate the detail within the dataset, but they are averages of the whole 3-yr period. The true value of the dataset is in the spatial and temporal resolution, which is used in section 4 to model solar PV power output at all the sites across the contiguous United States. The dataset will be used in future research to model CSP power output over the contiguous United States and in detailed electric power system modeling.

4. Solar photovoltaic power estimates

In this section, the contiguous U.S. regression-derived solar irradiance estimates are applied to a power output algorithm for a specific solar PV configuration. The formulation of the power model will be briefly outlined and a resource assessment for a specific configuration will be shown at the end.

To compute the solar photovoltaic power output, the total, direct, and diffuse solar irradiance estimates from section 3 were inserted into Eqs. (11)–(20) from King et al. (2004). In making the power estimates, a standard solar panel for the year of 2007 taken from the NREL System Advisor Model (SAM) version 2012.5.11 (https://sam.nrel.gov/) is assumed, namely the SunPower SPR-315E-WHT. It was assumed that the panels would be mounted on a single-axis tracker and would be orientated north to south while being tilted at latitude This results in the angle of incidence on the panels at all times of the day being the declination angle of the sun (Masters 2004). The generic constants used by the power generation algorithm were obtained from De Soto et al. (2006). The panel-specific constants were taken from the NREL SAM.

An important feature of solar PV panels is that the temperature of the cell greatly influences the power production potential. This effect is dealt with by computing the back of the module temperature using both the 10-m wind speeds and the 2-m ambient air temperature from the RUC assimilation model. There is no knowledge in the model of snow or ice covering the panels. Additionally, the panels are assumed to be placed far enough apart as to not create shadowing effects on neighboring panels.

The mathematical formulas for the algorithm of power production are all contained within King et al. (2004). An outline the major parts of the algorithm is described. First, one imports the solar irradiance estimates (GHI, DNI, DIF, and solar zenith angle) along with the meteorological data (wind speed at 10 m and temperature at 2 m). Second, we compute the cell temperature and the angle of incidence of the solar irradiance on the tilted and tracked panel. Third, we calculate the power falling onto the panel from the irradiance fields. Fourth, the current and voltages within the panel are approximated [the equations in King et al. (2004) and NREL SAM are empirically derived]. Finally, the current and voltage are combined to calculate the power for the panel. There are equations within the algorithm, which are based on NREL SAM, that compute the derating due to the panel structure and material. The output of the panel is restricted to 115% of the nameplate capacity. After the algorithm has finished, a post-processing derate factor of 95% is applied to estimate downtimes and other deficiencies such as inverter losses and bad wiring connections. The algorithm performs the

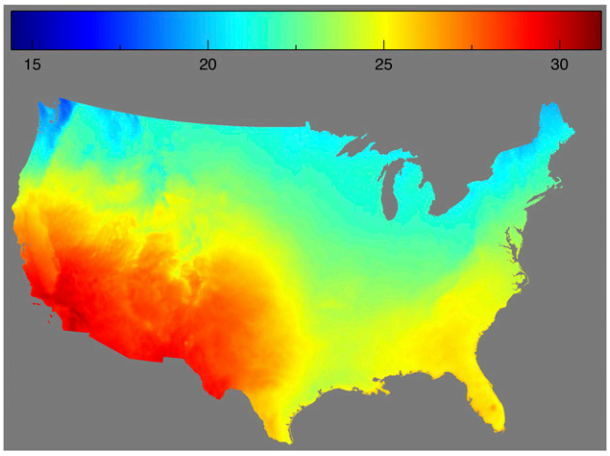

FIG. 12. The solar PV capacity factor map for the contiguous United States. The scale is from 14% to 33%. The capacity factor is for the individual panels described and tilted at latitude, tracking along one axis. Other solar PV panels will perform differently. One can see that the dynamic range over the United States is not large. The Pacific Northwest is particularly poor, and the Southwest particularly good.

process at every location within the domain at each time step and outputs the power estimate into a dataset.

Once the solar PV power estimate algorithm is finished, the average capacity factors were computed for the continental United States for the 3-yr period of 2006–08. The capacity factor maps show what a hypothetical solar PV plant made of SunPower SPR-315E-WHT panels would create as an average of the rated capacity in that model grid cell. For example, if the capacity factor in a grid cell was 10%, that means on average over the whole time period the solar PV plant will generate 10% of its rated capacity multiplied by the number of hours running. The efficiency of the panels chosen is 19.3%, which means it can turn 19.3% of the solar irradiance into electricity in optimal conditions. The whole power estimate algorithm can be altered, with a few constants, to produce similar datasets for different panels and different configurations of tilt, orientation, and tracking.

Figure 12 displays the capacity factor maps for the continental United States. The scale has a range of 14%–33%. Figure 12 shows that the Southwest region of the United States is the absolute best resource, but the structure is far from simple. The Southeast has great potential, particularly around Lake Okeechobee in Florida. The mountainous regions in Colorado have poorer resources along the Front Range, as a result of summertime clouds over the higher terrain. The Seattle area is particularly poor as a resource. The extreme southwest of California has the highest capacity factors, which is in agreement with the climatological data. What is striking is that the capacity factor map is not the same as any of the GHI, DNI, or DIF maps (Figs. 9–11), and that is because the capacity factor takes all three into accounts, as well as the temperature in the local area. A similar map for CSP, for example, would be expected to look to be well correlated to the DNI resource map because of its almost total reliance on that specific resource.

5. Discussion and conclusions

The present paper has provided a novel technique for obtaining solar irradiance species including direct normal and diffuse horizontal. The underlying engine for the procedure is a linear multiple multivariate regression scheme trained upon numerical weather prediction (NWP) assimilation model hydrometeors, satellite measurements where available, calculated top-of-the-atmosphere solar irradiance, and ground-based, high-quality, solar measurements. The choice of regressors is important, and in the present paper care was taken to choose, when possible, the best combination of model parameters to improve the solar irradiance. The solar irradiance estimates were processed through a solar PV power output algorithm to obtain a solar PV capacity factor resource map for the continental United States.

The method was verified against independent sites that were not in the training of the regression. The verification showed that the regression produced estimates that are representative of independent sites. An additional set of verification sites was acted upon when the full suite of regressions was applied (due to different satellite data available at different time steps). The results of the verification can be seen in Fig. 8, which shows that the use of the mixed regressions was less accurate than with all the data, but was consistent over the sites. The model performs as well as other current satellite models (Vignola et al. 2007).

The results from irradiance modeling indicate that the technique has a bias that could be due to the ground-based measurements, the weather data bias, or even the parallax effect from the satellite data in the regressions. The power of the regression procedure can be seen most clearly in Figs. 5–7, where the comparisons for GHI, DNI, and DIF for a summer and a winter period can be seen. There is a tendency for a negative bias in the procedure, but the estimates reproduced some difficult to handle features, such as rapid changes in irradiance, scattered cloud irradiance patterns, and morning fog events. In addition, since the datasets include almost every hour of the time periods, more analysis can be performed to investigate seasonal and geographic variations.

The resource maps of GHI, DNI, and DIF, along with the capacity factor maps, illustrate the best and worst resource sites. The accuracy of the data and the time interval over which the regression model was trained give the images some credibility. There are still going to be

errors in the model. Future work will be aimed at increasing the resolution of the weather data to 3 km, incorporating more satellite data, computing the training over longer time periods, and assimilating more ground-based observations to include more climate regimes. Further future work will be to include path integral calculations of attenuation that will take into account neighboring cell properties. In an effort to determine if a saturated training set was produced, regressions were performed for the contiguous United States repeatedly to train the regression scheme and to see if there was an improvement. Each time a new site and more data were added, the overall training set performance improved; however, some specific sites were made worse. In particular, when all the verification sites were included in the training set and the regression was performed, the estimates improved substantially. However, those results were not used because no sites would be left to validate against. The adjusted correlation coefficient for GHI at each site remained around 92%, the RMSE and CV decreased to around 17%–19%, and the MBE was 1%–2%. Future work will incorporate many more training and validation sites over a wide geographic region.

The entire dataset that was created for the present paper is available online (esrl.noaa.gov/gsd/renewable/newsresults/usstudy/Weather_Inputs/). The files also contain the spatially and temporally aligned wind dataset (Clack et al. 2016). The wind and solar PV power estimates from these datasets were utilized in studies of the U.S. electric grid (Clack et al. 2015; MacDonald and Clack et al. 2016).

Acknowledgments. The author thanks A. Alexander, J. Wilczak, and A. Sitler for their helpful recommendations for the study that is described in this paper.

REFERENCES

Augustine, J. A., G. B. Hodges, C. R. Cornwall, J. J. Michalsky, and C. I. Medina, 2005: An update on SURFRAD—The GCOS Surface Radiation Budget Network for the continental United States. *J. Atmos. Oceanic Technol.*, **22**, 1460–1472, doi:10.1175/JTECH1806.1.

Clack, C. T. M., Y. Xie, and A. E. MacDonald, 2015: Linear programming techniques for developing an optimal electrical system including high-voltage direct-current transmission and storage. *Int. J. Electr. Power Energy Syst.*, **68**, 103–114, doi:10.1016/j.ijepes.2014.12.049.

——, A. Alexander, A. Choukulkar, and A. E. MacDonald, 2016: Demonstrating the effect of vertical and directional shear for resource mapping of wind power. *Wind Energy*, **19**, 1687–1697, doi:10.1002/we.1944.

Deshmukh, M. K., and S. S. Deshmukh, 2008: Modeling of hybrid renewable energy systems. *Renewable Sustainable Energy Rev.*, **12**, 235–249, doi:10.1016/j.rser.2006.07.011.

De Soto, W., S. A. Klein, and W. A. Beckman, 2006: Improvement and validation of a model for photovoltaic array performance. *Sol. Energy*, **80**, 78–88, doi:10.1016/j.solener.2005.06.010.

Djebbar, R., R. Morris, D. Thevenard, R. Perez, and J. Schlemmer, 2012: Assessment of SUNY version 3 global horizontal and direct normal solar irradiance in Canada. *Energy Procedia*, **30**, 1274–1283, doi:10.1016/j.egypro.2012.11.140.

Dominguez-Ramos, A., M. Held, R. Aldaco, M. Fischer, and A. Irabiena, 2010: Carbon footprint assessment of photovoltaic modules manufacture scenario. *Proc. ESCAPE-20: European Symp. on Computer Aided Process Engineering*, Ischia, Naples, Italy, Italian Association of Chemical Engineering. [Available online at http://www.aidic.it/escape20/webpapers/305Dominguez-Ramos.pdf.]

Geuder, N., F. Trieb, C. Schillings, R. Meyer, and V. Quaschning, 2003: Comparison of different methods for measuring solar irradiation data. *Third Int. Conf. on Experiences with Automatic Weather Stations*, Torremolinos, Spain, Instituto de Nacional Meteorologia. [Available online at http://www.dlr.de/tt/Portaldata/41/Resources/dokumente/institut/system/publications/Automatic_Weather_Stations_2003_NGeuder.pdf.]

Hammer, A., D. Heinemann, E. Lorenz, and B. Lockehe, 1999: Short-term forecasting of solar radiation: A statistical approach using satellite data. *Sol. Energy*, **67**, 139–150, doi:10.1016/S0038-092X(00)00038-4.

Hicks, B. B., J. J. DeLuisi, and D. R. Matt, 1996: The NOAA Integrated Surface Irradiance Study (ISIS): A new surface radiation monitoring program. *Bull. Amer. Meteor. Soc.*, **77**, 2857–2864, doi:10.1175/1520-0477(1996)077<2857:TNISIS>2.0.CO;2.

Houborg, R., H. Soegaard, W. Emmerich, and S. Moran, 2007: Inferences of all-sky solar irradiance using Terra and Aqua MODIS satellite data. *Int. J. Remote Sens.*, **28**, 4509–4535, doi:10.1080/01431160701241902.

Huang, C., M. Huang, and C. Chen, 2013: A novel power output model for photovoltaic systems. *Int. J. Smart Grid Clean Energy*, **2**, 139–147, doi:10.12720/sgce.2.2.139-147.

King, D. L., S. Gonzalez, G. M. Galbraith, and W. E. Boyson, 2004: Performance model for grid-connected photovoltaic inverters. Sandia National Laboratories Tech. Rep., 41 pp. [Available online at http://energy.sandia.gov/wp/wp-content/gallery/uploads/043535.pdf.]

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

Kratzenberg, M. G., S. Colle, and H. G. Beyer, 2008: Solar radiation prediction based on the combination of a numerical weather prediction model and a time series prediction model. *Proc. First Int. Congress on Heating, Cooling, and Buildings: EuroSun 2008*, Lisbon, Portugal, Int. Solar Energy Society, 12 pp.

Lueken, C., G. E. Cohen, and J. Apt, 2012: Costs of solar and wind power variability for reducing CO_2 emissions. *Environ. Sci. Technol.*, **46**, 9761–9767, doi:10.1021/es204392a.

MacDonald, A. E., C. T. M. Clack, A. Alexander, A. Dunbar, J. Wilczak, and Y. Xie, 2016: Future cost-competitive electricity systems and their impact on US CO_2 emissions. *Nat. Climate Change*, **6**, 526–531, doi:10.1038/nclimate2921.

Masters, G. M., 2004: *Renewable and Efficient Electric Power Systems.* John Wiley and Sons, 680 pp.

Mathiesen, P., and J. Kleissl, 2011: Evaluation of numerical weather prediction for intra-day solar forecasting in the continental United States. *Sol. Energy*, **85**, 967–977, doi:10.1016/j.solener.2011.02.013.

——, C. Collier, and J. Kleissl, 2013: A high-resolution, cloud-assimilating numerical weather prediction model for solar irradiance forecasting. *Sol. Energy*, **92**, 47–61, doi:10.1016/j.solener.2013.02.018.

Michalsky, J. J., and Coauthors, 2003: Results from the first ARM diffuse horizontal shortwave irradiance comparison. *J. Geophys. Res.*, **108**, 4108, doi:10.1029/2002JD002906.

Mills, A., and R. Wiser, 2010: Implications of wide-area geographic diversity for short term variability of solar power. Lawrence Berkley National Laboratory Tech. Rep., 22 pp. [Available online at https://emp.lbl.gov/sites/all/files/presentation-lbnl-3884e-ppt.pdf.]

Myers, D. R., 2005: Solar radiation modeling and measurements for renewable energy applications: Data and model quality. *Energy*, **30**, 1517–1531, doi:10.1016/j.energy.2004.04.034.

Nottrott, A., and J. Kleissl, 2010: Validation of the NSRDB–SUNY global horizontal irradiance in California. *Sol. Energy*, **84**, 1816–1827, doi:10.1016/j.solener.2010.07.006.

Parida, B., S. Iniyanb, and R. Goic, 2011: A review of solar photovoltaic technologies. *Renewable Sustainable Energy Rev.*, **15**, 1625–1636, doi:10.1016/j.rser.2010.11.032.

Paulescu, M., E. Paulescu, P. Gravila, and V. Badescu, 2013: *Weather Modeling and Forecasting of PV Systems Operation.* Springer, 358 pp., doi:10.1007/978-1-4471-4649-0.

Pearson, K., 1908: On the generalized probable error in multiple normal correlation. *Biometrika*, **6**, 59–68, doi:10.1093/biomet/6.1.59.

Perez, R., and Coauthors, 2013: Comparison of numerical weather prediction solar irradiance forecasts in the US, Canada and Europe. *Solar Energy*, **94**, 305–326, doi:10.1016/j.solener.2013.05.005.

Solanki, C. S., 2009: *Solar Photovoltaics: Fundamentals, Technologies and Applications*. PHI Learning, 478 pp.

Spencer, J. W., 1971: Fourier series representation of the position of the sun. *Search*, **2**, 172. [Available online at http://www.mail-archive.com/sundial@uni-koeln.de/msg01050.html.]

Stanton, J. M., 2001: Galton, Pearson, and the peas: A brief history of linear regression for statistics instructors. *J. Stat. Educ.*, **9**, 3. [Available online at https://ww2.amstat.org/publications/jse/v9n3/stanton.html.]

Theil, H., 1961: *Economic Forecasts and Policy.* Contributions to Economic Analysis, Vol. 15, North-Holland, 567 pp.

Vignola, F., and R. Perez, 2004: Solar resource GIS data base for the Pacific Northwest using satellite data: Final report. Solar Radiation Monitoring Laboratory Tech. Rep., University of Oregon, 73 pp. [Available online at http://solardat.uoregon.edu/download/Papers/SolarResourceGISDataBaseforthePacificNorthwestusingSatelliteData–FinalReport.pdf.]

——, P. Harlan, R. Perez, and M. Kmiecik, 2007: Analysis of satellite derived beam and global solar radiation data. *Sol. Energy*, **81**, 768–772, doi:10.1016/j.solener.2006.10.003.

——, J. Michalsky, and T. Stoffel, 2012: *Solar and Infrared Radiation Measurements.* CRC Press, 410 pp.

Wang, K., J. Augustine, and R. E. Dickinson, 2012: Critical assessment of surface incident solar radiation observations collected by SURFRAD, USCRN and AmeriFlux networks from 1995 to 2011. *J. Geophys. Res.*, **117**, D23105, doi:10.1029/2012JD017945.

Wang, Z., F. Wang, and S. Su, 2011: Solar irradiance short-term prediction model based on BP neural network. *Energy Procedia*, **12**, 488–494, doi:10.1016/j.egypro.2011.10.065.

Wong, L., and W. Chow, 2001: Solar radiation model. *Appl. Energy*, **69**, 191–224, doi:10.1016/S0306-2619(01)00012-5.

Zhou, W., H. Yang, and Z. Fang, 2007: A novel model for photovoltaic array performance prediction. *Appl. Energy*, **84**, 1187–1198, doi:10.1016/j.apenergy.2007.04.006.