Solar Energy 176 (2018) 370–379

Contents lists available at ScienceDirect

Solar Energy

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

A stochastic downscaling approach for generating high-frequency solar irradiance scenarios

Wenqi Zhang[a,*], William Kleiber[a], Anthony R. Florita[b], Bri-Mathias Hodge[b], Barry Mather[b]

[a] Applied Mathematics Department, University of Colorado at Boulder, Boulder, CO 80309, United States
[b] Power Systems Engineering Center, National Renewable Energy Laboratory, Golden, CO 80401, United States

ARTICLE INFO

Keywords:
Irradiance modeling
High-resolution
Stochastic
Non-Gaussian

ABSTRACT

Solar power is increasingly cost viable with solar photovoltaic (PV) installations becoming commonplace. PV planning and operational studies, however, require high-frequency solar irradiance scenarios to understand potential electric grid impacts due to the variability and uncertainty of the underlying solar resource. Existing remote sensing solar data products are often available over large spatial domains, but are limited in temporal resolution. For example, the global horizontal irradiance (GHI) component contained within the National Solar Radiation Database (NSRDB) is available at time resolution of 30 min on an approximately four-kilometer grid. In contrast, substantial solar variability is present at finer time scales and this article describes an algorithm to stochastically generate one-minute GHI from widely available sub-hourly NSRDB. A generalized linear modeling (GLM) framework is proposed, which includes non-Gaussian mixtures, and extends the literature involving the synthesis of GHI data. The model is trained on a set of sample locations around Oregon, USA, and validated across the USA using both the Surface Radiation Budget Network (SURFRAD) dataset and Solar Radiation Monitoring Laboratory (SRML) network. Simulated ensembles show good coverage properties and temporal correlation structure. The resulting downscaled ensembles allow for understanding the unpredictable variability inherent in GHI at locations without direct measurements. Future work can leverage the algorithm as part of a stochastic optimization of electric grid operations with high-penetration PV systems.

1. Introduction

1.1. Motivation

Over 15% of total global electricity was generated from renewable energy in 2015 according to the REN21 (2016) renewable global report. In comparison to other renewables, solar photovoltaic (PV) capacity was relatively small; however, installations have increased substantially over the past few years (Cameron et al., 2017). Since solar PV is inherently diurnal, variable, and uncertain, its presence in both the bulk system power grid and as distributed rooftop installations makes reliable power systems operation more challenging (Lew et al., 2013). In particular, solar fluctuations over a short time period, especially in a high PV penetration scenario, may lead to inefficient and negative impacts on grid operations (Nguyen et al., 2016; Bright et al., 2017).

To avoid undesirable PV deployment scenarios, or to determine necessary changes in distribution system components (e.g. distribution line gauge and automatic voltage regulation equipment control settings), studies are typically completed for large interconnected system or many small systems once the aggregation hits a certain threshold. Traditionally, these PV integration studies were relatively simple and included power flow calculations at a limited number of salient operating points (Wang et al., 2017). The underlying data for the interconnected PV systems was also simple: all PV systems were typically fully on or off depending on which analyses were being completed (Lave et al., 2017). Modern analyses must consider hundreds and even thousands of individual PV systems per distribution circuit, and it is vital to increase the fidelity of PV interconnection studies. Quasi-static time-series (QSTS) analysis of distribution systems has thus been proposed as a method to improve PV interconnection study accuracy in scenarios with many interconnected PV systems (Reno et al., 2017). One of the challenges of QSTS distribution system analysis is the inherent need for accurate high temporal-resolution load and PV resource data (Diagne et al., 2013; Hansen et al., 2010).

Accurate solar PV power modeling not only plays an important role to ensure solar integration is economically feasible but also contributes to improvements in power quality (Perez et al., 2018). However, solar irradiance modeling is challenged by small scale cloud fronts and other

* Corresponding author.
E-mail address: wezh4530@colorado.edu (W. Zhang).

https://doi.org/10.1016/j.solener.2018.10.019
Received 2 August 2018; Received in revised form 28 September 2018; Accepted 5 October 2018

atmospheric conditions such as wind speed, temperature, and relative humidity. Uncertainty related to these conditions at the site-level is substantial (Lave and Kleissl, 2010; Perez et al., 2011; Schmalensee, 2015), and short-term weather volatility can seriously affect power generation. In Sayeef et al. (2012), solar intermittency was described as a fundamental barrier to the assimilation of large-scale solar power around the world. Leveraging in situ data is desirable, but these datasets are often plagued with missing data and inconsistent measurement techniques (Kumar et al., 2014). Methods to synthetically generate time series that honor observational statistical characteristics are thus useful for PV planning studies.

1.2. Current state of the literature

The literature on solar irradiance modeling for electric grid applications generally focuses on two major aspects: (1) forecasting and (2) resource assessment. In the past few years, solar irradiance forecasting models have become well-established (Wang et al., 2011; Baharin et al., 2014; Melzi et al., 2016; Jamaly and Kleissl, 2017). This work focuses on the latter problem, solar irradiance variability estimation and assessment.

When considering a individual site, there have been multiple approaches to synthetically generate irradiance time series. Lave et al. (2013) utilized a wavelet-based model to simulate the power output of a single solar PV plant. Morf (2013) developed a simulation method where irradiance was separated into a deterministic and stochastic component; Ångström-Prescott regression was used to generate the deterministic component while the stochastic component was derived from a cloud cover index. Ngoko et al. (2014) started with a clearness index and generated irradiance data using Markov models trained on empirical observations. Larrañeta et al. (2015) used an improved downscaling method from Polo et al. (2011) to generate direct normal irradiance time series at 10-min resolution. A Markov transition matrix was used to synthetically generate one-minute GHI with mean hourly meteorological observation inputs by Bright et al. (2015); however, to accurately determine cloud cover index, a minimum of 10 years of observational data were recommended for model training. Bright et al. (2017) improved the model by adding spatial variation in the synthetic generation; i.e., the model was then geographically flexible without the need of local irradiance data inputs. Larrañeta et al. (2018) introduced two approaches with different input accuracy requirements to downscale direct normal irradiance (DNI) time series. There are other downscaling methods including resampling and clustering techniques (Fernández-Peruchena et al., 2015; Fernández-Peruchena and Gastón, 2016; Peruchena et al., 2018). As observational data are geographically sparse, statistical space–time characteristics are of interest (Hoff and Perez, 2010; Lave et al., 2012; Perez and Fthenakis, 2015). Lave et al. (2012) suggested that solar variability cannot be assumed to be purely isotropic, and Perez and Fthenakis (2015) explored the spatial anisotropic nature of solar irradiance using cloud motion. Dise et al. (2013) and Perez et al. (2015) stochastically downscaled the National Solar Radiation Database (NSRDB) (Sengupta et al., 2018), which includes satellite derived estimates.

1.3. Novelty and structure

To contrast the proposed work with the existing literature, the closest approach is likely Grantham et al. (2018). These authors extended the method of Boland (2008), who develop classical stationary time series techniques for simulating daily and hourly GHI. Grantham et al. (2018)'s extension stochastically generates GHI using a nonparametric bootstrapping technique. However, as a resampling procedure, the proposed model of Grantham et al. (2018) requires a large training database, is only available at locations with high resolution data and is limited in its ability to generate noise that is not present in the training sample. Our proposal is a semiparametric model that combines both parametric distributions, which generate stochasticity, and nonparametric estimators of temporal correlation. Compared to Grantham et al. (2018), our model is less complex, requires less input data and is more computationally efficient. Specifically, crucial autocorrelation properties are maintained while low-order statistics such as means and variances are accurately reproduced. Moreover, our method results in simulations that are consistent with the popular NSRDB. In particular, we introduce a method for temporally downscaling GHI from NSRDB's 30-min resolution to one-minute resolution. NSRDB contains high quality estimates of historical solar irradiance based on processed satellite data, and is available at an approximately four km^2 resolution every 30 min over a large portion of North America. However, even high quality estimates every 30 min do not adequately represent or capture the variability of solar irradiance. For example, Fig. 1 shows linearly interpolated GHI from NSRDB along with in situ one-minute pyranometer measurements at a station in Eugene, OR. Generally the satellite data product follows the trend of directly measured values. However, there is substantial high frequency variability in the direct measurements that is not apparent from the satellite-based product and directly affect solar power generation. Locations without ground measurements require spatial prediction from those available elsewhere. Energy applications based on the proposed methodology may include small-scale PV supply and demand where the spatial element is not integral, for example, high-resolution irradiance inputs are required to calculate PV efficiencies based on electricity flows within residences (Darcovich et al., 2015).

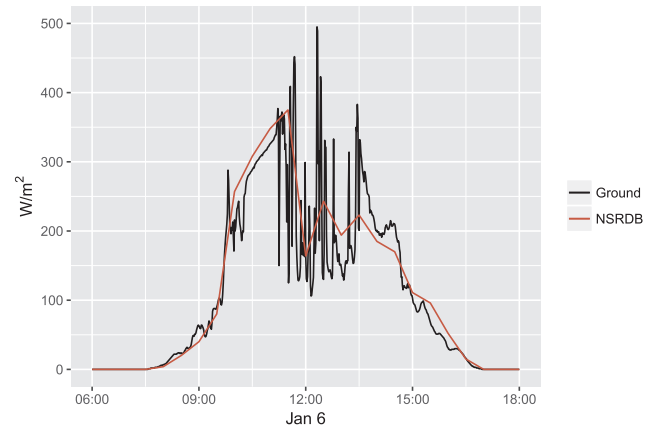

Fig. 1. Interpolated NSRDB data and in situ data for Eugene, OR, January 6, 2013.

A distinct aspect of the proposed methodology is the use of a generalized linear model (GLM) framework and properly transformed mixtures of random variables to capture the stochastic variability of high frequency irradiance. Furthermore, the first stage of the model distinguishes between clear and unclear days. A crucial element of our proposed downscaling process is allowing for seasonally nonstationary temporal correlation, yielding greater accuracy in capturing climatological variability.

The proposed method is validated at a number of locations with different climatic conditions throughout the USA. Generally, the downscaled ensembles exhibit proper statistical coverage properties, desired variability, and maintain critical temporal autocorrelation. The paper is structured as follows: Section 2 presents the in situ database used in this work. Section 3 describes the methodology proposed for generating stochastic one-minute GHI data based on the NSRDB. Validation based on a comprehensive set of statistical metrics are discussed in Section 4 before the work is concluded and future work suggested in Section 5.

Table 1
Data used to train the model.

Database and available series	Time span	Observations
Half-hourly NSRDB data sets		
Eugene, Oregon	January 1, 2013 to December 31, 2013	17,520
Global horizontal irradiance		
One-minute pyranometer data sets		
Eugene, Oregon	January 1, 2013 to December 31, 2013	525,600

2. Data

Two in situ observational time series datasets in the Pacific Northwest are used to train the model (see Table 1). The data consists of 30-min NSRDB estimates of GHI (in W/m^2) as well as direct pyranometer measurements of one-minute GHI at Eugene from the University of Oregon's Solar Radiation Monitoring Laboratory (SRML; (Vignola and Perez, 2004)). The method is additionally validated on two other locations in the SRML and a host of stations from the Surface Radiation Budget Network (SURFRAD) in 2013 (see Table 2), a network of high quality observation stations supported by NOAA's Office of Global Programs (Augustine et al., 2000). These locations offer climatically variable weather situations across the continental USA. The SRML and SURFRAD data are available at one-minute resolution.

Although the NSRBD contains satellite estimates of local irradiance, inherent biases and inconsistencies arise. There is a timing discrepancy on clear days between clear sky GHI and the ground measurements due to the spatial coarseness of the NSRDB product, which is not point source. To resolve this problem, a preprocessing step is necessary to match the sunrise time for each day. These satellites can measure cloud over at local scales, but solar irradiance at the site level is not available in most locations, and thus statistical models or parameterization schemes are required for generating irradiance scenarios (Clifton et al., 2018).

3. Methodology

Our proposed approach to stochastically downscaling NSRDB from the 30-min resolution to one-minute data on any particular day involves an initial decision tree determining what type of day to simulate. Intuitively, if atmospheric conditions are clear, then very little deviation from clear sky irradiance is expected, and thus those days are modeled separately. Otherwise, if substantial cloudiness or aerosols are present then a stochastic model is developed particular to those conditions. It is possible to consider sub-daily decision rules for "clear hours", but our results suggest that such complications are not needed (see the following section). The following subsection discusses the initial decision tree splitting rule.

Table 2
Model validation locations.

Site	Source	Climate
Portland, Oregon	NREL-SRML	Marine West Coast
Salem, Oregon	NREL-SRML	Mediterranean
Bondville, Illinois	NOAA-SURFRAD	Humid Continental
Boulder, Colorado	NOAA-SURFRAD	Semi-arid
Desert Rock, Nevada	NOAA-SURFRAD	Desert
Fort Peck, Montana	NOAA-SURFRAD	Semi-arid
Goodwin Creek, Mississippi	NOAA-SURFRAD	Humid Subtropical
Penn State University, Pennsylvania	NOAA-SURFRAD	Humid Subtropical
Sioux Falls, South Dakota	NOAA-SURFRAD	Humid Subtropical

3.1. Clear day rule

On a clear day, linearly interpolated NSRDB GHI is expected to provide a better estimate of actual GHI than the approach of the next subsection that incorporates stochastic variability. Thus, clear days need to be distinguished based on NSRDB alone. We adopt a clear day decision rule from a previous study (Zhang et al., 2018), described here for completeness, and note that other authors have considered more complicated algorithms (Reno and Hansen, 2016).

For any given day, GHI and clear sky GHI at the 30 min resolution are available from the NSRDB. The following rule is based on the heuristic that on clear days, variability of GHI and clear sky GHI should be similar, while on other days raw GHI will be more variable. Let X_t denote linearly interpolated GHI at time point t, and $X_{C,t}$ denote the clear sky GHI. To quantify variability, we compute first-order differences,

$$\Delta X_t = X_t - X_{t-1}, \tag{1}$$

$$\Delta X_{C,t} = X_{C,t} - X_{C,t-1}, \tag{2}$$

and consider the quantity

$$\gamma = \max_t |\Delta X_t - \Delta X_{C,t}|, \tag{3}$$

where the maximum is taken over a particular day. If this value is smaller than a given tolerance, that day is denoted as a clear day, and otherwise it is not clear. A tolerance for the quantity γ is set to be 0.1, with values less than this indicating a clear day, and values greater indicating a day that relies on the stochastic downscaling model (Zhang et al., 2018). Fig. 2 illustrates the outcome of this decision rule for two candidate days.

3.2. Non-clear days

Days that do not satisfy the clear sky decision rule are modeled as a correlated process that is marginally non-Gaussian, allowing for stochastic downscaling of the 30-min NSRDB data product. Let Y_t denote the in situ pyranometer measurement at time t, and let X_t be the linear interpolation of GHI from the NRSDB. We suppose a model of the form

$$\log(Y_t) = \log(X_t) + \varepsilon_t \tag{4}$$

for some stochastic error ε_t that represents variability due to atmospheric conditions and biases due to the linear interpolation of the NSRDB product. Note that this log-additive model is equivalent to modeling the clear sky index (Bright et al., 2015). Exploratory analysis confirms that the errors are correlated in time, which we model as

$$\varepsilon_t = \sum_{i=1}^{n} \ell_{ti}\omega_i, \tag{5}$$

where $\omega_1, \ldots, \omega_n$ are independent mean zero random variables, and $(\ell_{t1}, \ell_{t2}, \ldots, \ell_{tn})$ is the tth row of the Cholesky factor of a positive-definite covariance matrix. For n time points of interest, the model for the stochastic errors can be succinctly written as

$$\boldsymbol{\varepsilon} = L\boldsymbol{\omega}, \tag{6}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$, $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^{\mathrm{T}}$ and $L = (\ell_{ij})_{i,j=1}^{n}$. The ε_t errors are correlated in time due to being weighted combinations of the same latent ω_t random variables. It is straightforward to show that if L is the Cholesky factor of a positive definite matrix Σ, then $\boldsymbol{\varepsilon}$ has exactly Σ as its covariance matrix. Note that $\boldsymbol{\varepsilon}$ is a vector of correlated random variables, while $\boldsymbol{\omega} = L^{-1}\boldsymbol{\varepsilon}$ are uncorrelated; we refer to $\boldsymbol{\omega}$ as the *decorrelated residuals*.

To motivate the choice of distribution for $\boldsymbol{\omega}$, we show some empirical estimates from Eugene, OR for the year 2013. In particular, for each day we calculate

$$\hat{\boldsymbol{\omega}} = L^{-1}\boldsymbol{\varepsilon}, \tag{7}$$

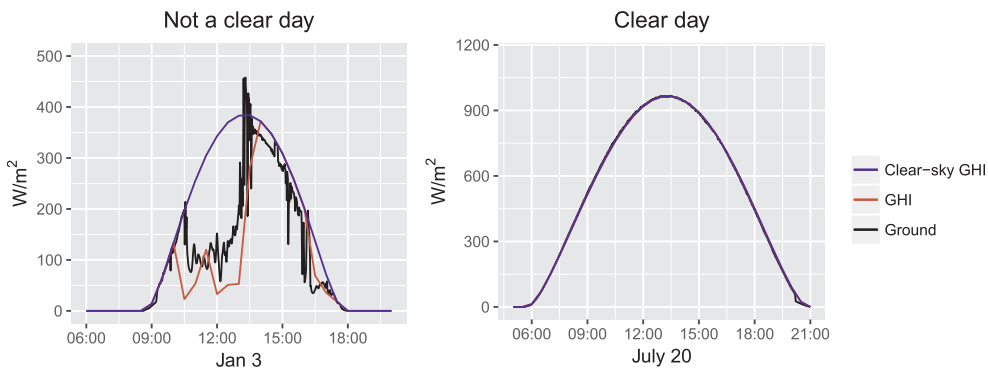

Fig. 2. Interpolated NSRDB and in situ data for Eugene, OR, January 3 and July 20, 2013, respectively. The clear day decision rule classifies January 3 as not clear, and July 20 as clear.

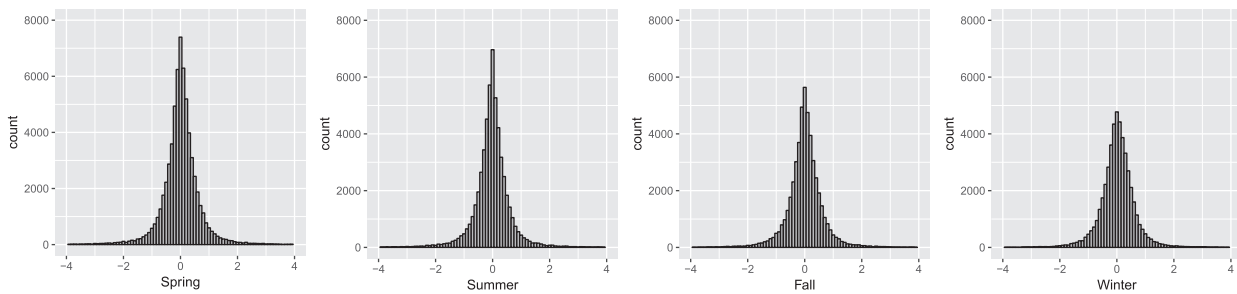

Fig. 3. Empirical distribution of $\{\omega_t\}$ for Eugene, OR, 2013.

where L is the Cholesky factor of the empirical autocorrelation matrix of $\boldsymbol{\varepsilon}$ based on residuals from that day. Fig. 3 shows histograms of the estimated decorrelated residuals broken up by season. Generally most values of the decorrelated residuals are concentrated on the interval $[-2, 2]$. We propose a mixture of distributions to model $\boldsymbol{\omega}$ that changes over time of the day, taking account the influence of solar zenith angle (Pecenak et al., 2016). In particular, ω_t is set to be a standard normal random variable if t is a time point corresponding to either the first two, or last two hours during the light hours. For every other time, ω_t is modeled as a centered and scaled Beta(α, β) random variable whose support is $[-2, 2]$. After testing on data at Eugene, we opt for the following parameters: $\alpha = 5.5$ and $\beta = 5.6$. This non-Gaussian mixture follows from initial data exploration; we compare performance against a solely Gaussian model in the Appendix.

4. Validation

Our model is illustrated by stochastically downscaling NSRDB estimates of GHI and comparing to in situ pyranometer measurements at Eugene, OR, as well as nine stations from the SRML/SURFRAD in the year of 2013 (see Table 2). The Beta distribution parameters and empirical autocorrelations are based on data from Eugene throughout, but could easily be updated to include historical in situ measurements from other climatic regions, where available.

4.1. Autocorrelation and coverage properties

We begin by validating the temporal correlation and coverage of the estimated GHI. The validation based on autocorrelations are split into two sections, the first at measurement locations where the empirical autocovariance is available, the second where the autocovariance is predicted. In either case, autocovariances are estimated separately over each 30 min time interval coinciding with the hour, split by season. For example, all data from 1 pm-1:29 pm during summer (June, July, August) based on empirical residuals from Eugene are compiled and their empirical covariance is stored, similarly for spring, fall and winter, and every other 30 min interval. We split validation statistics by season and time of the day as solar irradiance exhibits substantial diurnal and seasonal variability, and aggregation of verification statistics can misrepresent actual local performance (Kleiber et al., 2011).

4.1.1. In-sample testing

We first consider validation at Eugene, where direct pyranometer measurements can inform model parameters and correlations. Fig. 4 shows empirical and model autocorrelations on clear sky index split by season and time of the day (morning, noon and afternoon) at Eugene, OR. Although correlations are modeled on the log-residual scale, the method accurately captures the relatively high temporal correlation on the clear sky index scale. Moreover, there is apparent seasonal and diurnal variability in temporal structure. The model adequately adapts to this seasonal and diurnal variation.

The clear sky index represents a normalization of GHI time series to a theoretical clear sky irradiance, which is given in the NSRDB. A typical feature of the clear sky index is that the probability density function (PDF) of clear sky index time series is bimodal with the two peaks respectively corresponding to cloud cover and clear sky states (Suehrcke and McCormick, 1988). Fig. 5 shows the clear sky index PDF based on simulated and in situ data in 2013 for the validation station in Eugene. Our model accurately reproduces the bimodal behavior of the empirical pdf, but moves some probability from entirely clear sky days to slightly lower values. Note that the model is not trained to reproduce this density, and thus the goodness of fit in Fig. 5 is encouraging.

Because the method is stochastic, we do not expect any downscaled realization to exactly replicate of the true irradiance at any particular time point or location, but the statistical distribution of the simulations should capture the truth. For example, if we generate and order 100 downscaled realizations (at a single time point), we would expect the true GHI to fall above the lowest five and below the highest five about 90% of the time. Indeed, having accurate coverage intervals, known as calibration, is critical not only in probabilistic forecasting (Gneiting et al., 2007), but also in probabilistic resource assessment (Zhang et al., 2015). To assess this, we consider reliability plots; these plot empirical

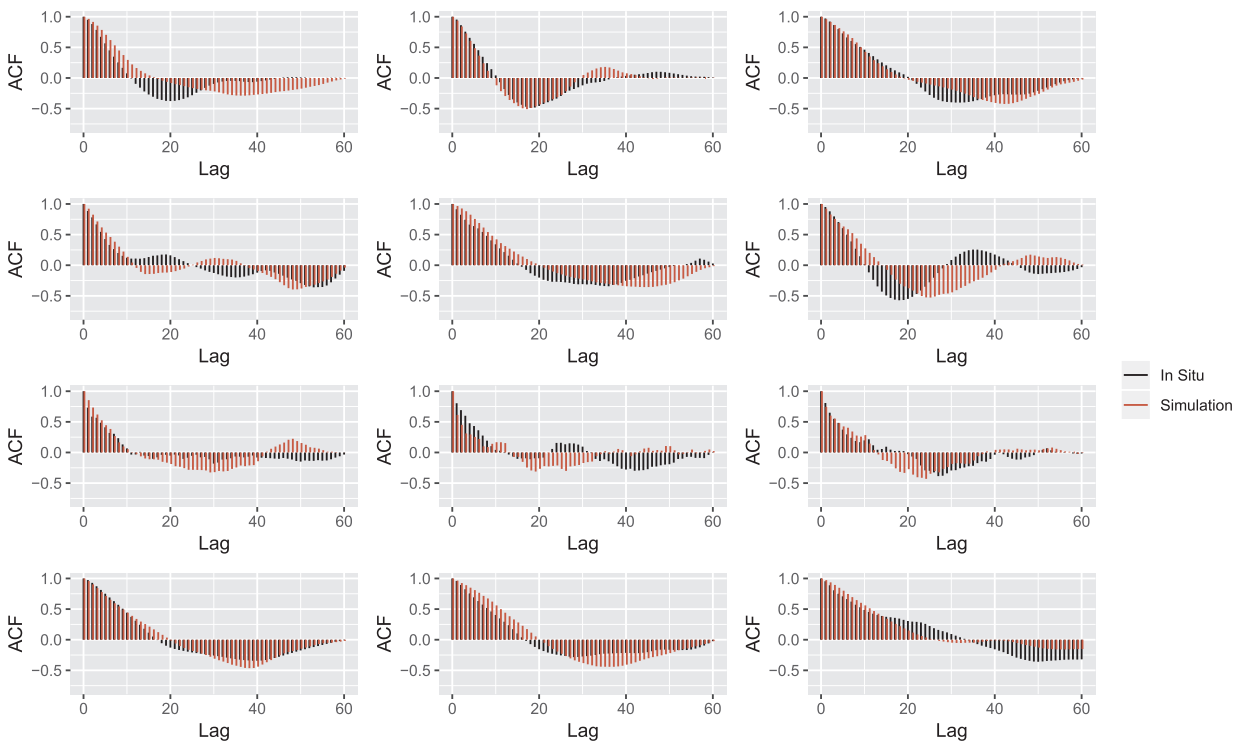

Fig. 4. Empirical and model autocorrelations on clear sky index at 1 min time resolution split by season and time of the day. Columns indicate examples of 60 min time intervals for a representative morning (9:00am-9:59am), noon (12:00 pm-12:59 pm) and afternoon hour (3:00 pm-3:59 pm). Rows index a representative day from each season: spring (March 20), summer (June 20), fall (September 20) and winter (December 20) at Eugene, OR.

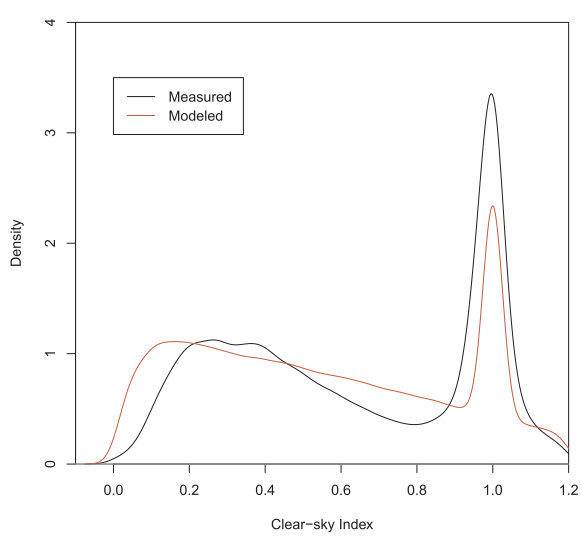

Fig. 5. Probability density distribution of clear sky index over Eugene's simulated and in situ data in 2013.

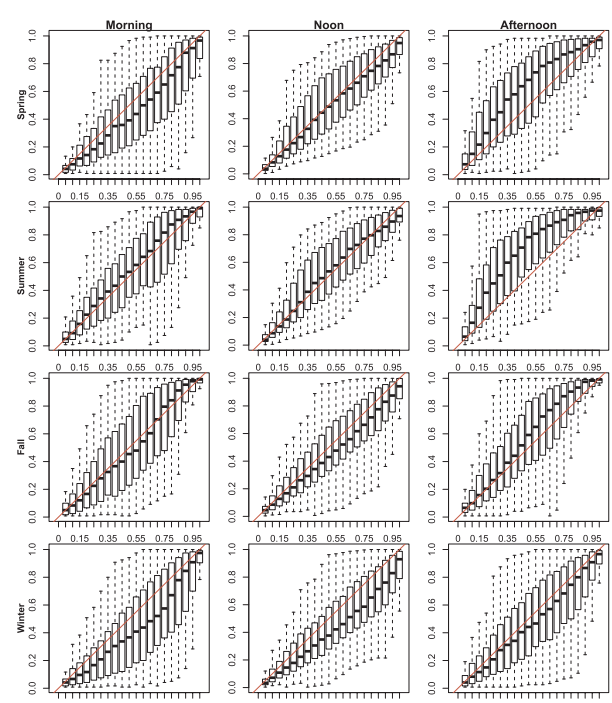

Fig. 6. Reliability plots for Eugene based on 1000 independent downscaled realizations at 1-min time resolution during the year 2013. Columns index morning, noon and afternoon periods and rows index spring, summer, fall and winter. Axis units are percentage.

coverage against nominal coverage (e.g., the empirical coverage of the 95% nominal confidence interval based on an ensemble of simulations). A calibrated downscaling ensemble will exhibit good coverage properties at all nominal levels, indicated by following the identity line on the reliability plot.

Fig. 6 shows reliability plots for downscaling at Eugene in 2013 split out by time of day and season. Empirical coverages are calculated based on 1000 independent downscaled realizations; we repeat this procedure 100 times, represented by boxplots. The median of each boxplot is generally close to the identity line indicating accurate coverage, while the worst coverage is apparent at Eugene during summer in the afternoon where the simulated ensembles are overdispersed. This may be related to the fact that summer tends to have less variability compared to other seasons. However, overall the simulated ensembles exhibit accurate coverages.

4.1.2. Out-of-sample testing

We now turn to simulation at locations without direct ground

Table 3
Statistics for the remaining validation stations based on 1-min downscaled ensembles.

Site	Mean squared errors for ACFs	Nominal levels		
		50%	75%	90%
Eugene, Oregon	5.24×10^{-4}	49.89%	75.12%	90.22%
Portland, Oregon	8.65×10^{-4}	49.76%	75.24%	90.34%
Salem, Oregon	7.30×10^{-4}	50.12%	74.33%	89.95%
Bondville, Illinois	3.74×10^{-3}	51.25%	76.27%	90.28%
Boulder, Colorado	6.02×10^{-3}	47.19%	73.25%	87.23%
Desert Rock, Nevada	9.43×10^{-3}	62.23%	80.54%	94.33%
Fort Peck, Montana	4.77×10^{-3}	48.34%	71.34%	88.78%
Goodwin Creek, Mississippi	5.59×10^{-3}	51.78%	76.27%	90.35%
Penn State University, Pennsylvania	4.50×10^{-3}	51.36%	76.18%	89.98%
Sioux Falls, South Dakota	3.09×10^{-3}	52.56%	76.10%	91.04%

The second column represents mean squared error in replicating empirical autocorrelation. Last three columns indicate average coverage probabilities at different nominal levels.

measurements where model parameters and autocovariances are not directly estimable. We swap in those model estimates from Eugene, and test the generalizability of the methodology at a number of other locations in Portland, Salem and locations in the SURFRAD dataset, respectively. We show results for Portland here, which is representative of the typical goodness-of-fit of our model, but refer the reader to Table 3 further validation statistics at other locations.

Fig. 7 contains autocorrelation functions of clear sky index based on in situ measurements and simulations during the morning, noon and afternoon hours on March 1, June 1, September 1 and December 1, 2013 respectively. Note that although the statistical model parameters are the same as in Fig. 4, the simulated autocorrelations differ due to differences in the NSRDB estimates for these two locations – thus, some temporal structure is informed by the 30 min estimates. Our method generally captures the shape of the in situ clear sky index correlations,

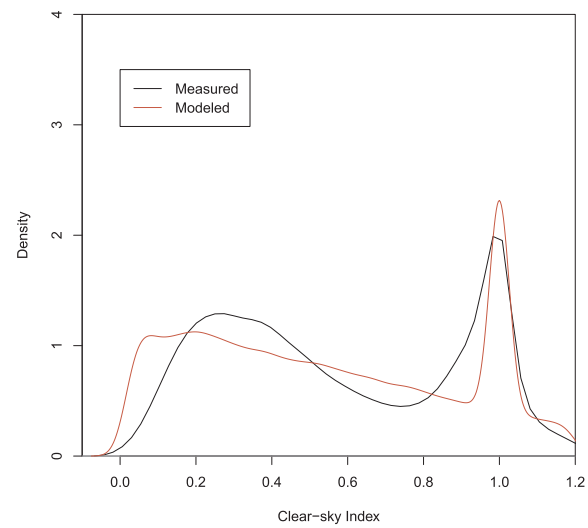

Fig. 8. Probability density distribution of clear sky index over Portland's simulated and in situ data in 2013.

but sometimes fails to capture small-scale structure. However, the next figure suggests that capturing such small-scale changes in correlation do not drastically impact the statistical calibration of the model.

Fig. 8 illustrates the clear sky index probability density function based on simulated and in situ data in 2013 for location Portland. Similar to Fig. 5, we capture the bimodality of this variable, but tend to slightly oversimulate clear sky days.

Fig. 9 shows statistical reliability at Portland using simulations based on model estimates from Eugene. Overall the stochastically downscaled ensembles still show adequate coverage properties. This is perhaps unsurprising as Eugene and Portland share similar climatological properties, but Figs. 7 and 9 are representative of model performance in other climates mentioned in Table 2.

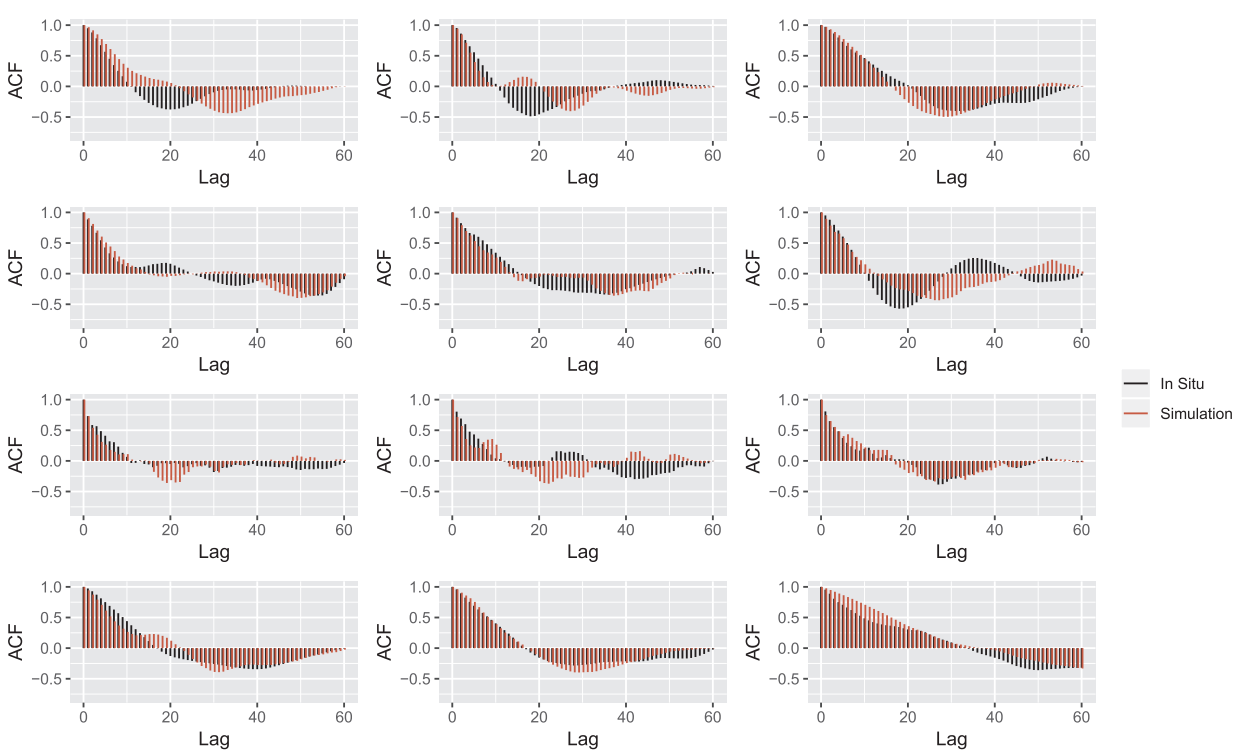

Fig. 7. Empirical and model autocorrelations on clear sky index at 1 min time resolution split by season and time of the day. Columns indicate examples of 60 min time intervals for a representative morning (9:00am-9:59am), noon (12:00 pm-12:59 pm) and afternoon hour (3:00 pm-3:59 pm). Rows index a representative day from each season: spring (March 1), summer (June 1), fall (September 1) and winter (December 1) at Portland, OR.

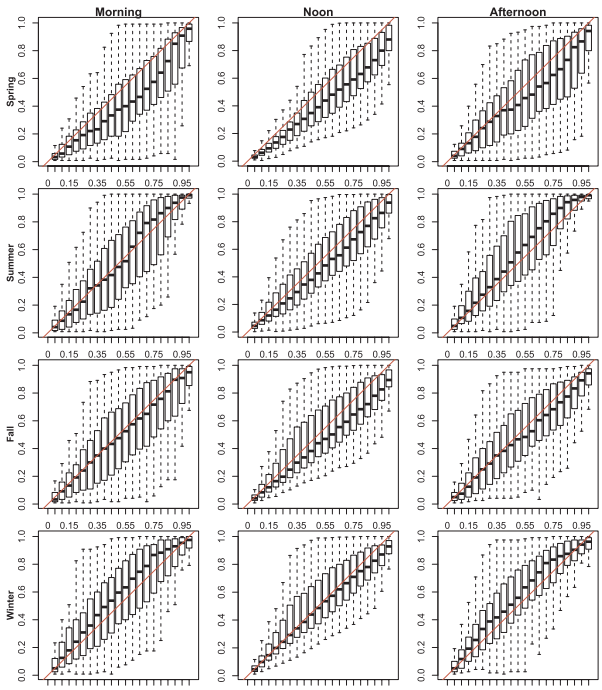

Fig. 9. Reliability plots for Portland based on 1000 independent downscaled realizations at 1-min time resolution during the year 2013 using estimated model parameters from Eugene. Columns index morning, noon and afternoon periods and rows index spring, summer, fall and winter. Axis units are percentage.

Although the figures for Portland are generally representative for all validation stations, we include statistics for the remaining validation stations in Table 3 to demonstrate the applicability to other climate regions. The table includes average squared error in replicating empirical autocorrelation,

$$\sum_{h} (\gamma(h) - \hat{\gamma}(h))^2 \tag{8}$$

where $\gamma(\cdot)$ is the autocorrelation function based on the simulated process and $\hat{\gamma}(\cdot)$ is the empirical autocorrelation function at a particular validation location. The autocorrelation functions are well-estimated based on this approach, in part due to the inherent autocorrelation present in the satellite data product. Table 3 also includes empirical reliability coverages for the validation locations, in particular average coverage probabilities at the nominal 50%, 75% and 90% levels. While generally the model translates to all stations in the SURFRAD dataset, the desert location at Desert Rock, NV is not fit as well.

4.2. Half-hourly validation

Although our method adequately captures 1-min statistics and correlations, comparison metrics which evaluate variability over a longer period of time such as a day or month are desirable. For example, hourly and monthly validation were performed in Bright et al. (2015) and Larrañeta et al. (2018), respectively.

We compare in situ and simulated mean half-hourly average irradiance at all validation locations. Fig. 10 shows a scatterplot of half-hourly means averaged across all locations in Table 2. Compared with the in situ data, averaging the minutely irradiance generated by the model over half-hourly timesteps shows a strong correlation ($R^2 = 0.981$). This suggests that our method adequately captures first-moment statistics of actual 30-min aggregated irradiance.

Fig. 11 shows the frequency distribution for mean half-hourly irradiance across all validation locations. The previous section examined behavior of the model at particular time points and seasons. The

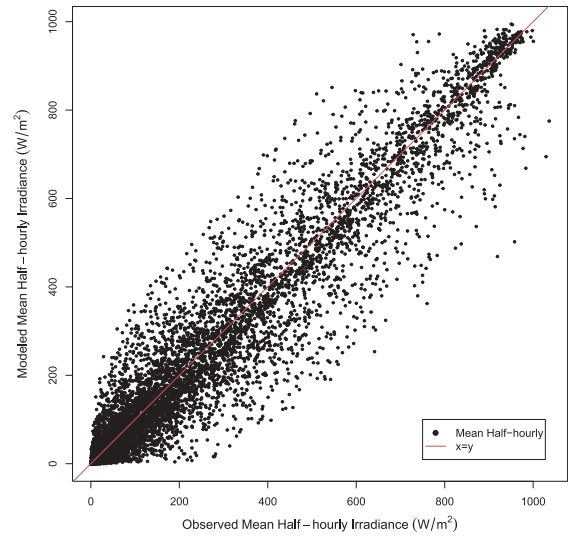

Fig. 10. Scatterplot of average half-hour GHI over the year 2013, observed (Obs) versus simulated (Mod). Each point represents half-hour averaged GHI for one of 9 validation locations for a particular half-hour interval. The identity line is added for reference.

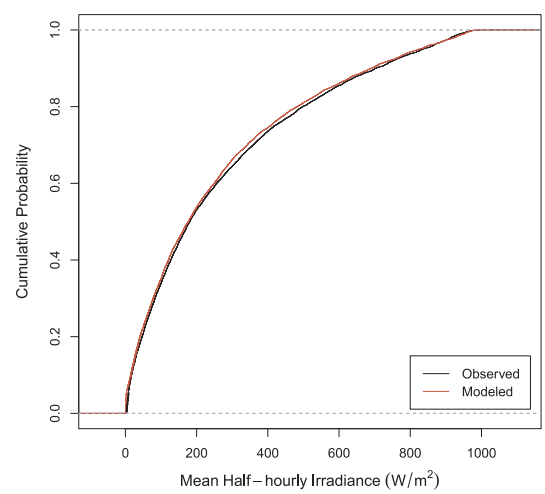

Fig. 11. Frequency distributions for mean half-hour GHI averaged over the 9 validation locations' simulated and in situ data in 2013.

frequency distribution of Fig. 11 is an aggregated statistic over all times, seasons and locations. The correlation between the simulated and observational averages is $R^2 = 0.999$, indicating very good performance of our stochastic downscaling approach. This approach preserves the low frequency mean by definition because the model conditions on the NSRDB data, which is the mean function.

While mean characteristics are properly maintained, variability over short time frames is also important to understand and capture. Fig. 12 shows average 30 min simulated variabilities in August, 2013 for three SRML locations along with the corresponding observed variability. Boxplots represent uncertainty in the model variability, and are based on 1000 simulations. Empirical variances are generally within the interquartile range. As in the previous section, the performance in this figure is representative of performance at the other tested locations.

4.3. Scattering events

Due to the method used to create the downscaled samples, the satellite-derived high-frequency irradiance samples are not expected to

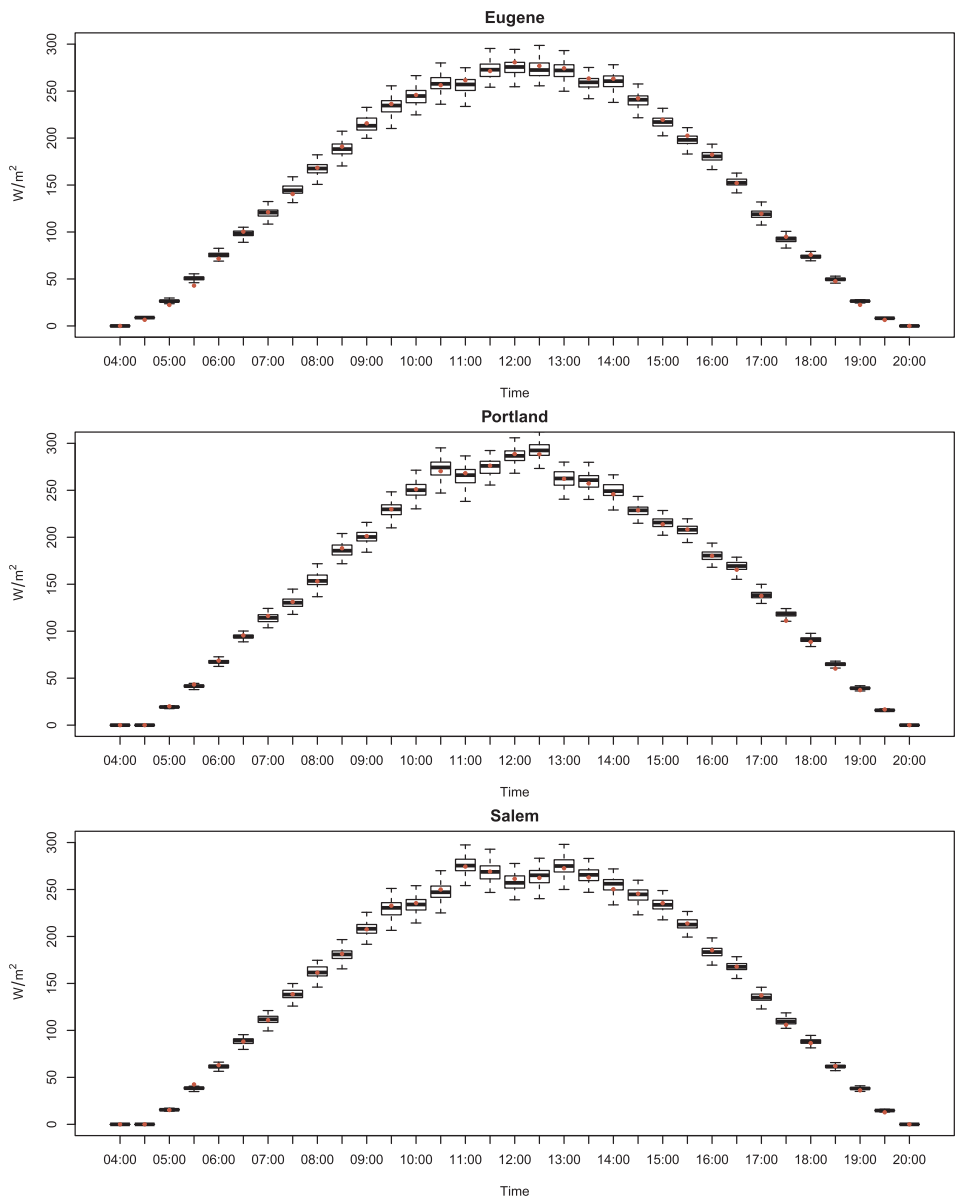

Fig. 12. Boxplots for 30 min standard deviation estimates and observed 30 min standard deviation (red points) during daytime based on the data of August, 2013 for select SRML locations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

exactly match the ground measured irradiance, but rather provide a statistically representative sample. Specifically, the timing of clouds may not match between downscaled and ground-measured irradiance variability. Pecenak et al. (2016) showed that irradiance enhancement is not due to cloud focusing, but scattering. Scattering events are those when less solar radiation is scattered and more solar radiation reaches the earth's surface, which usually occurs around midday with greater GHI.

To visualize how the model behaves during scattering events, the left panels in Fig. 13 show a selection of example days with simulations, in situ measurements and clear sky GHI for Eugene. The stochastic realizations are reasonable compared to the ground measurements, although not every scattering event is captured. The proposed stochastic model is not expected to replicate the in situ irradiance at the any given time point, whether or not scattering events occur. The right panels show pointwise 95% confidence bounds for the same days based on our downscaled estimates. The coverage of the grey band reveals the fact that the stochastic model plausibly captures the true irradiance process, even during midday periods when maximal GHI is experienced.

5. Conclusions and future work

The proposed stochastic downscaling model aims to capture the variability at a much higher temporal frequency than is available in data products such as the NSRDB. The statistical model varies with time of the day and season in order to capture diurnal and seasonal non-stationarity. We validate the model based on several metrics of variability indices for stations across the continental USA with different climates. The method shows good statistical calibration at numerous nominal levels. Moreover, low-order moments are reproduced very well, including means, variances and temporal correlations. The main output of this approach is an ensemble of estimated one-minute GHI, and the test results suggest the method appropriately captures the variability and temporal correlation of the true, unobserved irradiance process.

Future work may continue in a number of directions. We tested the model on the popular NSRDB dataset, but other satellite product datasets may be of interest. Temporally downscaling to an even finer time resolution, say one second, can be important to understand very high

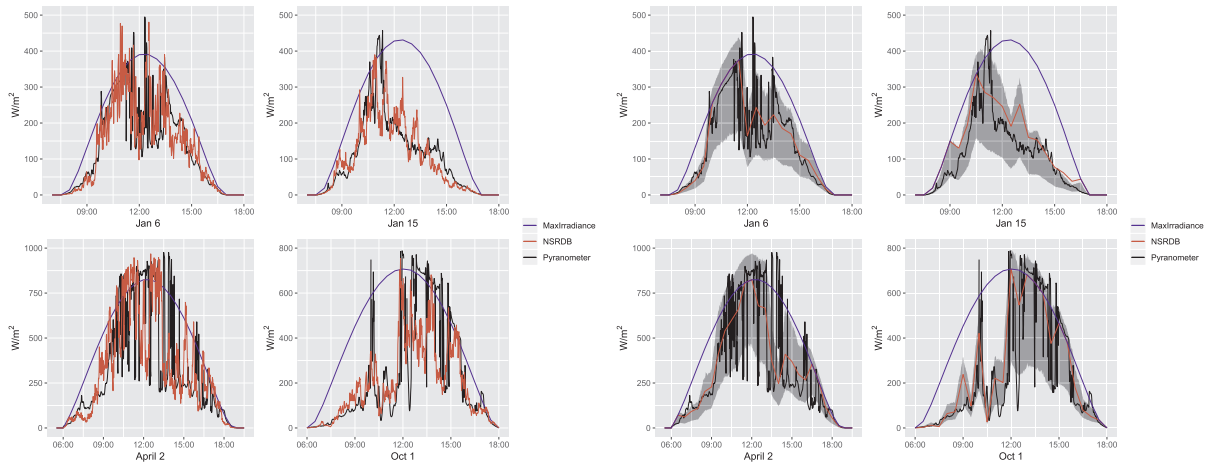

Fig. 13. Left panels: simulations, in situ data and clear sky GHI for sample days in Eugene, OR. Right panels: 95% pointwise confidence bands and linearly interpolated NSRDB for the same spatial location and days. Units are in Watts per meter squared.

frequency variability. It remains to be seen if this model is appropriate for such timescales, but would require very high frequency in situ data for training. The main drawback of the method is that it is currently point-level; spatially consistent simulations are important to understand how variability smooths over regions, this is a focus of our current efforts. The current methodology is a building block toward a coherent framework.

Even though this work has demonstrated geographic flexibility by validating in different climates, there is still more that can be done to allow for geographic flexibility. Lave et al. (2012) used geographic smoothing techniques to include a spatial dimension. Geographic flexibility was represented as a function of both longitude and latitude to analyze clear sky index distribution with cloud over for different climatic regions in (Smith et al., 2017). Sigrist et al. (2015) introduced an approach to generating spatially consistent ensembles using an advection–diffusion stochastic partial differential equation (SPDE). Future research may be devoted to incorporating the non-Gaussian mixture proposal of this manuscript with spatially-consistent generative approaches.

Acknowledgements

This work was authored by Alliance for Sustainable Energy, LLC, the Manager and Operator of the National Renewable Energy Laboratory for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Solar Energy Technologies Office. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes. Kleiber and Zhang's research was partially supported by NSF grants DMS-1406536, DMS-1417724, BCS-146576 and DMS-1811294.

Appendix A. Comparison of fit between different distributions to the empirical data

In this appendix we compare the performance of using different distributions for ω_t on statistical calibration. Fig. 14 shows reliability plots split over diurnal times (morning, noon and afternoon) and season for three possible stochastic models for ω_t: (1) the proposed normal-Beta mixture, (2) only normal random variables and (3) only Beta random variables. The plots suggest that using normal random variables during the midday period yields overdispersed simulations, while beta random variables yield underdispersion in the early and late day. Thus, the mixture model is a desirable balance between these two extremes.

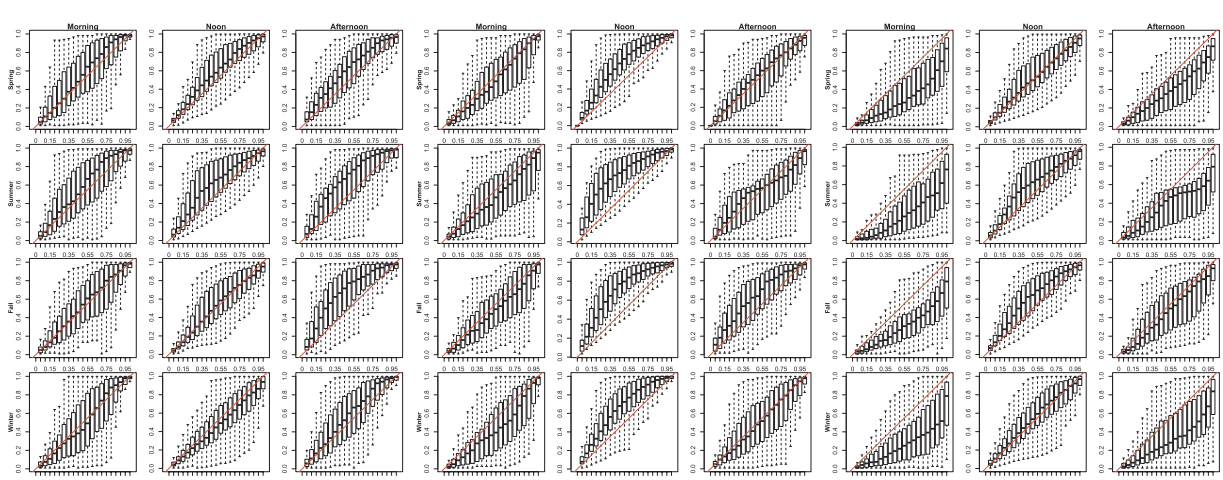

Fig. 14. Reliability plots using mixture model (left 4 × 3 grid), normal model (middle 4 × 3 grid) and Beta model (right 4 × 3 grid).

References

Augustine, J.A., DeLuisi, J.J., Long, C.N., 2000. SURFRAD–A national surface radiation budget network for atmospheric reserach. Bull. Am. Meteorol. Soc. 81, 2341–2357.

Baharin, K.A., Rahman, H.A., Hassan, M.Y., Gan, C.K., 2014. Hourly irradiance forecasting in Malaysia using support vector machine. In: 2014 IEEE Conference on Energy Conversion (CENCON), pp. 185–190.

Boland, J.W., 2008. Time series modelling of solar radiation. In: Modeling Solar Radiation at the Earth's Surface. Springer, pp. 283–312.

Bright, J.M., Babacan, O., Kleissl, J., Taylor, P.G., Crook, R., 2017. A synthetic, spatially decorrelating solar irradiance generator and application to a LV grid model with high PV penetration. Sol. Energy 147, 83–98.

Bright, J.M., Smith, C., Taylor, P.G., Crook, R., 2015. Stochastic generation of synthetic minutely irradiance time series derived from mean hourly weather observation data. Sol. Energy 115, 229–242.

Cameron, D.R., Crane, L., Parker, S.S., Randall, J.M., 2017. Solar energy development and regional conservation planning. In: Energy Sprawl Solutions. Springer, pp. 66–75.

Clifton, A., Hodge, B., Draxl, C., Badger, J., Habte, A., 2018. Wind and solar resource data sets. Wiley Interdiscipl. Rev.: Energy Environ. 7 (2), e276.

Darcovich, K., Kenney, B., MacNeil, D., Armstrong, M., 2015. Control strategies and cycling demands for li-ion storage batteries in residential micro-cogeneration systems. Appl. Energy 141, 32–41.

Diagne, M., David, M., Lauret, P., Boland, J., Schmutz, N., 2013. Review of solar irradiance forecasting methods and a proposition for small-scale insular grids. Renew. Sustain. Energy Rev. 27, 65–76.

Dise, J., Kankiewicz, A., Schlemmer, J., Hemker, K., Kivalov, S., Hoff, T., Perez, R., 2013. Operational improvements in the performance of the SUNY satellite-to-solar irradiance model using satellite infrared channels. In: 2013 IEEE 39th Photovoltaic Specialists Conference (PVSC). IEEE, pp. 0960–0963.

Fernández-Peruchena, C.M., Blanco, M., Gastón, M., Bernardos, A., 2015. Increasing the temporal resolution of direct normal solar irradiance series in different climatic zones. Sol. Energy 115, 255–263.

Fernández-Peruchena, C.M., Gastón, M., 2016. A simple and efficient procedure for increasing the temporal resolution of global horizontal solar irradiance series. Renew. Energy 86, 375–383.

Gneiting, T., Balabdaoui, F., Raftery, A.E., 2007. Probabilistic forecasts, calibration and sharpness. J. Roy. Statist. Soc.: Series B (Statist. Methodol.) 69, 243–268.

Grantham, A.P., Pudney, P.J., Boland, J.W., 2018. Generating synthetic sequences of global horizontal irradiation. Sol. Energy 162, 500–509.

Hansen, C.W., Stein, J.S., Ellis, A., 2010. Statistical criteria for characterizing irradiance time series. Sandia Report, SAND2010-7314.

Hoff, T.E., Perez, R., 2010. Quantifying PV power output variability. Sol. Energy 84 (10), 1782–1793.

Jamaly, M., Kleissl, J., 2017. Spatiotemporal interpolation and forecast of irradiance data using kriging. Sol. Energy 158, 407–423.

Kleiber, W., Raftery, A.E., Gneiting, T., 2011. Geostatistical model averaging for locally calibrated probabilistic quantitative precipitation forecasting. J. Am. Statist. Assoc. 106, 1291–1303.

Kumar, A., Gomathinayagam, S., Giridhar, G., Mitra, I., Vashistha, R., Meyer, R., Chhatbar, K., 2014. Field experiences with the operation of solar radiation resource assessment stations in India. Energy Procedia 49, 2351–2361.

Larrañeta, M., Fernandez-Peruchena, C., Silva-Pérez, M., Lillo-Bravo, I., 2018. Methodology to synthetically downscale DNI time series from 1-h to 1-min temporal resolution with geographic flexibility. Sol. Energy 162, 573–584.

Larrañeta, M., Moreno-Tejera, S., Silva-Pérez, M.A., Lillo-Bravo, I., 2015. An improved model for the synthetic generation of high temporal resolution direct normal irradiation time series. Sol. Energy 122, 517–528.

Lave, M., Kleissl, J., 2010. Solar variability of four sites across the state of Colorado. Renew. Energy 35 (12), 2867–2873.

Lave, M., Kleissl, J., Arias-Castro, E., 2012. High-frequency irradiance fluctuations and geographic smoothing. Sol. Energy 86 (8), 2190–2199.

Lave, M., Kleissl, J., Stein, J.S., 2013. A wavelet-based variability model (WVM) for solar PV power plants. IEEE Trans. Sustain. Energy 4 (2), 501–509.

Lave, M., Quiroz, J., Reno, M.J., Broderick, R.J., 2017. Temporal Resolution Load Variability Compared to PV Variability. In: 2017 IEEE 44th Photovoltaic Specialist Conference (PVSC). IEEE, pp. 1–6.

Lew, D., Brinkman, G., Ibanez, E., Hodge, B., Hummon, M., Florita, A., Heaney, M., 2013. The western wind and solar integration study phase 2. Technical report, National Renewable Energy Laboratory (NREL), Golden, CO.

Melzi, F.N., Touati, T., Same, A., Oukhellou, L., 2016. Hourly solar irradiance forecasting based on machine learning models. In: 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, pp. 441–446.

Morf, H., 2013. A stochastic solar irradiance model adjusted on the ångström–prescott regression. Sol. Energy 87, 1–21.

Nguyen, A., Velay, M., Schoene, J., Zheglov, V., Kurtz, B., Murray, K., Kleissl, J., 2016. High PV penetration impacts on five location distribution networks using high resolution solar resource assessment with sky imager and quasi-steady state distribution system simulations. Sol. Energy 132, 221–235.

Ngoko, B.O., Sugihara, H., Funaki, T., 2014. Synthetic generation of high temporal resolution solar radiation data using Markov models. Sol. Energy 103, 160–170.

Pecenak, Z.K., Mejia, F.A., Kurtz, B., Evan, A., Kleissl, J., 2016. Simulating irradiance enhancement dependence on cloud optical depth and solar zenith angle. Sol. Energy 136, 675–681.

Perez, M.J., Fthenakis, V.M., 2015. On the spatial decorrelation of stochastic solar resource variability at long timescales. Sol. Energy 117, 46–58.

Perez, R., Kivalov, S., Schlemmer, J., Hemker Jr., K., Hoff, T., 2011. Parameterization of site-specific short-term irradiance variability. Sol. Energy 85 (7), 1343–1353.

Perez, R., Lauret, P., Perez, M., David, M., Hoff, T.E., Kivalov, S., 2018. Solar resource variability. Wind Field and Solar Radiation Characterization and Forecasting: A Numerical Approach for Complex Terrain, pp. 149.

Perez, R., Schlemmer, J., Hemker, K., Kivalov, S., Kankiewicz, A., Gueymard, C., 2015. Satellite-to-irradiance modeling-a new version of the SUNY model. In: 2015 IEEE 42nd Photovoltaic Specialist Conference (PVSC). IEEE, pp. 1–7.

Peruchena, C.F., Larrañeta, M., Blanco, M., Bernardos, A., 2018. High frequency generation of coupled GHI and DNI based on clustered dynamic paths. Sol. Energy 159, 453–457.

Polo, J., Zarzalejo, L., Marchante, R., Navarro, A., 2011. A simple approach to the synthetic generation of solar irradiance time series with high temporal resolution. Sol. Energy 85 (5), 1164–1170.

REN21, 2016. Renewable Global Status Report; 2016.

Reno, M.J., Deboever, J., Mather, B., 2017. Motivation and requirements for quasi-static time series (QSTS) for distribution system analysis. In: IEEE PES General Meeting.

Reno, M.J., Hansen, C.W., 2016. Identification of periods of clear sky irradiance in time series of GHI measurements. Renew. Energy 90, 520–531.

Sayeef, S., Heslop, S., Cornforth, D., Moore, T., Percy, S., Ward, J., Berry, A., Rowe, D., 2012. Solar intermittency: Australia's clean energy challenge. characterising the effect of high penetration solar intermittency on australian electricity networks.

Schmalensee, R., 2015. The future of solar energy: A personal assessment. Energy Econ. 52, S142–S148.

Sengupta, M., Xie, Y., Lopez, A., Habte, A., Maclaurin, G., Shelby, J., 2018. The national solar radiation data base (NSRDB). Renew. Sustain. Energy Rev. 40, 51–60.

Sigrist, F., Künsch, H.R., Stahel, W.A., 2015. Stochastic partial differential equation based modelling of large space–time data sets. J. Roy. Statist. Soc.: Series B (Statist. Methodol.) 77 (1), 3–33.

Smith, C.J., Bright, J.M., Crook, R., 2017. Cloud cover effect of clear sky index distributions and differences between human and automatic cloud observations. Sol. Energy 144, 10–21.

Suehrcke, H., McCormick, P.G., 1988. The frequency distribution of instantaneous insolation values. Sol. Energy 40 (5), 413–422.

Vignola, F., Perez, R., 2004. Solar Resource GIS Data Base for the Pacific Northwest using Satellite Data–Final Report. Technical report.

Wang, Z., Wang, F., Su, S., 2011. Solar irradiance short-term prediction model based on bp neural network. Energy Procedia 12, 488–494.

Wang, J., Zhu, X., Lubkeman, D., Lu, N., Samaan, N., 2017. Continuation power flow analysis for PV integration studies at distribution feeders. In: 2017 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT). IEEE, pp. 1–5.

Zhang, J., Draxl, C., Hopson, T., Delle Monache, L., Vanvyve, E., Hodge, B., 2015. Comparison of numerical weather prediction based deterministic and probabilistic wind resource assessment methods. Appl. Energy 156, 528–541.

Zhang, W., Kleiber, W., Florita, A.R., Hodge, B., Mather, B., 2018. Modeling and simulation of high frequency solar irradiance, IEEE J. Photovolt. (submitted for publication).