SoDa: An Irradiance-Based Synthetic Solar Data Generation Tool

Ignacio Losada Carreño
Raksha Ramakrishna
Anna Scaglione
School of Electrical Engineering
Arizona State University
Tempe, Arizona
{ilosadac,rramakr6,ascaglio}@asu.edu

Daniel Arnold
Ciaran Roberts
Sy-Toan Ngo
Sean Peisert
Energy Technologies Area
Lawrence Berkeley National Laboratory
Berkeley, California
{dbarnold,CMRoberts,SyToanNgo,sppeisert}@lbl.gov

David Pinney
Analytics, Resiliency and
Reliability Work Group
NRECA
Washington, D.C
{dpinney}@me.com

***Abstract*—In this paper, we present SoDa, an irradiance-based synthetic Solar Data generation tool to generate realistic sub-minute solar photovoltaic (PV) output power time series, that emulate the weather pattern for a certain geographical location. Our tool relies on the National Solar Radiation Database (NSRDB) to obtain irradiance and weather data patterns for the site. Irradiance is mapped onto a PV model estimate of a solar plant's 30-min power output, based on the configuration of the panel. The working hypothesis to generate high-resolution (e.g. 1 second) solar data is that the conditional distribution of the time series of solar power output given the cloud density is the same for different locations. We therefore propose a stochastic model with a switching behavior due to different weather regimes as provided by the cloud type label in the NSRDB, and train our stochastic model parameters for the cloudy states on the high-resolution solar power measurements from a Phasor Measurement Unit (PMU). In the paper we introduce the stochastic model, and the methodology used for the training of its parameters. The numerical results show that our tool creates synthetic solar time series at high resolutions that are statistically representative of the measured solar power and illustrate how to make use of the tool to create synthetic data for arbitrary sites in the footprint covered by the NSRDB.**

***Index Terms*—Solar PV data, synthetic models, NSRDB**

I. Introduction

There is an increasing need for realistic simulations using high-resolution solar power generation datasets that explore the impact of the penetration of solar power and devise control strategies [1]. For this, it is important to effectively model the uncertainty and variability in solar energy at fast time resolutions and produce realistic synthetic data.

Several papers have proposed different approaches to model the stochastic nature of solar energy [2] that can be broadly classified as physics based, model-based and model-free. Physics based methods such as [3] are based on the underlying physics that governs solar irradiance and sometimes incorporate numerical weather predictions (NWP). Model-based methods like in [4] use statistical models such as Markov models to capture the uncertainty and variability. Model-free methods are black-box methods, like artificial neural networks [5] and support vector machines to mimic the variability in solar data by using large training datasets. There also exist hybrid methods that are a combination of aforementioned approaches. However, high-resolution solar power measurements, that could be used to analyze a feeder where there are multiple sources are scarce, and forecasting methods do not lend themselves directly to producing synthetic time series, since they rely on historical solar PV data. More abundant are, however, historical solar irradiance data that can be used to produce solar PV data as shown in [6] for example.

Although solar irradiance measurements can be recorded using ground-based sensing instruments like pyranometers and pyrheliometers, their accuracy is not very high and strongly depends on data acquisition and calibrations methods [7], [8]. Additionally, the frequency of sampling varies widely from minutes to seconds and these datasets are known to contain spurious or missing instances, thus, they are not suited for power systems modeling.

Thus, solar irradiance measurements from geostationary satellites and state-of-the-art Physical Solar Models (PSMs) are used to develop the National Solar Radiation Database (NSRDB) [9]. NSRDB has achieved spatial resolution up to 4 kilometers and temporal resolution of 30 minutes. However, the temporal resolution is wholly inadequate to model the uncertainty and variability of solar resources in the electric power grid as it pertains to the interaction with other controllable devices on the grid. Sub-minute solar variability may impact the power quality of grid-connected systems [10], inverter sizing [11], or may lead to suboptimal inverter control policies [12].

Contribution: The development of SoDa is motivated by the need of using realistic high-resolution solar PV data to train algorithms that can ensure voltage stability via Distributed Energy Resources (DER) control. In this paper we provide

This research was supported in part by the Director, Cybersecurity, Energy Security, and Emergency Response, Cybersecurity for Energy Delivery Systems program, of the U.S. Department of Energy, under contract DE-AC02-05CH11231. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors of this work.

a method to fill the gap highlighted and synthetically generate statistically accurate solar data time series at high resolution. It leverages the NSRDB and constructs a stochastic model for the cloud behavior that fill the 30 minutes intervals with realistic solar PV trends, obtained by matching the statistics with those of a site where high resolution solar power data from PMUs are available. The stochastic model used is inspired by our prior modeling work in [13], [14] which is a statistical model that is physics-inspired. Our model, unlike traditional PSMs, scales for high resolutions, from seconds to minutes.

Note that although previous works as in [5] also generate synthetic solar generation scenarios, they require large sets of training data and may require retraining when generating scenarios for a different geographical region or different panel characteristics. Our SoDa tool is adaptive and could be applied to any geographical region in the ambit of NSRDB, without the need for additional training.

This manuscript is organized as follows. In Section II , we leverage the NSRDB and present a method to generate power time series from weather variables. Section III details the cloud regime parametrization and stochastic models used for each regime. Section IV presents the results of a study compared against measurements from a PMU. We conclude the paper and outline future directions in Section V.

II. Modeling PV power generation from an irradiance-based solar model

Satellite remote sensing yields high-density measurements that are ingested by PSMs, and can be used to create solar datasets like the NSRDB. Although these datasets may not capture the variability of interest to model DER control, they can model the seasonal and geographical variability in larger time scales that a stochastic model may not be able to reproduce from training data. The NSRDB is the current state-of-the-art solar dataset in the US, and therefore our algorithm to generate synthetic solar power time series for a certain location leverages the NSRDB for the development of our tool. Prior to detailing our model, we describe the dataset.

A. The National Solar Radiation Database (NSRDB)

The NSRDB is a serially complete state-of-the-art collection of meteorological data developed at the National Renewable Energy Lab (NREL). It uses a physics-based solar model that leverages satellite-based measurements to generate meteorological variables, covering the entire United States along with other international locations. Time series in the NSRDB are provided at every 4-km and 30 minutes from 1998 to 2017. The dataset includes measurements of solar radiation: global horizontal (GHI), direct normal (DNI), diffuse horizontal irradiance (DHI) and other weather variables such as atmospheric pressure, ground-level temperature, wind speeds or cloud coverage. NREL offers a free Python API to retrieve NSRDB data from any location available which we have integrated in our tool, providing flexibility to operators looking to obtain solar times series at specific locations.

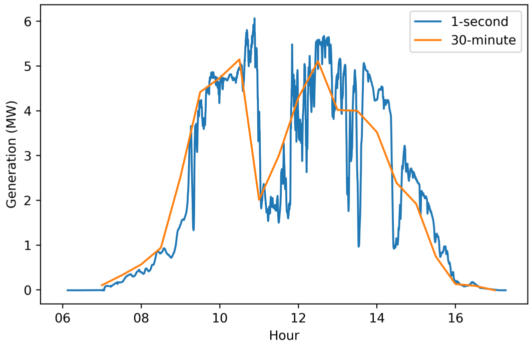

Fig. 1. A comparison between a 30-min and 1-second solar profiles showing how 30-minute data are insufficient to capture the underlying variability.

Although the NSRDB complements ground-based solar measurements by making time series available at multiple locations, it does not, however, match the sampling frequency of pyranometers. Consequently, the direct use of the NSRDB in dynamic studies may mislead the operator by not capturing the variability under 30-minute intervals, resulting in bad control policies. An alternative solution is to interpolate low-resolution data to obtain a more granular dataset. In practice, this method is not suited for the generation of high-resolution data since the variability of solar irradiance in sub-hourly times frames may strongly deviate from the interpolated values. This is illustrated in Fig. 1.

Prior to introducing the variability due to clouds of the solar PV output via a stochastic model, we generate a low resolution solar power time series using the 30-minute irradiance data from the NSRDB. We use the approach presented by [6], validated against real PV panel performance [15], [16] and used by NREL's System Advisor Model (SAM) [17]. This model ingests DNI, DHI, wind speed and temperature data from the NSRDB to produce 30-minute solar power time series. The method is explained in the next subsection.

B. Mapping irradiance into PV power generation

Solar irradiation is received by a panel with an angle of incidence θ. The angle of incidence θ in a fixed-tilt solar panel is given by

$$\theta = \cos^{-1}\left(\sin\theta_s \cos\left(\gamma - \gamma_s\right) \sin\beta + \cos\theta_s \cos\beta\right) \quad (1)$$

where $\theta_s, \gamma, \gamma_s$ and β are the solar zenith, surface azimuth, solar azimuth and surface tilt angles, respectively. A similar algorithm to calculate θ in one-axis tracker panels is available in the work from [18]. To model the shading, we introduce an attenuation $f(\theta)$:

$$f(\theta) = b_0 + b_1\theta + b_2\theta^2 + b_3\theta^3 + b_4\theta^4 + b_5\theta^5 \quad (2)$$

where $(b_0, \ldots, b_5)$ are coefficients of a polynomial fit that is specific to the glass in the panel. The transmitted irradiance on the plane-of-array $I_t[k]$ at time $k = 0, \ldots, n-1$, is:

$$I_t[k] = f \cos\theta I_n[k] + I_{ds}[k] + I_{dg}[k] \quad (3)$$

where $I_n[k], I_{ds}[k], I_{dg}[k]$ are the normal, sky diffuse and ground diffuse irradiance, respectively. The normal and diffuse components of solar irradiance are often given as a result of running climate models. In our work, these values are obtained from the NSRDB. To calculate the DC PV power we introduce an efficiency factor for the panel due to temperature:

$$\xi[k] = p_0^{dc}(1 + \kappa(\tau_c[k] - \tau_r)) \tag{4}$$

where p_0^{dc} is the solar panel DC nameplate capacity given in watts, $\kappa = -0.5\%/C$ is a temperature coefficient, and τ_c, τ_r are the cell and reference temperatures, respectively. It should be noted that $\tau_c[k]$ is generated as a result of running a thermal model [19] that accounts for the ambient temperature and convection losses due to air flow. Then, we can define the vector $\boldsymbol{\iota}[k] = [I_n[k], I_{ds}[k], I_{dg}[k]]^\top$ and $\boldsymbol{\nu}[k] = [f(\theta_k)\cos(\theta_k), 1, 1]^\top$ and express the resulting DC PV power produced as:

$$\hat{p}^{dc}[k] = \xi[k] \left(\boldsymbol{\iota}[k]\right)^\top \boldsymbol{\nu}[k] \tag{5}$$

We also model the effect of the inverter as a low-pass filter, clipped at the nameplate capacity of the inverter $\bar{p} = \frac{p_0^{dc}}{\delta}$ where $\delta \geq 1$ is the *DC-to-AC ratio*. Thus, the resulting AC power at time instant $k \in \mathcal{T}_h$ and $n = |\mathcal{T}_h| = 48$ is defined as follows,

$$\hat{p}^{ac}[k] = \begin{cases} \epsilon \hat{p}^{dc}[k] & 0 < \hat{p}^{dc}[k] < \bar{p} \quad k_u \leq k \leq k_d \\ \epsilon \bar{p} & \hat{p}^{dc}[k] \geq \bar{p} \quad\quad\; k_u \leq k \leq k_d \\ 0 & \text{else} \end{cases} \tag{6}$$

where k_u, k_d are the sunrise and sunset times, respectively and $\epsilon < 1$ is an coefficient that accounts for AC losses, e.g. inverter losses, wiring, step-up transformer etc. The temporal resolution of the generated time series is that of the NSRDB, i.e. 30 minutes or 48 intervals in 24 hours. In the following section, we address the issue of augmenting the resolution of the data by using a stochastic model that realistically reproduces the trends expected due to the weather information. Specifically, we define a model that will allow us to leverage the power time series from a PMU as training data to learn the parameters for the given weather pattern labels that the NSRDB data provide for the corresponding location, as explained in more detail next.

III. Using stochastic models to generate high-resolution solar power time series

In this section, we propose a method to learn the parameters from PMU measurements that characterize the statistics of solar power in minute and sub-minute time scales under different weather regimes. Based on these estimated parameters, samples of solar data are drawn from the appropriate distributions. We argue that climatic and topographic effects on solar power are captured by the NSRDB, and that the conditional distribution of sub-minute variability of solar power given the cloud type is the *same* for different locations. The model is loosely inspired by our previous modeling efforts [13] for probabilistic forecasting of solar PV outputs [14].

A. Cloud regime parametrization

We denote the AC component of the power produced by the PV plant and measured by the PMU on day d and time instant $k \in \mathcal{T} = \{0, 1, \ldots, nT-1\}$ as $p_d[k]$, where T is the number of samples in a 30-minute interval. The superscript in $\hat{p}_d^{ac}[k]$ has been omitted for readability in the high resolution signal. Solar irradiation is attenuated by clouds and aerosols, modeled as a random mask that subtracts a percentage of the incoming light rays at any given time. We model $p_d[k]$ as follows

$$p_d[k] = s_d[k] - p_d^b[k] + p_d^e[k] \tag{7}$$

where $s_d[k]$ is an upsampled version of the 30 minute solar PV generation calculated from (1)-(6), and $p_d^b[k], p_d^e[k]$, capture the attenuation of the direct component and edge-of-cloud effect, respectively. We argue that any attenuation of the diffuse component happens over longer than 30 minute periods and therefore, such effects are encapsulated in $s_d[k]$. To obtain $s_d[k]$ at high resolution we up-sample ($\uparrow T$) and use linear interpolation filter $g[k] = (1 - k/T)rect_T(k)$, where $\mathrm{rect}_T(k)$ is a rectangular pulse between $[0, T]$, that is:

$$s_d[k] = \sum_{\ell=0}^{n-1} \hat{p}^{ac}[\ell] g[k - \ell T] \tag{8}$$

The attenuation of solar irradiation depends on the cloud type which is provided by the NSRDB [9]. However, the cloudy type label corresponds to an interval of 30 minutes. Let $\vartheta[\ell] \in \{0, 1, \ldots, 12\}$ be the cloud type for the ℓ^{th} interval, $k \in \mathcal{T}_\ell = \{(\ell-1)T, \ldots, \ell T\}$. We consider two cases, the sunny regime and cloudy regime that yield for an interval $k \in \mathcal{T}_\ell$ the model:

$$p_d[k] = \begin{cases} s_d[k] - p_d^b[k] & \vartheta[\ell] \leq 1 \\ s_d[k] - p_d^b[k] + p_d^e[k] & \vartheta[\ell] > 1 \end{cases} \tag{9}$$

It should be noted that $p_d^e[k]$ cannot exist in the sunny weather due to the absence of low and mid-level clouds that may cause this effect. Similarly, the existence of $p_d^b[k]$ in the sunny weather is very sparse. Such a model allows the separation of the components and study a plausible stochastic model for them. In the following subsection, we present a stochastic model that encompasses all the cloud regimes represented as labels in NSRDB dataset.

B. Parametrization of the attenuation components

For any weather regime, $\forall \vartheta[\ell]$, the power components are:

$$p_d^b[k] \approx \sum_q \tilde{h}_w[q] z_w[k-q], \;\; p_d^e[k] \approx \sum_q \tilde{h}_m[q] z_m[k-q] \tag{10}$$

where the direct attenuation, $p_d^b[k]$, and edge-of-cloud effect, $p_d^e[k]$, components are modelled as the convolution of a one-dimensional filter $\tilde{h}_m[q], \tilde{h}_w[q]$, with stochastic input $z_m[k]$ and $z_w[k]$, respectively. The filter $\tilde{h}_w[q]$ models the attenuation of power due to clouds and aerosols represented as drop in direct solar power and is chosen to be a Hamming window:

$$\tilde{h}_w[q] = 0.54 - 0.46 \cos\left(\frac{2\pi q}{M-1}\right) \quad 0 \leq q \leq M-1 \tag{11}$$

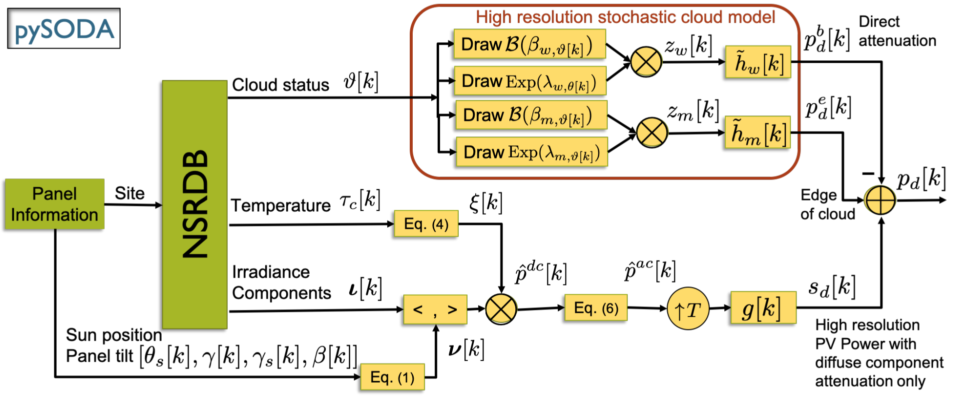
pySODA
High resolution stochastic cloud model
Draw $\mathcal{B}(\beta_{w,\vartheta[k]})$
Draw $\mathrm{Exp}(\lambda_{w,\theta[k]})$
Draw $\mathcal{B}(\beta_{m,\vartheta[k]})$
Draw $\mathrm{Exp}(\lambda_{m,\vartheta[k]})$
$z_w[k]$
$z_m[k]$
$\tilde{h}_w[k]$
$\tilde{h}_m[k]$
$p_d^b[k]$
Direct attenuation
$p_d^e[k]$
Edge of cloud
$-$
$p_d[k]$
Cloud status $\vartheta[k]$
NSRDB
Panel Information
Site
Temperature $\tau_c[k]$
Eq. (4)
$\xi[k]$
Irradiance Components $\boldsymbol{\iota}[k]$
$<\ ,\ >$
$\hat{p}^{dc}[k]$
Eq. (6)
$\hat{p}^{ac}[k]$
$\uparrow T$
$g[k]$
$s_d[k]$
High resolution PV Power with diffuse component attenuation only
Sun position
Panel tilt $[\theta_s[k], \gamma[k], \gamma_s[k], \beta[k]]$
Eq. (1)
$\boldsymbol{\nu}[k]$

Fig. 2. A block diagram that summarizes the modelling and generation of synthetic 1-second resolution solar data. We obtain a deterministic solar time series at 30-minute resolution using the NSRDB. We upsample the signal and and use the high-resolution stochastic cloud model with parameters trained on PMU data to generate the final results at 1-second resolution.

Here, the large size of a megawatt-scale PV plant has a geographical "smoothing" effect on the attenuation, and the Hamming window best reflects this behaviour, instead of our prior model that considers a sudden drop in power as in [13]. The filter $\tilde{h}_m[q]$ is a variant of the Morlet wavelet as follows:

$$\tilde{h}_m[k] = e^{\left(\frac{2k}{M}-1\right)^2} \left|\cos\left(-20(1+2k/M)\right)\right|, \quad 0 \le k \le M-1$$

and it is used to capture the edge of cloud effect [20] followed by the attenuation of the direct power component. The filters are shown in Fig. 3.

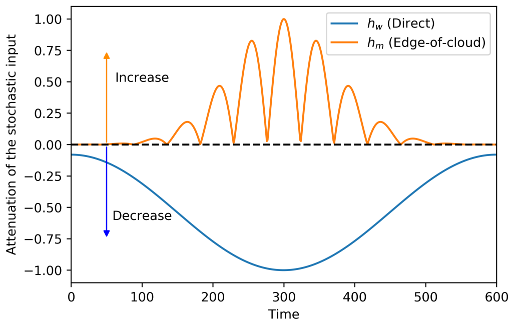

Fig. 3. Hamming window and Morlet wavelets

The stochastic input $z_m[k]$ is the product of a Bernoulli random variable $b_m[k]$ and an exponentially distributed random variable $z'_m[k]$ whose parameters depend on the cloud type index $\vartheta[\ell]$. Input $z_w[k]$ is also similar,

$$z_m[k] = b_m[k] z'_m[k], \qquad z_w[k] = b_w[k] z'_w[k] \quad k \in \mathcal{T}_\ell \tag{12}$$

$$b_m[k] \sim \mathcal{B}(\beta_{m,\vartheta[\ell]}), z'_m[k] \sim \mathrm{Exp}(\lambda_{m,\vartheta[\ell]}), \tag{13}$$

$$b_w[k] \sim \mathcal{B}(\beta_{w,\vartheta[\ell]}), z'_w[k] \sim \mathrm{Exp}(\lambda_{w,\vartheta[\ell]}), \tag{14}$$

C. Learning the parameters of the stochastic models

To learn the parameters from the PMU data we ignore the Bernoulli factors in fitting the observations, and regularize the problem inducing sparsity in $z_m[k]$ and $z_w[k]$ solving a LASSO problem [21]. Let $\mathbf{y}_d \in \mathbb{R}^{nT}$ denote the vector of attenuated power for a day d with entries defined as $y_d[k] = s_d[k] - p_d[k]$. The assumption is that:

$$\mathbf{y}_d = \mathbf{\Phi}_m \mathbf{z}_m + \mathbf{\Phi}_w \mathbf{z}_w + \boldsymbol{\varepsilon}, \quad \mathbf{y}_d \in \mathbb{R}^{nT}, \mathbf{z}_m, \mathbf{z}_w \in \mathbb{R}^{(nT-M)}$$

where $\boldsymbol{\varepsilon}$ is modeling error and $\mathbf{\Phi}_w, \mathbf{\Phi}_m \in \mathbb{R}^{nT\times(nT-M)}$ are the Toeplitz matrices of the convolution operations $\mathbf{y}_w = \mathbf{h}_w * \mathbf{z}_w$, $\mathbf{y}_m = \mathbf{h}_m * \mathbf{z}_m$ of a one dimensional filter $\mathbf{h}_w, \mathbf{h}_m \in \mathbb{R}^M$ with stochastic input $\mathbf{z}_w, \mathbf{z}_m \in \mathbb{R}^{(nT-M)}$. The Toeplitz matrices $\mathbf{\Phi}_w, \mathbf{\Phi}_m$ are characterized by the first columns $[h_w[0], \ldots, h_w[M-1], \mathbf{0}^{nT-M}]$ and $[h_m[0], \ldots, h_m[M-1], \mathbf{0}^{nT-M}]$, and first rows $[h_w[0], \mathbf{0}^{nT-M-1}]$ and $[h_m[0], \mathbf{0}^{nT-M-1}]$, respectively. Here, we want to estimate the stochastic input vectors $\mathbf{z}_m$, $\mathbf{z}_w$ regularized in a LASSO formulation to avoid overfitting:

$$\begin{aligned} \min_{\mathbf{z}_m, \mathbf{z}_w} & \|\mathbf{y} - \mathbf{\Phi}_m \mathbf{z}_m - \mathbf{\Phi}_w \mathbf{z}_w\|^2 + \rho_m(\mathbf{1}^\top \mathbf{z}_m) + \rho_w(\mathbf{1}^\top \mathbf{z}_w) \\ & \text{subject to} \ \ \mathbf{z}_m, \mathbf{z}_w \ge 0, \end{aligned} \tag{15}$$

After obtaining the estimates of $\mathbf{z}_m, \mathbf{z}_w$, we estimate the parameters of the Bernoulli and exponential distributions for each cloud type. More specifically, given the cloud type $\vartheta[\ell] \in \{0, 1, \ldots, 12\}$, the unknown distribution parameters are

$$\Theta \triangleq \{\lambda_{m,0}, \lambda_{m,1}, \ldots \lambda_{m,12}, \lambda_{w,0}, \lambda_{w,1}, \ldots \lambda_{w,12}\} \tag{16}$$

$$B \triangleq \{\beta_{m,0}, \beta_{m,1}, \ldots \beta_{m,12}, \beta_{w,0}, \beta_{w,1}, \ldots \beta_{w,12}\} \tag{17}$$

To estimate, we segment the data based on the cloud type $\mathcal{Z}_m^i = \{\mathbf{z}_m[k] | k \in \mathcal{T}_\ell, \vartheta[\ell] = i\}$ and $\mathcal{Z}_w^i = \{\mathbf{z}_w[k] | k \in \mathcal{T}_\ell, \vartheta[\ell] = i\}, \forall \ell, i \in \{2, \ldots, 12\}$ and set the values below a certain threshold τ to zero. Also, by definition, $\lambda_{m,0} = \lambda_{m,1} = \infty$ since these correspond to sunny regime. The remaining $\lambda_{m,i}, \lambda_{w,i}$ are maximum likelihood (ML) estimates, i.e.

$$\lambda_{m,i} = \frac{1}{\sum_{\mathbf{z}_m[k]\in\mathcal{Z}_m^i, \mathbf{z}_m[k]>\tau_m} \mathbf{z}_m[k]}, \tag{18}$$

$$\lambda_{w,i} = \frac{1}{\sum_{\mathbf{z}_w[k]\in\mathcal{Z}_w^i, \mathbf{z}_w[k]>\tau_w} \mathbf{z}_w[k]} \quad i \in \{2, \ldots, 12\} \tag{19}$$

TABLE I

SUMMARY OF THE STOCHASTIC INPUT PARAMETERS LEARNED FROM THE 1-SECOND RESOLUTION PMU DATA

NSRDB Cloud Type	**Label ID**	$\lambda_{m,i}$	$\lambda_{w,i}$	$\beta_{m,i}$	$\beta_{w,i}$
Clear, Unknown Type	0, 10	∞	5.98	0.0013	0.0019
Probably Clear	1	∞	5.80	0.0028	0.0090
Fog, Dust, Smoke	2, 11, 12	3.29	6.50	0.0097	0.0035
Water, Overshooting	3,9	3.90	6.07	0.0055	0.0028
Super-Cooled Water	4	3.25	5.88	0.0189	0.0041
Opaque Ice	6	4.19	4.83	0.0004	0.0011
Cirrus	7	3.10	5.15	0.0074	0.0042
Mixed, Overlapping	4, 8	4.09	6.66	0.0036	0.0080

Similarly, we estimate the parameters of the corresponding Bernoulli distribution as

$$\beta_{m,i} = |\mathcal{Z}_m^i|^{-1} \sum_{\mathbf{z}_m[k] \in \mathcal{Z}_m^i, \mathbf{z}_m[k] > \tau_m} 1 \tag{20}$$

$$\beta_{w,i} = |\mathcal{Z}_w^i|^{-1} \sum_{\mathbf{z}_w[k] \in \mathcal{Z}_w^i, \mathbf{z}_w[k] > \tau_w} 1 \tag{21}$$

We depict the block diagram describing the overall stochastic modelling in Fig 2.

IV. RESULTS

Learning the parameters for the stochastic model using PMU data at 1-second resolution:
We train the stochastic model using data from a solar site in Riverside, CA, where a PMU measures the generation at a rate of 1-second per sample. We use a filter of length $M = 600$ seconds to account for the effects of clouds traveling across the PV site. Our assumptions are based on an average cloud speed of 40 kmph and we use PMU data from a 7.5 MWdc plant with an extension of approximately 1.5 km^2. It would take 5 minutes minutes for the cloud to pass directly over the plant, and the tails of the wavelets (i.e. the remaining 5 minutes) model the transition between clouds. The parameters are calculated as a result of solving the problem in (15). We use the regularizers ρ_m, ρ_w to ensure the input to be sparse. In particular, we use $\rho_m \gg \rho_w$ since we expect the expect the stochastic input of the Morlet wavelet filter to be sparser that of the Hamming window, e.g. $\rho_m = 0.1$, $\rho_w = 0.01$. The estimation of parameters using (18) and (20) yields the parameters presented in Table I. We use $\tau_m = 0.2, \tau_w = 0.1$ to threshold the data. The NSRDB contains 13 different labels for the cloud classification. Some of these cloud types are rarely present in the data, e.g. dust and smoke conditions. Furthermore, the closest NRSDB site to the PMU location did not contain any data corresponding to label IDs 4,9,10,11 and 12. However, the NSRDB includes a description of the cloud type that we leveraged to merge different labels as shown in Table I. Fig. 5 shows a plot a histogram of the non-zero components of vectors $\mathbf{z}_m, \mathbf{z}_w$ to validate our assumption that the non-zero components of the stochastic input follows a quasi exponential distribution as in (12).

Next, we showcase the stochastic models by drawing samples from the appropriate distributions. We show the results for two different sites where PMU data is sampled every second and every minute, respectively.

Generating synthetic solar data at 1-second resolution: In Fig. 4, we show the 1-second resolution results for different weather conditions throughout the day, namely a sunny, partly-cloudy and two cloudy days. Our results are statistically representative of those measured by the PMU. We show slightly above average variability in the early and late hours of the day. We argue that this may be an artifact of the surroundings of the PMU site, such as buildings, trees or walls that may shade the panels during the sunrise and sunset hours.

Generating synthetic solar data at 1-minute resolution: We prove the validity of our models for a different site in Berkeley, CA where a PMU measures solar generation with a 1-minute sampling rate. Our models are trained on 1-second resolution data, thus, we first generate solar time series at that resolution. Then, we downsample the solar generation to match the resolution of the PMU. The results are shown in Fig. 6.

Potential improvements: We acknowledge some of the limitations of our model. For instance, the model does not account for time-of-day variations in the variability of solar generation. This may create undesired fluctuations of solar power when the generation is low. It should be noted that this model is trained with MW-scale PMU data, and that further corrections may be needed if single kW-scale installations are modeled. For instance, when modeling rooftop solar, the land covered by the panels (e.g. a neighborhood) should be large enough in order for our assumptions to hold. Also, this model leverages the NSRDB to generate synthetic data and although the footprint of the NSRDB is rather large, the use of this tool may be limited to some areas in the world. Moreover, a mismatch between the NSRDB classification and the actual cloud regime may exist.

However, note that, the scope of our modeling approach can be extended beyond using the NSRDB. If there is sufficient input from the user to accurately describe sunny day pattern such as information about the latitude, longitude, panel orientation, capacity of the system, we can account for geographical variability. The cloud-type input can be drawn from weather prediction databases. With only these two inputs, it is possible to generate synthetic high-resolution solar power dataset for any desired location. The NSRDB gives us more information than what is needed for the success of our method and therefore, we incorporate that in our generation (or interpolation) of solar profiles. Finally, we have validated our models using data from two plants, in Riverside and Berkeley. To reaffirm the validity of our assumptions and models, a more thorough validation is left for future work.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented stochastic models that, trained on PMU data, can generate synthetic statistically-representative solar time series at 1-second resolution. This model, unlike traditional PSM, can scale for high resolutions. Furthermore, we test the performance of our tool in two different locations and show that its use is not limited to a single temporal solution, e.g. 1-second or 1-minute.

Going forward, we want to address the aforementioned limitations and extend the capabilities of our tool by allowing

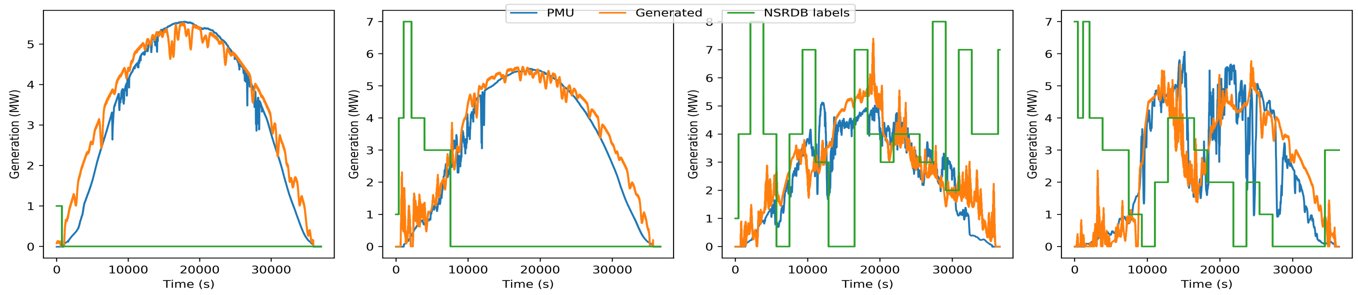

Fig. 4. Generation of solar time series as a result of using the stochastic models with the parameters learned from the regression problem provided in (15). We show three different days characterized by sunny (first), partly cloudy (second) and cloudy (third and fourth) weather. Results are shown for a one dimensional filter of length $M = 600$, i.e. 10 minutes.

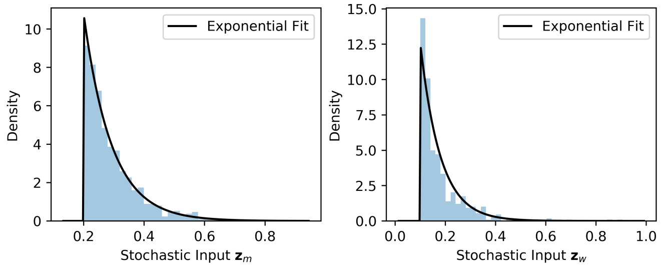

Fig. 5. Histogram of the non-zero components of stochastic input. On top of the histogram, we show the exponential distribution that best fits the histogram data.

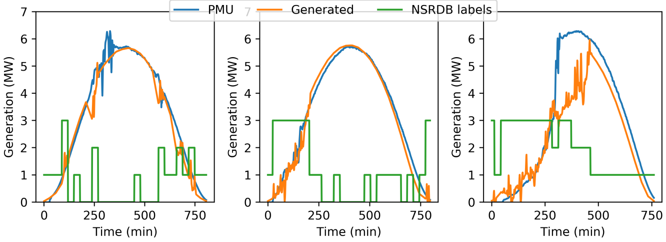

Fig. 6. Generation of stochastic 1-minute resolution solar time series for a site in Berkeley, CA. The right figure shows a mismatch between the generated and measures solar data that can be attributed to inaccurate cloud labeling.

the user to input its own low-resolution solar profile. A Python version of this tool will soon be available in *pip* and *Github*.

References

[1] A. Mills, "Understanding variability and uncertainty of photovoltaics for integration with the electric power system," 2009.

[2] R. H. Inman, H. T. Pedro, and C. F. Coimbra, "Solar forecasting methods for renewable energy integration," *Progress in energy and combustion science*, vol. 39, no. 6, pp. 535–576, 2013.

[3] C. B. Martinez-Anido, B. Botor, A. R. Florita, C. Draxl, S. Lu, H. F. Hamann, and B.-M. Hodge, "The value of day-ahead solar power forecasting improvement," *Solar Energy*, vol. 129, pp. 192–203, 2016.

[4] M. D. Tabone and D. S. Callaway, "Modeling Variability and Uncertainty of Photovoltaic Generation: A Hidden State Spatial Statistical Approach," *IEEE Transactions on Power Systems*, vol. 30, no. 6, pp. 2965–2973, Nov 2015.

[5] Y. Chen, Y. Wang, D. Kirschen, and B. Zhang, "Model-free renewable scenario generation using generative adversarial networks," *IEEE Transactions on Power Systems*, vol. 33, no. 3, pp. 3265–3275, 2018.

[6] A. P. Dobos, "Pvwatts version 5 manual," National Renewable Energy Lab.(NREL), Golden, CO (United States), Tech. Rep., 2014.

[7] I. Reda, T. Stoffel, and D. Myers, "A method to calibrate a solar pyranometer for measuring reference diffuse irradiance," *Solar Energy*, vol. 74, no. 2, pp. 103–112, 2003.

[8] T. Stoffel, I. Reda, D. Myers, D. Renne, S. Wilcox, and J. Treadwell, "Current issues in terrestrial solar radiation instrumentation for energy, climate, and space applications," *Metrologia*, vol. 37, no. 5, p. 399, 2000.

[9] M. Sengupta, Y. Xie, A. Lopez, A. Habte, G. Maclaurin, and J. Shelby, "The national solar radiation data base (nsrdb)," *Renewable and Sustainable Energy Reviews*, vol. 89, pp. 51–60, 2018.

[10] M. Patsalides, D. Evagorou, G. Makrides, Z. Achillides, G. E. Georghiou, A. Stavrou, V. Efthimiou, B. Zinsser, W. Schmitt, and J. H. Werner, "The effect of solar irradiance on the power quality behaviour of grid connected photovoltaic systems," in *International Conference on Renewable Energy and Power Quality*, 2007, pp. 1–7.

[11] S. Chen, P. Li, D. Brady, and B. Lehman, "Determining the optimum grid-connected photovoltaic inverter size," *Solar Energy*, vol. 87, pp. 96–116, 2013.

[12] S. Kundu, S. Backhaus, and I. A. Hiskens, "Distributed control of reactive power from photovoltaic inverters," in *2013 IEEE International Symposium on Circuits and Systems (ISCAS2013)*. IEEE, 2013, pp. 249–252.

[13] R. Ramakrishna and A. Scaglione, "A compressive sensing framework for the analysis of solar photo-voltaic power," in *2016 50th Asilomar Conference on Signals, Systems and Computers*. IEEE, 2016, pp. 308–312.

[14] R. Ramakrishna, A. Scaglione, V. Vittal, E. Dall'Anese, and A. Bernstein, "A model for joint probabilistic forecast of solar photovoltaic power and outdoor temperature," *IEEE Transactions on Signal Processing*, vol. 67, no. 24, pp. 6368–6383, 2019.

[15] N. Blair, A. Dobos, and N. Sather, "Case studies comparing system advisor model (sam) results to real performance data," National Renewable Energy Lab.(NREL), Golden, CO (United States), Tech. Rep., 2012.

[16] J. Freeman, J. Whitmore, N. Blair, and A. P. Dobos, "Validation of multiple tools for flat plate photovoltaic modeling against measured data," in *2014 IEEE 40th Photovoltaic Specialist Conference (PVSC)*. IEEE, 2014, pp. 1932–1937.

[17] N. Blair, A. P. Dobos, J. Freeman, T. Neises, M. Wagner, T. Ferguson, P. Gilman, and S. Janzou, "System advisor model, sam 2014.1. 14: General description," National Renewable Energy Lab.(NREL), Golden, CO (United States), Tech. Rep., 2014.

[18] W. Marion and A. Dobos, "Rotation angle for the optimum tracking of one-axis trackers (research report no. tp-6a20-58891)," *National Renewable Energy Laboratory, Golden, CO, USA*, 2013.

[19] M. Fuentes, "A simplified thermal model for flat-plate photovoltaic arrays (no. sand-85-0330)," *Sandia National Labs., Albuquerque, NM (USA)*, 1987.

[20] A. Kankiewicz, M. Sengupta, and D. Moon, "Observed impacts of transient clouds on utility-scale pv fields," in *Solar 2010 Conference Proceedings*, vol. 2009, 2010.

[21] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.