energies

MDPI

Article

Generation Data of Synthetic High Frequency Solar Irradiance for Data-Driven Decision-Making in Electrical Distribution Grids

Mohammad Rayati [1,*], Pasquale De Falco [2], Daniela Proto [3], Mokhtar Bozorg [1] and Mauro Carpita [1]

[1] Institut d'Energie et Systèmes Electriques (IESE), Haute École d'Ingénierie et de Gestion du Canton de Vaud (HEIG-VD), Haute École Spécialisée de Suisse Occidentale (HES-SO), 1041 Yverdon-les-Bains, Switzerland; mokhtar.bozorg@heig-vd.ch (M.B.); mauro.carpita@heig-vd.ch (M.C.)
[2] Department of Engineering, University of Naples Parthenope, 80133 Naples, Italy; pasquale.defalco@uniparthenope.it
[3] Department of Electrical Engineering, Università Federico II of Napoli, 80138 Naples, Italy; danproto@unina.it
* Correspondence: mohammad.rayati@heig-vd.ch

Abstract: In this paper, we introduce a model representing the key characteristics of high frequency variations of solar irradiance and photovoltaic (PV) power production based on Clear Sky Index (CSI) data. The model is suitable for data-driven decision-making in electrical distribution grids, e.g., descriptive/predictive analyses, optimization, and numerical simulation. We concentrate on solar irradiance data since the power production of a PV system strongly correlates with solar irradiance at the site location. The solar irradiance is not constant due to the Earth's orbit and irradiance absorption/scattering from the clouds. To simulate the operation of a PV system with one-minute resolution for a specific coordinate, we have to use a model based on the CSI of the solar irradiance data, capturing the uncertainties caused by cloud movements. The proposed model is based on clustering the days of each year into groups of days, e.g., (i) cloudy, (ii) intermittent cloudy, and (iii) clear sky. The CSI data of each group are divided into bins of magnitudes and the transition probabilities among the bins are identified to deliver a Markov Chain (MC) model to track the intraday weather condition variations. The proposed model is tested on the measurements of two PV systems located at two different climatic regions: (a) Yverdon-les-Bains, Switzerland; and (b) Oahu, Hawaii, USA. The model is compared with a previously published N-state MC model and the performance of the proposed model is elaborated.

Keywords: data analysis; electrical distribution grids; Markov Chain (MC) model; numerical simulation; photovoltaic (PV) systems; solar irradiance

check for updates

Citation: Rayati, M.; De Falco, P.; Proto, D.; Bozorg, M.; Carpita, M. Generation Data of Synthetic High Frequency Solar Irradiance for Data-Driven Decision-Making in Electrical Distribution Grids. *Energies* **2021**, *14*, 4734. https://doi.org/10.3390/en14164734

Academic Editor: Dimitrios Katsaprakakis

Received: 23 June 2021
Accepted: 28 July 2021
Published: 4 August 2021

1. Introduction

The energy transition pathway has been proposed in global and regional scales for decarbonization of the energy sector and its transformation to a sustainable system, keeping global warming below the permissible limit. The 21st Conference of the Parties (COP21) Paris Agreement of 2015 has individuated the global scale pathway [1]; as an example of a regional pathway, the Swiss Federal Office of Energy (SFOE) has developed the "Energy Strategy Act 2050" to prepare Switzerland for the energy transition [2]. It is supposed to enable Switzerland's grid to withstand a high capacity of renewable energy through the promotion of small photovoltaic (PV) systems in electrical distribution grids by exploiting the benefits of power industry digitalization [3]. However, the diffusion of PV systems and the high frequency variations of power flow in electrical distribution grids bring operational challenges such as voltage deviations due to the excess/deficit of generation, power quality decrement, and reliability issues [4]. To address these challenges, the power industry digitalization enables us to use data-driven decision-making [5].

For analyzing and solving the adverse impacts of the high penetration of PV systems in electrical distribution grids, data-driven decision-making based on numerical simulation is practical. For this purpose, we have to simulate the operation of an electrical distribution grid over time under the high penetration of PV systems. A numerical simulation requires a model representing the key characteristics and behaviors of the physical systems and processes [6]. It is worth mentioning that using a model to generate synthetic data instead of using the recorded real measurement data for numerical simulation has the following benefits:

1. In contrast to recorded measurement data and to approaches used in commercial software [7], a model is more general and non-biased against the occurrence of a specific scenario.
2. For stochastic numerical simulation, an infinite amount of data can be generated.
3. To investigate the sensitivity of a numerical simulation to specific model parameters, we can change the parameters of the model.

In this paper, we address the problem of finding a suitable model for PV systems' power production that may be used for the numerical simulation of an electrical distribution grid. Since the power production of a PV system in an electrical distribution grid strongly correlates with the solar irradiance data, we have to model the high frequency variations of solar irradiance at the site coordinate. The solar irradiance variations depend on the Earth's orbit and irradiance absorption/scattering from the clouds.

An exemplary application of the numerical simulation of an electrical distribution grid has been presented in [8], which is based on bucket testing. The robustness of a treatment strategy in electrical distribution grids for solving the issues of high PV penetration has been evaluated in [8], using a numerical simulation based on a stochastic model of PV systems power production. Numerical simulation has been used in [9] to perform a predictive analysis of the grid imbalance risks in the presence of a Markov Chain (MC) model of PV systems power production. Finally, the applications of numerical simulation have been reviewed in [10] with the objective of optimizing PV and battery storage systems integration.

The performance of data-driven decision-making based on numerical simulation for an electrical distribution grid highly depends on the accuracy of the PV system model. The uncertainties of PV systems power production, autocorrelation, and correlation with other variables directly affect the quality of data-driven decision-making in electrical distribution grids. To model and calculate the PV systems power production variations, the uncertainty related to the solar irradiance is normally modeled through the Clear Sky Index (CSI). The CSI is defined as the ratio between the Global Horizontal Irradiance (GHI) and corresponding clear sky GHI at the same time and at the same coordinate [11]. The CSI considers the condition of the sky, particularly cloud movements in a short-term scale, depending on the location and seasonal time.

In what follows, a review of previous works dealing with models proposed to reproduce PV power production variations is presented.

1.1. Literature Review

A recent literature survey has been given in [4] for describing the models of PV systems power production variations, which are used in the numerical simulation and data analysis studies. The most important features of PV systems power production variations are Probability Density Functions (PDFs) and Temporal Autocorrelation Functions (TAFs). The Beta PDF as presented in [12–14] and Gaussian PDF as given in [15,16] have been used for modeling the PV systems power production variations. The TAF of CSI time-series has been analyzed in [17–19], where the following four frequencies of TAFs have been discovered in the measurement data: (i) very short-term variations with a period of one minute; (ii) short-term variations with a period of one day; (iii) mid-term variations with a period of one month; and (iv) long-term variations with a period of one year.

The PV systems power production may be simulated with limited accuracy when we have both PDF and TAF data of solar irradiance variations. In order to improve the accuracy of numerical simulation, more sophisticated models of PV systems power production variations have been developed in recent years with the objective of accurate modeling. In this regard, there are three directions for improving the models of the PV systems power production variations: (i) using white-box models that return numerical weather predictions; (ii) adding other sources of data such as recorded meteorological input data; and (iii) using a grey- or black-box modeling strategy.

In the direction of the first strategy (i.e., using white-box models), a prediction model is presented in [20] that takes into account the thermal effects that occur in a PV system as well as the temperature-dependent nature of the PV system efficiency. In the direction of the second strategy (i.e., adding other sources of meteorological input data), Bright et al. [21] used the MC model to produce stochastic time-series of cloud cover depending on the seasons and prevailing pressure system; then, the transition probability matrices were trained using 10 years of coarse meteorological data (excluding the solar irradiance). Bright et al. [22] demonstrated that the synthetic one-minute resolution of solar irradiance time-series, varying on a spatial dimension, may be generated with the following inputs: (i) hourly average meteorological observations of *okta* (in meteorology, an *okta* is a unit of measurement used to describe the amount of cloud cover at any given location such as a weather station); (ii) wind speed; (iii) cloud height; and (iv) atmospheric pressure. A synthetic downscaling simulation of real CSI data based on clustering and an MC model has been developed in [23], in which the hidden patterns and trends in very short-term resolution have been discovered. The output of [23] is a simulated one-minute GHI dataset downscaled from 10 to 20 min real recorded data. (The standard of available meteorological data of next-generation satellites would be in 10 to 15 min.) Finally, the prerequisite on long historical data for downscaling and data discovery of solar irradiance variation has been relaxed in [24], which has used the datasets existing in the World Radiation Monitoring Center (WRMC), the central archive of the Baseline Surface Radiation Network (BSRN) [25]. In the direction of the third strategy (i.e., using grey- or black-box modeling strategy), the most recent models are copula [26], neural-network based [27], ensemble methods [28], deep learning [29], and MC [30,31]. Moreover, a synthetic long-term dataset of CSI time-series has been generated in [32] that is statistically indistinguishable from the observed data. The model of [32] requires an input from 10 to 15 annual time-series of hourly values that may be retrieved from satellite-based GHI data. The Hidden Markov Models (HMMs) with Gaussian observation have been used in [33] to generate synthetic CSI data. In [34], a PV systems power production model based on MC has been used for optimizing the size of a battery storage system.

The third strategy has two main advantages compared to the first and second ones as follows: (i) regarding the third strategy, the use of white-box models is computationally intensive, and numerical weather predictions are typically obtained from third-party providers; and (ii) there is not any presumption on the behaviour of PV systems power production in the third strategy and, as a result, it is more accurate. Due to the above-mentioned advantages, we concentrate on the third strategy for modeling PV systems power production variations. In this paper, we use a grey-box modeling strategy, which is only based on solar irradiance data input. (The solar irradiance data at a location may be measured easily by a pyranometer.) Therefore, it is in the direction of the mentioned third strategy. According to the dependency of PV systems power production and the solar irradiance, the data of PV systems power production are generated to be used for the numerical simulation of electrical distribution grids.

1.2. Contribution

We propose a model based on a number of individual N-state MCs for ensuring the consistency of synthetic and real data across different time resolutions, i.e., one minute, daily, and monthly. Since the proposed model is suitable for short- to mid-term variation

frequencies, we concentrate on one-minute, daily, and monthly temporal variations of solar irradiance, whereas we neglect the yearly variations.

Compared to the models of [21–23,26], where many variables such as wind speed, cloud cover, and weather temperature are needed, the model proposed in this paper only requires the measurement of GHI. Compared to [24], where the data of many locations are used for training the model simulating the downscaled one-minute PV systems power production of one day, the proposed model of this paper requires the data of one location and they may be used for the numerical simulation of PV systems power production throughout one year. Compared to the neural network model in [27], where the data of many years are required for training the model, the proposed model of this paper needs minimum measurements of one year. Compared to [30], where a two-state MC model has been developed, the proposed model of this paper uses N-state MC, which is more accurate. In addition, compared to [31], we develop an N-state MC model for each group of days with (i) cloudy, (ii) intermittent cloudy, and (iii) clear sky behavior, whereas a single N-state MC model has been developed in [31]. Compared to [32], where the synthetic data of PV systems power production have been used for analyzing the profitability/risk of financing PVs in long-term resolution by means of a model based on 10 to 15 years of real data, the model proposed in this paper is advantageous for numerical simulation in short- to mid-term analysis and it is trained by a limited amount of input measurements.

The contributions of this paper are as follows:

- We provide a comprehensive model of PV systems power production that is suitable for the numerical simulation of electrical distribution grids.
- The dependency of temporal weather variations, including one-minute, daily, and monthly variations on the PV systems power production, is considered. A model is developed to generate a sequence of PV systems power production with a random average based on the N-state MC model of that day.
- A distinct N-state MC model is proposed for each group of days characterized by cloudy, intermittent cloudy, and clear sky. The proposed model is tested on two locations with "warm and temperate" and "tropical wet and dry/savanna" climates.

1.3. Organization

The structure of the paper is organized as follows: the proposed model, which is used for generating the synthetic data of PV systems power production, is explained in Section 2. Possible applications of the proposed model are described in Section 3. The experimental results, analyses, and discussions are presented in Section 4. In addition, the proposed model is compared with the state-of-the-art one in Section 4. The conclusions and suggestions for future lines of work are given in Section 5.

2. Proposed Model

The details of the proposed model for generating synthetic data of PV systems power production are described in Figure 1. The procedure is represented in three steps, namely data preprocessing, model training, and synthetic data generation. Afterwards, as depicted in Figure 1, there is an evaluation step to analyze the performance of the proposed model.

In this study, it is assumed that we have data of measured GHI, i.e., $G(m, d, t)$, in one-minute resolution as an input of the model. The month is indexed by $m \in \{1, \ldots, 12\}$, the day is denoted by $d \in \{1, \ldots, D_m\}$, where D_m is the number of days in month m, and the time step of a day is indexed by $t \in \{1, \ldots, 1440\}$. The objective is to generate synthetic one-minute data of GHI, i.e., $\hat{G}(m, d, t)$, capturing the main features of original time-series, including the impacts of (i) one-minute, (ii) daily, and (iii) monthly weather variations on GHI. Then, the PV systems power production data are generated based on the model of the PV system, including the power production efficiency.

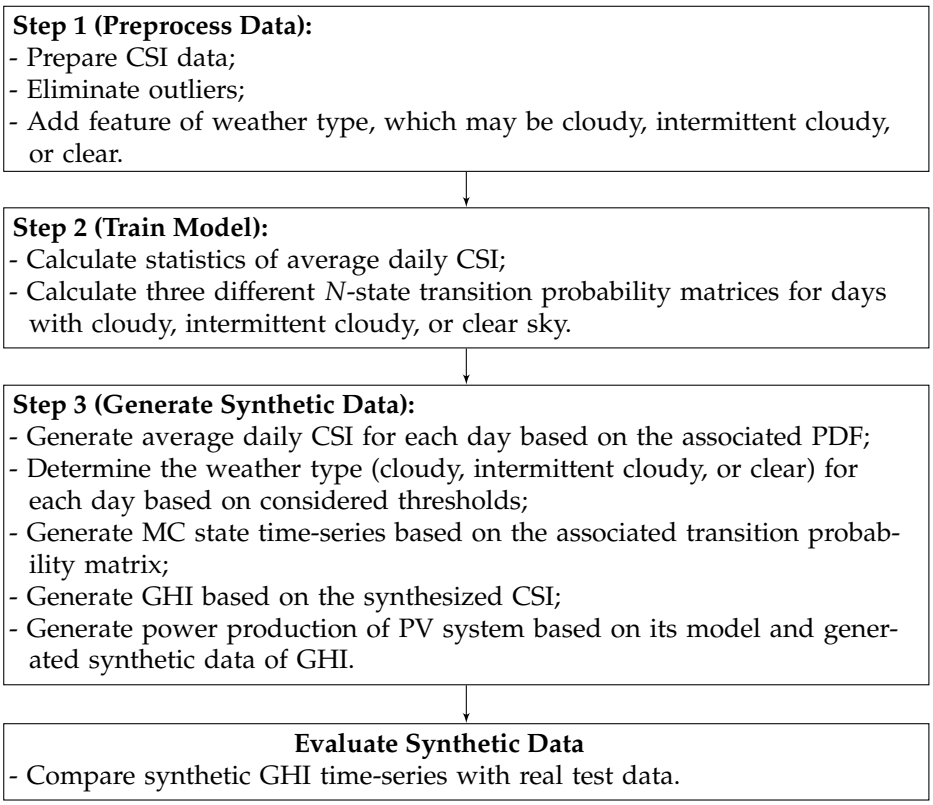


Figure 1. Flow diagram of the proposed model for generation and evaluation of synthetic data of a PV system's power production.

2.1. Step 1 (Preprocess Data)

In most cases, the input data are incomplete, inconsistent, and contain outlier errors. Furthermore, measurement equipment may have inaccuracies. In order to make the proposed model resistant to flaws in the input data, we remove the outliers by means of Tukey's test [35]. Based on this test, the tolerance band is between $\underline{\mathrm{G}}$ and $\overline{\mathrm{G}}$, where

$$\underline{\mathrm{G}} = \mathrm{G}^{\langle 0.25 \rangle} - 3.(\mathrm{G}^{\langle 0.75 \rangle} - \mathrm{G}^{\langle 0.25 \rangle}), \tag{1}$$

$$\overline{\mathrm{G}} = \mathrm{G}^{\langle 0.75 \rangle} + 3.(\mathrm{G}^{\langle 0.75 \rangle} - \mathrm{G}^{\langle 0.25 \rangle}), \tag{2}$$

$\mathrm{G}^{\langle 0.25 \rangle}$ and $\mathrm{G}^{\langle 0.75 \rangle}$ are the 0.25 quantile (25 percentile) and the 0.75 quantile (75 percentile) of the input data of measured GHI, i.e., $G(m, d, t)$, respectively.

Then, we evaluate the missing values. Since the aim is to capture temporal variations of GHI in the final model, we calculate the average of two points around a missing point and replace the missing point with it. If there are consecutive missing points, we replace them all with the average of two points around the missing points. (Note that if there are more than six consecutive missing points, we neglect the data of that specific day for building our model.)

Next, to model the stochastic temporal variations of instantaneous GHI, the CSI data are generated, defined as the uncertain variable of the solar irradiance when the impact of deterministic variations of the sun's position in the sky has been removed. Mathematically, it is defined as

$$\mathrm{CSI}(m, d, t) = \frac{G(m, d, t)}{G_c(m, d, t)}\ , \tag{3}$$

where $G_c(m, d, t)$ is the deterministic clear sky GHI over time step t of day d of month m, calculated for latitude/longitude coordinates of the PV site.

Then, we split the CSI data of each month m into training and testing datasets. We build our model using the training datasets of all months to capture the monthly weather variations of CSI.

Finally, we define the feature of day type. Each day may be clustered into cloudy, intermittent cloudy, or clear. (We may have a different number of clusters (other than three) depending on the input data. Here, we present the model for three clusters without loss of generality as we may present a similar model with another number of clusters.) To this end, we define the following rule for determining the type of each day.

$$\text{Type}(m,d) = \begin{cases} \text{Cloudy}, & \text{if } \eta(m,d) \leq \eta_{th1} \text{ and } \sigma(m,d) \leq \sigma_{th1}, \\ \text{Clear}, & \text{if } \eta(m,d) \geq \eta_{th2} \text{ and } \sigma(m,d) \geq \sigma_{th2}, \\ \text{Intermittent cloudy}, & \text{otherwise}, \end{cases} \tag{4}$$

where $\eta(m,d)$ and $\sigma(m,d)$ are the average and standard deviation of CSI(m,d,t) over day d at month m; η_{th1} and η_{th2} are two thresholds to group the day type based on the CSI average; σ_{th1} and σ_{th2} are two thresholds to group the day type based on the CSI standard deviation.

The thresholds η_{th1}, η_{th2}, σ_{th1}, and σ_{th2} are tuned such that the cloudy, intermittent cloudy, and clear days are distinguished. To this end, the following clustering problem must be solved.

$$\arg \min_{\eta_{th1},\eta_{th2},\sigma_{th1},\sigma_{th2}} \sum_{m\in\{1,2,\ldots,12\}} \sum_{d\in\{1,2,\ldots,D_m\}} \left(r_\eta^2(m,d) + r_\sigma^2(m,d) \right), \tag{5}$$

where

$$\begin{aligned} r_\eta(m,d) &= \min\{|\eta(m,d) - \tfrac{\eta_{th1}}{2}|, |\eta(m,d) - \tfrac{\eta_{th1}+\eta_{th2}}{2}|, |\eta(m,d) - \tfrac{1+\eta_{th2}}{2}|\}, \\ r_\sigma(m,d) &= \min\{|\sigma(m,d) - \tfrac{\sigma_{th1}}{2}|, |\sigma(m,d) - \tfrac{\sigma_{th1}+\sigma_{th2}}{2}|, |\sigma(m,d) - \tfrac{1+\sigma_{th2}}{2}|\}. \end{aligned} \tag{6}$$

Problem (5) is a conventional clustering problem with a predetermined number of clusters [36]. In the objective, the distances of CSI mean and standard deviation throughout the year from the clusters' centers, i.e., $\frac{\eta_{th1}}{2}$, $\frac{\sigma_{th1}}{2}$, $\frac{\eta_{th1}+\eta_{th2}}{2}$, $\frac{\sigma_{th1}+\sigma_{th2}}{2}$, $\frac{1+\eta_{th2}}{2}$, and $\frac{1+\sigma_{th2}}{2}$, are minimized.

It is worth noting that the formulation of (5) does not imply that weather variations throughout the day are ignored. The intraday weather variations are captured in the model with MC in another layer of the model described in the second step.

2.2. Step 2 (Train Model)

The average daily CSI is an uncertain variable, depending on the daily weather variation. We capture this uncertain variable by 12 different PDFs referring to the months of the year. Using parametric PDFs fitted upon the available data for the given location, the parameter $\eta(m,d)$ of each day is determined.

The selection of the most appropriate PDF family for the average daily CSI at a given location is an open research topic. In this paper, we concentrate on investigation of several PDF families taken from the relevant literature and choose those that best fit the empirical samples. The parameters of the PDFs are estimated through a Maximum Likelihood Estimation (MLE). The most adequate PDF for the specific location is then chosen through a Goodness of Fitting (GoF) evaluation score based on a Determination Coefficient (DC), i.e., R^2 [37,38].

2.2.1. Pool of Selected Probability Density Functions

Several PDFs have been considered for modeling the average daily CSI. For the sake of brevity, six of them are presented here, which are more precise for this type of data. Three of them (Beta, Kumaraswamy (Kuma) and Logit-Normal (LogitN)) are suitable for fitting random variables bounded within $[0,1]$, which is the typical range of average daily CSI values [13,39]. The other three (Logistic (Logist), Log-Logistic (Log-Log), and Generalized

Extreme Value (GEV)) are suitable instead for modeling unbounded random variables; however, they may also be useful to characterize bounded variables such as average daily CSI [40]. The formulations of the PDFs are briefly recalled below.

(i) The Beta PDF is

$$f_{Beta}(x|a,b) = \frac{x^{a-1}\cdot(1-x)^{b-1}\cdot\Gamma(a+b)}{\Gamma(a)\cdot\Gamma(b)}, 0 < x < 1, a > 0, \text{ and } b > 0, \tag{7}$$

where a and b are shape parameters.

(ii) The Kumaraswamy PDF is

$$f_{Kuma}(x|c,d) = c\cdot d\cdot x^{c-1}\cdot(1-x^c)^{d-1}, 0 < x < 1, c > 0, \text{ and } d > 0, \tag{8}$$

where c and d are shape parameters as well.

(iii) The Logit-Normal PDF is

$$f_{LogitN}(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\frac{1}{x(1-x)}e^{-\frac{(logit(x)-\mu)^2}{2\sigma^2}}, 0 < x < 1, \mu \in \mathbb{R}, \text{ and } \sigma > 0, \tag{9}$$

where μ and σ are the average and standard deviation of normally distributed $logit(x)$.

(iv) The Logistic PDF is

$$f_{Logist}(x|\lambda,s) = \frac{1}{s\left(1+e^{-\frac{x-\lambda}{s}}\right)^2}e^{-\frac{x-\lambda}{s}}, x \in \mathbb{R}, \lambda \in \mathbb{R}, \text{ and } s > 0, \tag{10}$$

where λ and s are the average and scale parameter, respectively.

(v) The Log-Log PDF is related to the Logistic PDF. The logarithm of generated variable by Log-Log has Logistic PDF. The Log-Log PDF is

$$f_{Log-Log}(x|\tau,v) = \frac{1}{v}\frac{1}{x}\frac{e^z}{(1+e^z)^2}, z = \frac{log(x)-\tau}{v}, x \in \mathbb{R}, \tau > 0, \text{ and } v > 0, \tag{11}$$

where μ and v are the Log average and Log scale parameter.

(vi) The GEV PDF is

$$f_{GEV}(x|k,\rho,c) = \frac{1}{c}e^{-\left(1+k\frac{x-\rho}{c}\right)^{-\frac{1}{k}}}\left(1+k\frac{x-\rho}{c}\right)^{-1-\frac{1}{k}}, \rho \in \mathbb{R}, k \in \mathbb{R}, \text{ and } c \geq 0\ , \tag{12}$$

where ρ is the location parameter, c is the scale parameter, and k is the shape parameter.

2.2.2. Parameter Estimation of Probability Density Function and Goodness of Fitting Test

The parameters of the PDFs are estimated through an MLE, differentiated for each month m. For example, the parameters $a(m)$ and $b(m)$ of the Beta PDF are estimated by solving the following optimization problem:

$$a(m), b(m) = \arg\max_{a,b} \Pi_{d\in\{1,2,\ldots,D_m\}} f_{Beta}(\eta(m,d)|a,b)\ , \tag{13}$$

that maximizes the likelihood of the monthly data on the considered PDF. It is trivially extended to all other PDFs for consistency.

The DC is calculated for each PDF to assess the GoF score quantitatively. The DC is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{B}(N_{bi}-p_{bi})^2}{\sum_{i=1}^{B}(N_{bi}-\overline{N}_{bi})^2}\ , \tag{14}$$

where B is the number of bins in which the domain of the random variable is divided for the GoF score assessment, p_{bi} is the probability of the random variable to lie within the bin determined from the estimated PDF, N_{bi} is the number of observed samples within the bth bin, and $\overline{N}_{bi} = \frac{1}{B}\sum_{i=1}^{B} N_{bi}$. A greater value of the DC determines better fit and indicates preferable PDF selection.

2.2.3. Markov Chain Training

For each group of days, an N-state (It is important to note that the number of states, i.e., N, must be less than $1/\mathrm{CSI}_{res}$, where CSI_{res} is the maximum possible resolution at which we can measure CSI.) MC model is developed. Each model, developed for a day type, is trained on data including CSI time-series. We discretize the data based on the number of states, i.e., N. Thus, we define the State of CSI (SCSI) as follows:

$$\mathrm{SCSI}(m,d,t) = \begin{cases} 1, & \text{if } 0 \leq \mathrm{CSI}(m,d,t) < \frac{1}{N}, \\ 2, & \text{if } \frac{1}{N} \leq \mathrm{CSI}(m,d,t) < \frac{2}{N}, \\ \dots & \\ i, & \text{if } \frac{i-1}{N} \leq \mathrm{CSI}(m,d,t) < \frac{i}{N}, \\ \dots & \\ N, & \text{if } \frac{N-1}{N} \leq \mathrm{CSI}(m,d,t) < \frac{N}{N}. \end{cases} \tag{15}$$

Based on this configuration of states, the transition matrix is calculated as

$$P(\mathrm{Type}) = \left[\frac{n_{ij}}{\sum_{k=1}^{N} n_{ik}}\right], \tag{16}$$

where n_{ij} is the number of transitions from state i to state j from one time step to the next. It is worth mentioning that the transition matrix $P(\mathrm{Type})$ is a function of the day type (cloudy, intermittent cloudy, or clear). Thus, three transition matrices $P(\mathrm{Type})$ are computed using three separate sets of data for days with cloudy, intermittent cloudy, and clear sky. To compute n_{ij}, the data of the day type are built. Then, the number of transitions from state i to state j in these data are enumerated.

2.3. Step 3 (Generate Synthetic Data)

Once the models for the three mentioned clusters of days are trained, they are used to generate the synthetic data with the following procedure.

First of all, for generating the data of day d of month m, we have to decide on the average weather of that day, whether the sky is cloudy, intermittent cloudy, or clear. To this end, we generate a random number using the PDF with the associated parameters of month m. This generated random number is the average $\eta(m,d)$. Then, using the computed thresholds resulting from the optimization problem (5), the type of day d of month m is determined.

Next, for that day, we consider the initial $\mathrm{CSI}(m,d,0)$ equal to the random number $\eta(m,d)$ generated by the PDF. We use the associated transition probability matrix of that day type for calculating the state transition for the next time steps until the whole 1440 time steps are generated. It is worth mentioning that with an initial value $\mathrm{CSI}(m,d,0)$ equal to the random number $\eta(m,d)$ and generating the time-series based on the transition matrix, the average CSI of that day would be $\eta(m,d)$ [31].

2.4. Evaluate Synthetic Data

Upon generating the synthetic data of CSI, the GHI data and PV systems power production data are generated using the CSI model of GHI and the model of the PV system. The generated synthetic GHI is compared with the real test data in terms of:

- similarity of PDFs in different months to evaluate the impacts of monthly weather variation on GHI;
- TAF similarity of average daily GHI for evaluating the impacts of daily weather variation on GHI; and
- TAF similarity of GHI time-series for evaluating the impacts of one-minute-resolution weather variations.

By comparing the standard TAF for real and synthetic data, three mentioned similarities are evaluated. The standard TAF is defined as

$$\mathrm{TAF}(\tau) = \mathbb{E}\left[\frac{(\mathrm{CSI}(m,d,t) - \hat{\mu}) \cdot (\mathrm{CSI}(m,d,t+\tau) - \hat{\mu})}{\hat{\sigma}^2}\right], \tag{17}$$

where $\hat{\sigma}$ is the standard deviation of CSI over the year, τ is the lag of TAF in minutes, and $\hat{\mu}$ is the average of CSI over the whole data. In addition, $\mathbb{E}[.]$ is the expectation operator. Since $\mathrm{TAF}(\tau)$ does not capture monthly and daily variations, the standard TAF is not the correct index for our study.

To validate the performance of the proposed model, we define the following metric, which we refer to as the modified version of TAF.

$$\mathrm{MTAF}(\tau) = \mathbb{E}\left[\frac{(\mathrm{CSI}(m,d,t) - \eta(m,d)) \cdot (\mathrm{CSI}(m,d,t+\tau) - \eta(m,d))}{\sigma^2(m,d)}\right], \tag{18}$$

where $\eta(m,d)$ and $\sigma(m,d)$ are the average and standard deviation of CSI at day d of month m, respectively.

The difference between this modified version of TAF and the standard one is that in the modified version, the average daily CSI, i.e., $\eta(m,d)$, and the standard deviation of CSI at each day, i.e., $\sigma(m,d)$, are used, whereas the average $\hat{\mu}$ and standard deviation $\hat{\sigma}$ over the whole year are used in the standard TAF. Since the model of each month for weather variations and the transitional behavior of each day are different, the modified version of TAF is the best representative index for evaluating the synthetic data introduced in this study.

2.5. Benchmark Model

To evaluate the performance, the proposed model is compared with the state-of-the-art model in [31]. The procedure for generating synthetic data based on this model includes three steps: (i) data preprocessing; (ii) model training; and (iii) synthetic data generation. In the data preprocessing step, the CSI data are prepared and the outliers are eliminated. In the model training step, a single transition probability matrix is calculated with N-state for all days. In the synthetic data generation step, the MC state time-series is generated based on the calculated transition probability matrix. Finally, the GHI and power production of PV systems are calculated based on the synthesized CSI.

3. Exemplary Applications of the Proposed Model

A possible application of the proposed model is Probabilistic Load Flow (PLF) analysis for studying the uncertainties of voltage/current profiles. The PV systems power production and other uncertain inputs such as Electric Vehicle (EV) charging and load demand are the electrical distribution grid's variables, whose models are useful for numerical simulation purposes. Figure 2 depicts a typical illustration of the process of PLF analysis. The PLF simulation requires a number of synthetic samples, extracted on the basis of predefined PDFs and TAFs. Then, the PDF of voltage and current magnitudes of medium-voltage (MV) and low-voltage (LV) distribution grids are determined using the power flow equations.

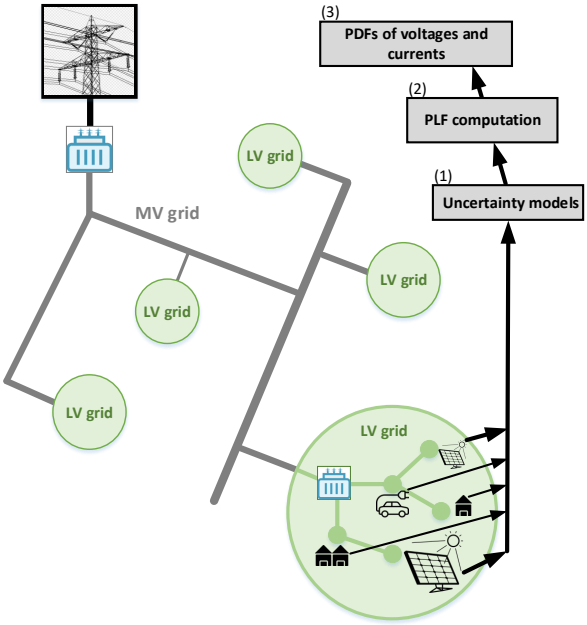


Figure 2. Process of PLF analysis in electrical distribution grids.

Another application of the proposed model is the validation of developed algorithms in future electrical distribution grids. Such an application is investigated in the project "Grid Data Digger" [41]. The innovative aspect of this project is that it valorizes available data to system operators through a data-driven and automated process. It particularly aims at providing to the system operators all necessary information and analysis for a secure and optimal operation of the electrical distribution grids. It is achieved by developing a big-data platform with descriptive, diagnostic, predictive, and preventive data analysis applications dedicated to the electrical distribution grid's operation. A part of this emulated big-data platform is the modeling of PV systems power production, which is based on the proposed N-state MC model.

4. Experimental Results

4.1. Input Data

To test the proposed procedure, the measured data of a PV system installed at Yverdon-les-Bains, Switzerland (location (a)) were used. In addition, the proposed model was tested on the data obtained from the National Renewable Energy Laboratory (NREL) radiometer array in Oahu, Hawaii, USA (location (b)) [42].

The PV system at location (a) is a solar carport, i.e., a car shelter equipped with PV systems and with charging stations as illustrated in Figure 3. In total, 85 PV systems Bisol 295 W are installed. Twelve of them are dedicated to smart home production (i.e., 3540 W) and the rest is connected to a smart-grid laboratory via three-phase converters with the possibility to generate capacitive or inductive current. This infrastructure is used for validating data-driven decision-making such as PLF analysis.

The input data include one-minute time-series of measured GHI from 1 January 2016 to 31 December 2017 at location (a) and one-minute time-series of measured GHI from 1 April 2010 to 31 October 2011 at location (b). In addition, using the "pvlib" package [43] in "Python", we obtained the clear sky Ineichen model of GHI, i.e., $G_c(m, d, t)$, for the same locations and timings. Based on (3), we calculated $\mathrm{CSI}(m, d, t)$ for the time-series. The data of 2016 were used for training the model of location (a) and the data of 2017 were used as a testing dataset. For location (b), the data of 1 April 2010 to 31 March 2011 were used for training and the rest was used for testing.

Figure 3. View of solar carport at location (a).

4.2. Model Training on Data of Location (a)

The DC values calculated on the estimated PDFs for the average daily CSI at location (a) are shown in Table 1; bold values indicate the best fit for each month. The Beta, LogitN, and Kuma PDFs are the best fit in eleven of twelve months, with the GEV being instead the most suitable PDF for the remaining month.

The (a) PDFs and (b) Cumulative Density Functions (CDFs) of the average daily CSI for the months of February and June are illustrated in Figures 4 and 5, respectively. The figures include the Beta, LogitN, Kuma, and GEV PDFs. Note that the Beta and Kuma PDFs and CDFs practically overlapped due to the theoretical similarity between the two PDFs.

Table 1. DC of the estimated PDFs for the average CSI at location (a).

Month	DC					
	Beta	**LogitN**	**Kuma**	**Logist**	**Loglog**	**GEV**
Jan.	0.9481	**0.9541**	0.9515	0.8664	0.9288	0.8974
Feb.	0.9490	**0.9574**	0.9503	0.8954	0.9173	0.8921
Mar.	0.9854	0.9803	**0.9856**	0.9485	0.9236	0.9454
Apr.	0.9518	**0.9763**	0.9384	0.8877	0.9494	0.9150
May	**0.9917**	0.9845	0.9916	0.9493	0.9291	0.9590
Jun.	0.9796	0.9591	**0.9813**	0.9609	0.8950	0.9591
Jul.	0.9911	0.9810	**0.9912**	0.9667	0.9268	0.9648
Aug.	**0.9725**	0.9669	0.9713	0.9317	0.8493	0.9289
Sep.	**0.9814**	0.9812	0.9803	0.9295	0.9211	0.8068
Oct.	0.9706	**0.9757**	0.9679	0.9026	0.8676	0.8902
Nov.	0.9234	**0.9639**	0.9274	0.8854	0.9389	0.9236
Dec.	0.8260	0.8924	0.8339	0.7704	0.8855	**0.9307**

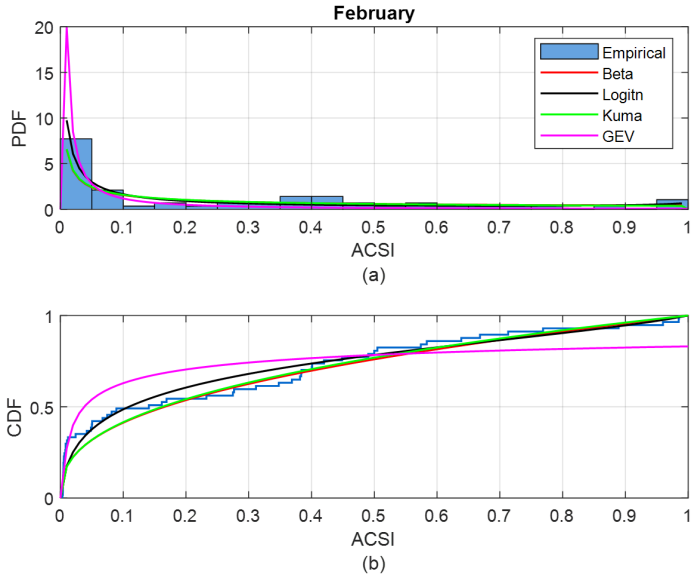


Figure 4. Comparing the pool of different models for average daily CSI: (**a**) PDFs and (**b**) CDFs at location (a) in February.

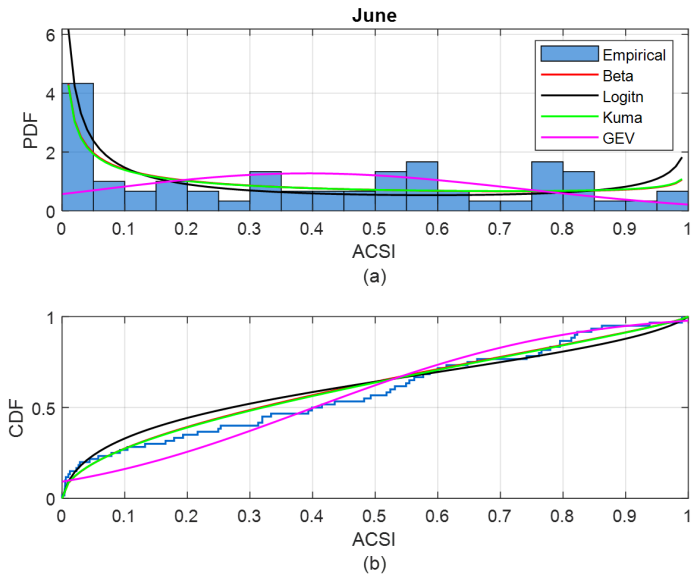
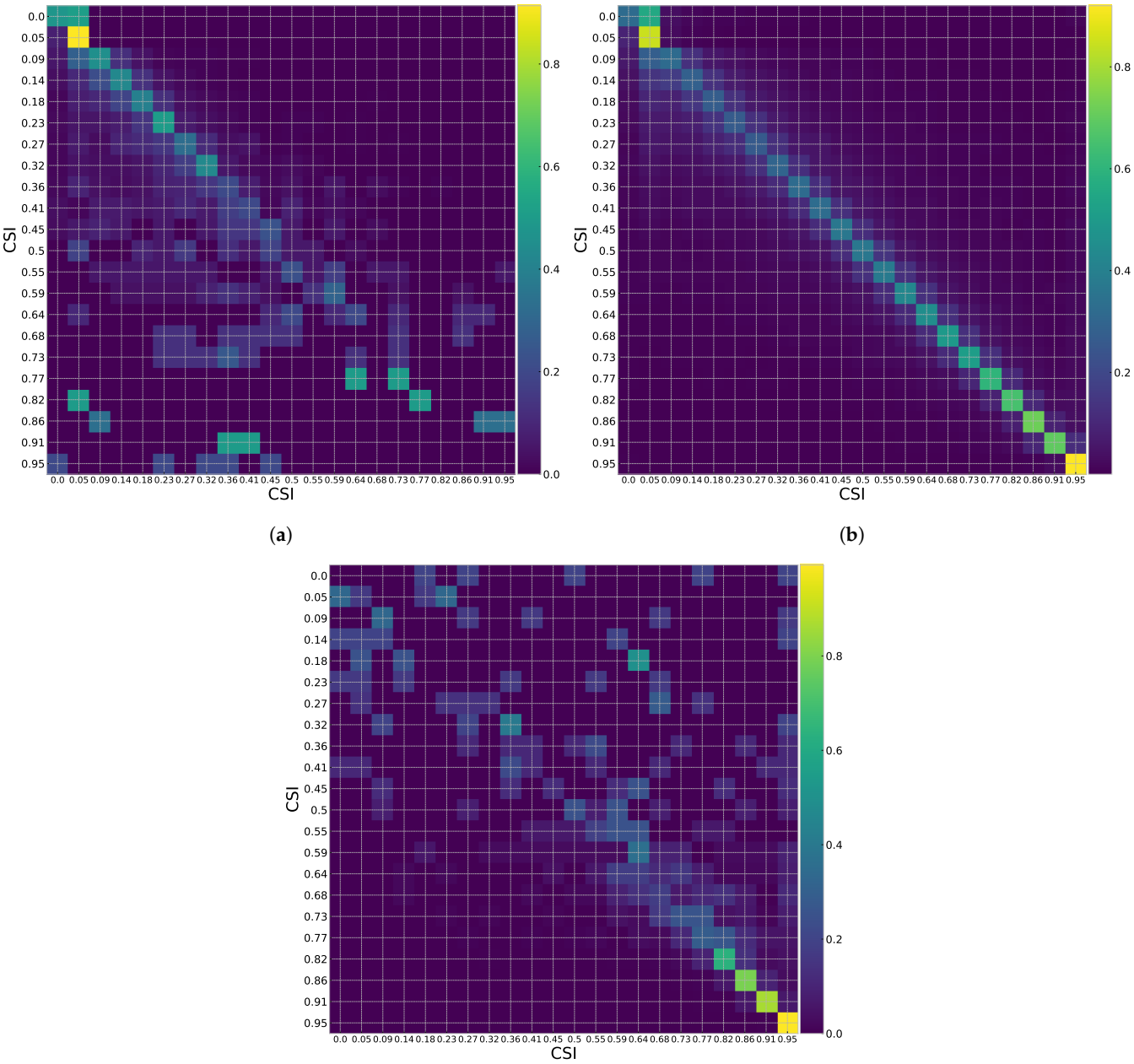


Figure 5. Comparing the pool of different models for average daily CSI: (**a**) PDFs and (**b**) CDFs at location (a) in June.

Three transition matrices are calculated with $N = 21$, i.e., the number of states is equal to 21. The transition matrices are presented in Figure 6a–c with colorbar plots for days with cloudy, intermittent cloudy, and clear sky, respectively. We use the "quantecon" package in "Python" for producing the transition probabilities from the input measured data.

The sum of probabilities of each column and row of transition matrices in Figure 6a–c must add up to unity based on the Markov requirement for transition probabilities. In Figure 6a, the state of CSI for the values less than 0.33 is persistent since the CSI is close to zero. On the other hand, the state of CSI for the days with clear sky is persistent for values more than 0.81. Finally, by observing the transition probabilities for the days with intermittent cloudy sky, one may see that the CSI transition within zero to one is possible. It is noticeable that strong persistence of CSI, i.e., high values of the diagonal elements and of those close to the diagonal elements in the transition matrix, is expectable for one-minute data since the weather variations in one-minute resolution are not considerable.



(**a**) (**b**)

(**c**)

Figure 6. Surface plot for transition matrices of days with (**a**) cloudy sky, (**b**) intermittent cloudy sky, and (**c**) clear sky.

We calculated one transition matrix with $N = 21$ for the whole dataset based on the state-of-the-art model in [31]. The calculated transition matrix is presented in Figure 7. Here, we do not distinguish among the days with cloudy, intermittent cloudy, and clear sky. In addition, one model has been developed for the whole year and the monthly weather variations have been neglected. This state-of-the-art model is used as a benchmark in this study.

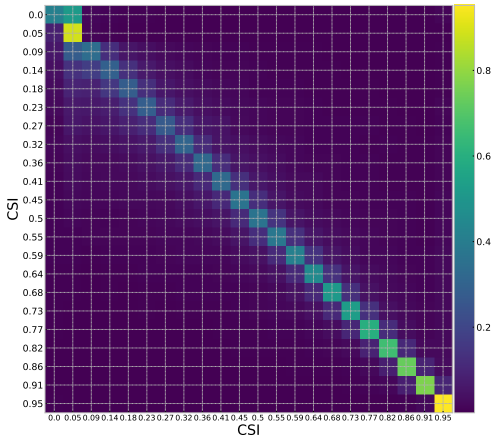


Figure 7. Surface plot for transition matrix of the state-of-the-art model [31].

4.3. Synthetic Data Generation at Location (a)

We generated synthetic data of CSI and GHI based on the two models (i.e., our proposed model and the state-of-the-art models in [31]). The data of real GHI and the synthetic GHI based on the two models for month June 2017 are illustrated in Figure 8a–c with red lines. In addition, the CSI model of GHI, i.e., $G_c(m, d, t)$, for month June 2017 is depicted with blue lines in Figure 8a–c.

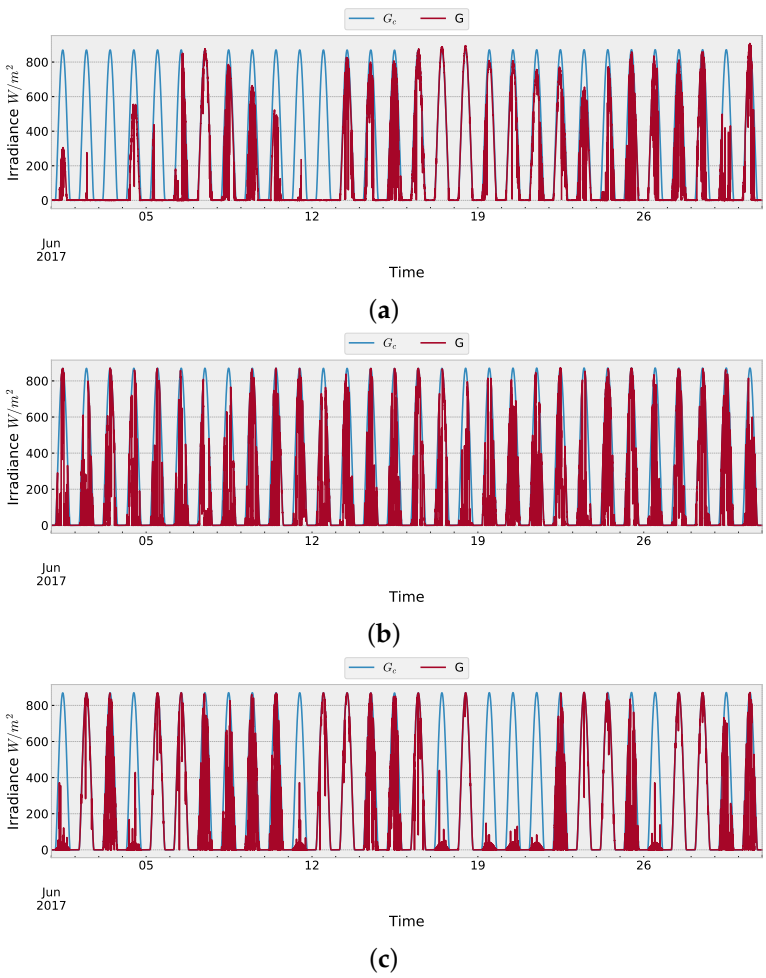


Figure 8. (**a**) Real data of G and G_c for month June 2017; (**b**) Synthetic data of G and G_c for month June 2016 based on state-of-the-art model [31]; (**c**) Synthetic data of G and G_c for month June 2016 based on our proposed model.

By comparing the real data with the two synthetic data visually, one may observe that the daily weather variations are not captured in the state-of-the-art model as all days behaved similarly. However, in our proposed model, some days have small CSI as they are cloudy and some of them have persistently high CSI as they are days with clear sky. It is worth mentioning that the CSI at each time step is calculated as the ratio of $G(m, d, t)$, given by the red lines, and $G_c(m, d, t)$, given by the blue lines in Figure 8a–c.

4.4. Evaluation of Synthetic Data of Location (a)

In Figure 9a,b, the standard TAF and the defined performance index (modified version of the TAF) are presented, respectively, for synthetic data based on the state-of-the-art model [31] and the proposed model of this paper. From the viewpoint of the modified metric, the synthetic data of the proposed model have a TAF closer to the real data. As one may see in Figure 9b, the modified TAF of the synthetic data based on the proposed model perfectly matches the real data. This was expected, since three different transition matrices were developed for days with cloudy, intermittent cloudy, and clear sky, while in the state-of-the-art model [31], one transition matrix was developed for the whole year data. In addition, the monthly weather variations are not considered in the state-of-the-art model. Finally, the introduced modified version of TAF captures all mentioned temporal variabilities of the data for comparison.

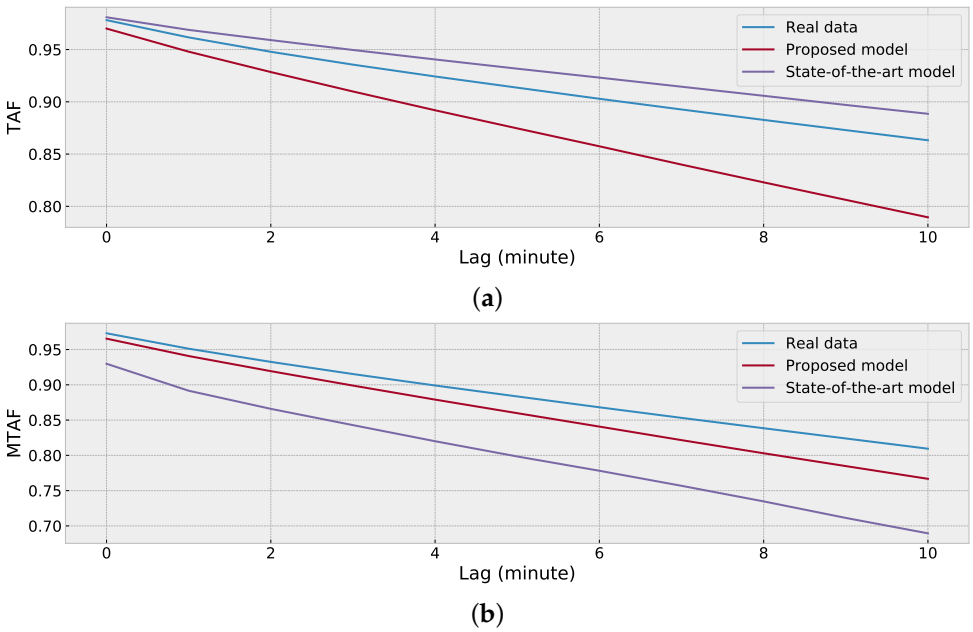


Figure 9. Comparing the proposed model, state-of-the-art model, and real data: (**a**) Standard TAF; (**b**) Modifed TAF.

4.5. Model Training on Data of Location (b)

Results of the model training on the location (b) data are presented in a compact form for conciseness. The DC values calculated on the estimated PDFs for the average daily CSI at the location (b) are shown in Table 2; bold values indicate the best fit for each month. The GEV is the best fit in four months, followed by the Beta (three months) and by the LogitN, Kuma, and Logist PDFs (two months each). Compared to the outcome evidence for location (a), the different statistical behavior of the average daily CSI at location (b) is due to the generally better weather conditions. This is particularly clear by observing the (a) PDFs and (b) CDFs of the average daily CSI for June, which are illustrated in Figure 10. The figure includes the Beta, LogitN, Kuma, and GEV PDFs only, for readability.

Table 2. DC of the estimated PDFs for the average CSI at location (b).

Month	DC					
	Beta	**LogitN**	**Kuma**	**Logist**	**Loglog**	**GEV**
Jan.	0.9534	0.9695	0.9612	0.9604	0.9196	**0.9756**
Feb.	0.9697	0.9726	0.9689	0.9562	0.9508	**0.9748**
Mar.	0.9275	0.9433	0.9477	**0.9577**	0.9144	0.9572
Apr.	**0.9889**	0.9870	0.9871	0.9828	0.9798	0.9858
May	0.9726	0.9816	0.9824	0.9770	0.9561	**0.9940**
Jun.	0.9851	0.9887	**0.9931**	0.9896	0.9786	0.9871
Jul.	**0.9948**	0.9944	0.9933	0.9915	0.9872	0.9948
Aug.	**0.9945**	0.9932	0.9924	0.9930	0.9903	0.9877
Sep.	0.9927	**0.9929**	0.9921	0.9888	0.9826	0.9896
Oct.	0.7810	0.6872	0.8091	**0.9905**	0.7960	0.9898
Nov.	0.9652	0.9749	**0.9825**	0.9769	0.9592	0.9779
Dec.	0.9239	**0.9405**	0.9233	0.9245	0.8780	0.9305

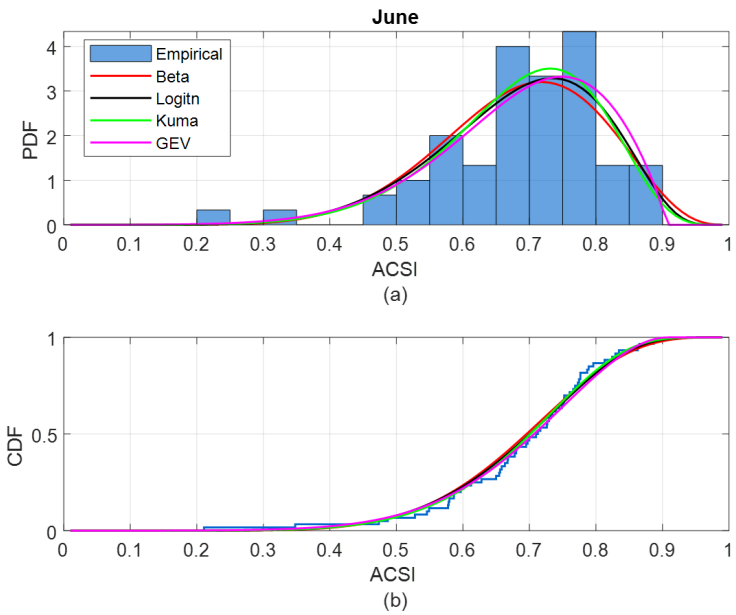


Figure 10. Comparing the pool of different models for average daily CSI: (**a**) PDFs and (**b**) CDFs at location (b) in June.

The real data of GHI for month June 2011, the synthetic data based on the state-of-the-art model, and the synthetic data based on the proposed model are depicted in Figure 11a–c, respectively. As one may see, there is not a daily weather variation in the real data due to the climatic condition. In other words, the weather type of most of the days is classified as intermittent cloudy in location (b). As a result, the advantage of using our proposed model is marginal compared to the state-of-the-art model. In Figure 12a,b, the standard TAF and the defined performance index (modified version of the TAF) are presented, respectively, for the data of location (b).

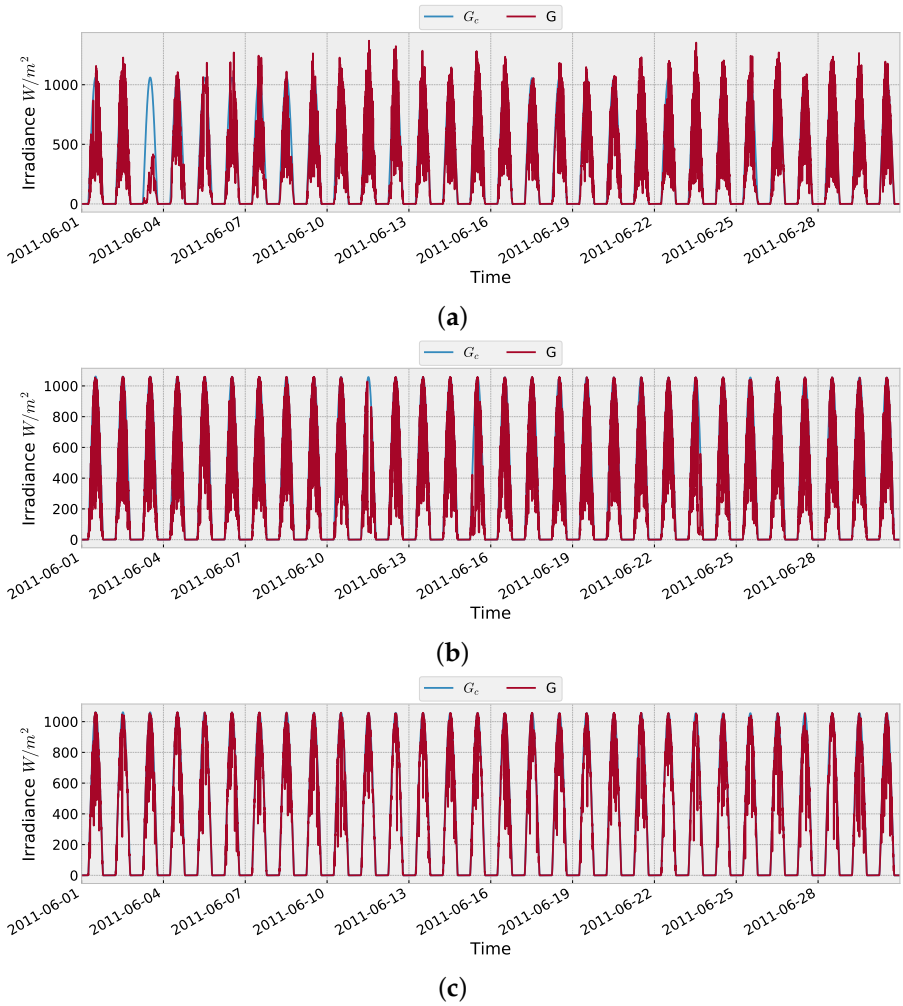


Figure 11. Comparing real data and synthetic data generated by the proposed model and state-of-the-art one: (**a**) Real data of G and G_c for month June 2011 in location (b); (**b**) Synthetic data of G and G_c for month June 2016 based on state-of-the-art model [31]; (**c**) Synthetic data of G and G_c for month June 2016 based on our proposed model.

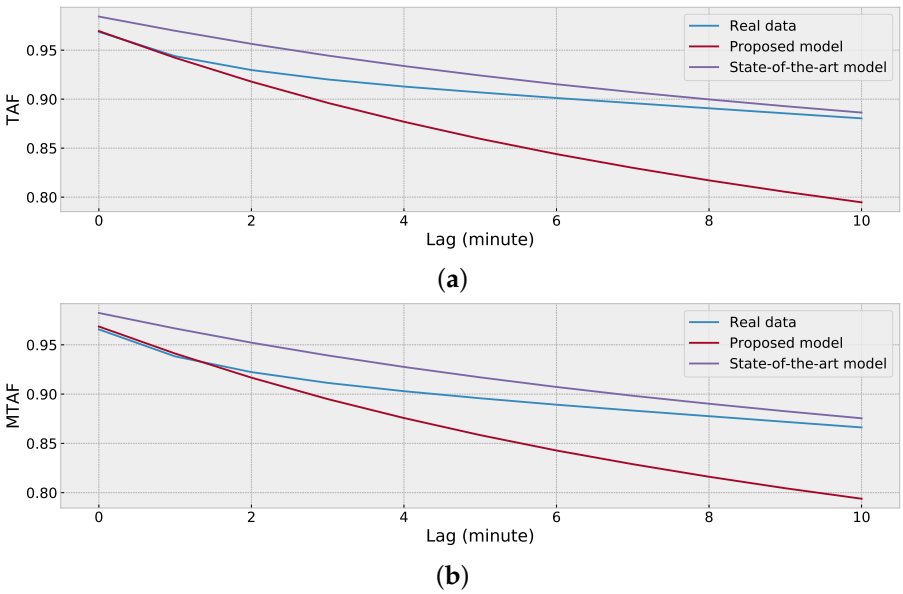


Figure 12. Comparing the proposed model, state-of-the-art model, and real data: (**a**) Standard TAF; (**b**) Modifed TAF.

5. Conclusions and Future Works

In this work, we present a model for the synthetic generation of data of PV systems power production for the numerical simulation of electrical distribution grids. The generated synthetic data are supposed to be used in the project "Grid Data Digger" for testing the data-driven algorithms in future electrical distribution grids. The proposed model is based on three different N-state MCs and a hierarchy of different temporal-scale weather variations, i.e., one-minute, daily, and monthly. The synthetic data based on the proposed model have high similarity with the real data in terms of PDF and TAF metrics compared to the synthetic data based on the state-of-the-art model. The added value of the proposed model is that it considers the seasonal, intra-month, and intra-day variability of the solar irradiance. Therefore, the proposed model is particularly suitable for the numerical simulation of electrical distribution grids.

In future works, the spatiotemporal correlation of CSI can be modeled using the real data of more locations. The Heliosat model [44], which is based on satellite data processing instead of local measuring stations, can be used to repeat the evaluation of proposed models for different geographical positions. In addition, the yearly weather variations can be addressed in future models using the long-term data retrieved from the satellite estimations. Finally, future studies may look into the effects of other meteorological factors such as ambient temperature and wind velocity on the efficiency of PV systems for generating synthetic datasets.

Author Contributions: Conceptualization, M.R. and P.D.F.; methodology, M.R.; software, M.R.; validation, M.R., D.P. and M.B.; resources, M.C.; writing—original draft preparation, M.R.; writing—review and editing, P.D.F., D.P. and M.B.; project administration, M.B.; funding acquisition, M.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: To ensure full reproducibility and to facilitate the future uptake of the proposed model, the Python code and data of experimental results section are accessible from the following link: https://github.com/mohammadrayati/CSI_MC_Model (accessed on 2 August 2021).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

BSRN	Baseline Surface Radiation Network
COP21	21st Conference of Parties
CSI	Clear Sky Index
DC	Determination Coefficient
EV	Electric Vehicle
GEV	Generalized Extreme Value
GHI	Global Horizontal Irradiance
GoF	Goodness of Fitting
HMM	Hidden Markov Model
Kuma	Kumaraswamy
LogitN	Logit-Normal
Logist	Logistic
Log-Log	Log-Logistic
LV	Low-Voltage
MC	Markov Chain
MLE	Maximum Likelihood Estimation
MTAF	Modified Temporal Autocorrelation Function
MV	Medium-Voltage
PDF	Probability Density Function

PLF	Probabilistic Load Flow
PV	Photovoltaic
SCSI	State of Clear Sky Index
SFOE	Swiss Federal Office of Energy
TAF	Temporal Autocorrelation Function
WRMC	World Radiation Monitoring Center

Nomenclature

Indices

m	Month
d	Day
t	Minute time-step
τ	Lag of TAF in minutes

Functions

$f_{Beta}(x\|a,b)$	Beta PDF
$f_{Kuma}(x\|c,d)$	Kumaraswamy PDF
$f_{LogitN}(x\|\mu,\sigma)$	Logit-Normal PDF
$f_{Logist}(x\|\lambda,s)$	Logist PDF
$f_{Log-Log}(x\|\tau,v)$	Log-Log PDF
$f_{GEV}(x\|k,\rho,c)$	GEV PDF
$\mathrm{TAF}(\tau)$	Standard TAF
$\mathrm{MTAF}(\tau)$	Modified version of TAF

Variables

$G(m,d,t)$	Measured GHI at month m, day d, and time step t
$\hat{G}(m,d,t)$	Synthetic GHI at month m, day d, and time step t
$G_c(m,d,t)$	Clear sky GHI at month m, day d, and time step t
$\mathrm{CSI}(m,d,t)$	Clear Sky Index at month m, day d, and time step t
$\mathrm{SCSI}(m,d,t)$	State of CSI at month m, day d, and time step t
$\underline{G}$, $\overline{G}$	Minimum and maximum of tolerance band
$G^{<0.25>}$, $G^{<0.75>}$	0.25 and 0.75 quantile of the measured GHI
$\mathrm{Type}(m,d)$	Type of day d of month m
$\eta(m,d)$	Average of CSI over day d at month m
$\sigma(m,d)$	Standard deviation of CSI over day d at month m
$r_\eta(m,d)$	Residual of CSI average from the centers of clusters
$r_\sigma(m,d)$	Residual of CSI standard deviation from the centers of clusters
$P(\mathrm{Type})$	Transition matrix of a day type

Parameters

D_m	Number of days in month m
η_{th1}, η_{th2}	Average thresholds to cluster the days
σ_{th1}, σ_{th2}	Standard deviation thresholds to cluster the days
a, b	Shape parameters of Beta PDF
c, d	Shape parameters of Kumaraswamy PDF
μ, σ	Average and standard deviation of of Logit-N PDF
λ, s	Average and scale parameter of of Logist PDF
τ, v	Log average and log scale parameter of Log-Log PDF

References

1. Rhodes, C.J. The 2015 Paris climate change conference: COP21. *Sci. Prog.* **2016**, *99*, 97–104. [CrossRef]
2. SFOE. *Potenzialabschätzung für Sonnenkollektoren im Wohngebäudepark*; Swiss Federal Office of Energy: Bern, Switzerland, 2018.
3. Hänni, J. Energy Transition in Switzerland. In *Energy Law and Economics*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 43–57.

4. Ramadhani, U.H.; Shepero, M.; Munkhammar, J.; Widén, J.; Etherden, N. Review of probabilistic load flow approaches for power distribution systems with PV generation and electric vehicle charging. *Int. J. Electr. Power Energy Syst.* **2020**, *120*, 106003. [CrossRef]
5. Bahaidarah, H.M.; Tanweer, B.; Gandhidasan, P.; Rehman, S. A combined optical, thermal and electrical performance study of a V-trough PV system—Experimental and analytical investigations. *Energies* **2015**, *8*, 2803–2827. [CrossRef]
6. Gunal, M.M. *Simulation for Industry 4.0*; Springer: Berlin/Heidelberg, Germany, 2019.
7. Mueller, S.C.; Remund, J.; Meteotest, A. Validation of the Meteonorm satellite irradiation dataset. In Proceedings of the 35th European Photovoltaic Solar Energy Conference and Exhibition, Brussels, Belgium, 24–28 September 2018; pp. 24–27.
8. Uniyal, A.; Sarangi, S. Optimal network reconfiguration and DG allocation using adaptive modified whale optimization algorithm considering probabilistic load flow. *Electr. Power Syst. Res.* **2020**, *192*, 106909. [CrossRef]
9. Da Silva, A.M.L.; de Castro, A.M. Risk assessment in probabilistic load flow via Monte-Carlo simulation and cross-entropy method. *IEEE Trans. Power Syst.* **2018**, *34*, 1193–1202. [CrossRef]
10. Prusty, B.R.; Jena, D. A critical review on probabilistic load flow studies in uncertainty constrained power systems with PV generation and a new approach. *Renew. Sustain. Energy Rev.* **2017**, *69*, 1286–1302. [CrossRef]
11. Engerer, N.; Mills, F. KPV: A clear-sky index for PVs. *Sol. Energy* **2014**, *105*, 679–693. [CrossRef]
12. Graham, V.; Hollands, K. A method to generate synthetic hourly solar radiation globally. *Sol. Energy* **1990**, *44*, 333–341. [CrossRef]
13. Sulaiman, M.Y.; Oo, W.H.; Abd Wahab, M.; Zakaria, A. Application of Beta distribution model to Malaysian sunshine data. *Renew. Energy* **1999**, *18*, 573–579. [CrossRef]
14. Ettoumi, F.Y.; Mefti, A.; Adane, A.; Bouroubi, M. Statistical analysis of solar measurements in Algeria using Beta distributions. *Renew. Energy* **2002**, *26*, 47–67. [CrossRef]
15. Kabir, M.N.; Mishra, Y.; Bansal, R. Probabilistic load flow for distribution systems with uncertain PV generation. *Appl. Energy* **2016**, *163*, 343–351. [CrossRef]
16. Dellino, G.; Laudadio, T.; Mari, R.; Mastronardi, N.; Meloni, C.; Vergura, S. Energy production forecasting in a PV plant using transfer function models. In Proceedings of the 2015 IEEE 15th International Conference on Environment and Electrical Engineering (EEEIC), Rome, Italy, 10–13 June 2015; pp. 1379–1383.
17. Biga, A.; Rosa, R. Statistical behaviour of solar irradiation over consecutive days. *Sol. Energy* **1981**, *27*, 149–157. [CrossRef]
18. Herrero, M. Autocorrelation coefficient of the daily solar irradiation series in Spain. *Int. J. Ambient Energy* **1995**, *16*, 11–16. [CrossRef]
19. Jain, P.; Lungu, E. Stochastic models for sunshine duration and solar irradiation. *Renew. Energy* **2002**, *27*, 197–209. [CrossRef]
20. Bálint, R.; Fodor, A.; Magyar, A. Model-based Power Generation Estimation of Solar Panels using Weather Forecast for Microgrid Application. *Acta Polytech. Hung.* **2019**, *16*, 149–165.
21. Bright, J.; Smith, C.; Taylor, P.; Crook, R. Stochastic generation of synthetic minutely irradiance time series derived from mean hourly weather observation data. *Sol. Energy* **2015**, *115*, 229–242. [CrossRef]
22. Bright, J.M.; Babacan, O.; Kleissl, J.; Taylor, P.G.; Crook, R. A synthetic, spatially decorrelating solar irradiance generator and application to a LV grid model with high PV penetration. *Sol. Energy* **2017**, *147*, 83–98. [CrossRef]
23. Frimane, A.; Soubdhan, T.; Bright, J.M.; Aggour, M. Nonparametric Bayesian-based recognition of solar irradiance conditions: Application to the generation of high temporal resolution synthetic solar irradiance data. *Sol. Energy* **2019**, *182*, 462–479. [CrossRef]
24. Frimane, Â.; Bright, J.M.; Yang, D.; Ouhammou, B.; Aggour, M. Dirichlet downscaling model for synthetic solar irradiance time series. *J. Renew. Sustain. Energy* **2020**, *12*, 063702. [CrossRef]
25. Driemel, A.; Augustine, J.; Behrens, K.; Colle, S.; Cox, C.; Cuevas-Agulló, E.; Denn, F.M.; Duprat, T.; Fukuda, M.; Grobe, H.; et al. Baseline Surface Radiation Network (BSRN): Structure and data description (1992–2017). *Earth Syst. Sci. Data* **2018**, *10*, 1491–1501. [CrossRef]
26. Munkhammar, J.; Widén, J.; Hinkelman, L.M. A copula method for simulating correlated instantaneous solar irradiance in spatial networks. *Sol. Energy* **2017**, *143*, 10–21. [CrossRef]
27. Voyant, C.; Muselli, M.; Paoli, C.; Nivet, M.L. Optimization of an artificial neural network dedicated to the multivariate forecasting of daily global radiation. *Energy* **2011**, *36*, 348–359. [CrossRef]
28. Lotfi, M.; Javadi, M.; Osório, G.J.; Monteiro, C.; Catalão, J.P. A novel ensemble algorithm for solar power forecasting based on kernel density estimation. *Energies* **2020**, *13*, 216. [CrossRef]
29. Aslam, M.; Lee, J.M.; Kim, H.S.; Lee, S.J.; Hong, S. Deep learning models for long-term solar radiation forecasting considering microgrid installation: A comparative study. *Energies* **2020**, *13*, 147. [CrossRef]
30. Munkhammar, J.; Widén, J. A Markov-Chains probability distribution mixture approach to the clear-sky index. *Sol. Energy* **2018**, *170*, 174–183. [CrossRef]
31. Munkhammar, J.; Widén, J. An B-state Markov-Chains mixture distribution model of the clear-sky index. *Sol. Energy* **2018**, *173*, 487–495. [CrossRef]
32. Larrañeta, M.; Fernandez-Peruchena, C.; Silva-Pérez, M.; Lillo-Bravo, I.; Grantham, A.; Boland, J. Generation of synthetic solar datasets for risk analysis. *Sol. Energy* **2019**, *187*, 212–225. [CrossRef]
33. Shepero, M.; Munkhammar, J.; Widén, J. A generative hidden Markov model of the clear-sky index. *J. Renew. Sustain. Energy* **2019**, *11*, 043703. [CrossRef]

34. Cervone, A.; Carbone, G.; Santini, E.; Teodori, S. Optimization of the battery size for PV systems under regulatory rules using a Markov-Chains approach. *Renew. Energy* **2016**, *85*, 657–665. [CrossRef]
35. Ben-Gal, I. Outlier detection. In *Data Mining and Knowledge Discovery Handbook*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 131–146.
36. Ye, M. *An Integer Programming Clustering Approach with Application to Recommendation Systems*; Iowa State University: Ames, Iowa, 2007
37. Draper, N.R.; Smith, H. *Applied Regression Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 1998; Volume 326.
38. Bracale, A.; Carpinelli, G.; De Falco, P. A new finite mixture distribution and its expectation-maximization procedure for extreme wind speed characterization. *Renew. Energy* **2017**, *113*, 1366–1377. [CrossRef]
39. Koudouris, G.; Dimitriadis, P.; Iliopoulou, T.; Mamassis, N.; Koutsoyiannis, D. A stochastic model for the hourly solar radiation process for application in renewable resources management. *Adv. Geosci.* **2018**, *45*, 139–145. [CrossRef]
40. La Salle, J.L.G.; Badosa, J.; David, M.; Pinson, P.; Lauret, P. Added-value of ensemble prediction system on the quality of solar irradiance probabilistic forecasts. *Renew. Energy* **2020**, *162*, 1321–1339. [CrossRef]
41. Bozorg, M.; Fatemi, N.; Andres Pena, C.; Mousavi, O.; Carpita, M. *L'intelligence Artificielle au Service des Réseaux*; Technical Report, 2020. Available online: https://www.bulletin.ch/fr/news-detail/lintelligence-artificielle-au-service-desreseaux.html (accessed on 2 August 2021).
42. Sengupta, M.; Andreas, A. *Oahu Solar Measurement Grid (1-Year Archive): 1-Second Solar Irradiance; Oahu, Hawaii (Data)*; Technical Report; National Renewable Energy Lab. (NREL): Golden, CO, USA, 2010.
43. Andrews, R.W.; Stein, J.S.; Hansen, C.; Riley, D. Introduction to the open source PV LIB for python Photovoltaic system modelling package. In Proceedings of the 2014 IEEE 40th Photovoltaic Specialist Conference (PVSC), Denver, CO, USA, 8–13 June 2014; pp. 0170–0174.
44. Qu, Z.; Oumbe, A.; Blanc, P.; Espinar, B.; Gesell, G.; Gschwind, B.; Klüser, L.; Lefèvre, M.; Saboret, L.; Schroedter-Homscheidt, M.; et al. Fast radiative transfer parameterisation for assessing the surface solar irradiance: The Heliosat-4 method. *Meteorol. Z.* **2017**, *26*, 33–57. [CrossRef]