Generation of synthetic solar datasets for risk analysis

Larrañeta, M.[1], Fernandez-Peruchena, C.[2], Silva-Pérez, M.A.[3], Lillo-bravo, I.[3], Grantham, A.[4], Boland, J.[4]

[1] Andalusian Association for Research and Industrial Cooperation (AICIA). Seville, Spain

[2] National Renewable Energy Centre (CENER). Seville, Spain

[3] University of Seville. Department of Energy Engineering. Seville, Spain.

[4] Centre for Industrial and Applied Mathematics, University of South Australia. Mawson Lakes, Australia

Corresponding author:

Miguel Larrañeta, Andalusian Association for Research and Industrial Cooperation, Camino de los Descubrimientos s/n. 41092, Seville, Spain.

Phone: (+34)954487237

E-mail: mlarraneta@gter.es

Abstract

In this paper, we present a method for the synthetic generation of long-term time series of coupled 1-min global horizontal solar irradiance (GHI) and direct normal solar irradiance (DNI). This method requires an input of 10-15 annual time series of hourly DNI and GHI values that can be retrieved from satellite-based irradiance databases, and produces 100 years of 1-min solar radiation values that can be used for risk analysis or as input for solar plants performance simulation in a wide range of scenarios.

The method consists of the conjunction of three steps. The first one, based on a stochastic procedure, is used to generate 100 years of monthly DNI and GHI values. The second step consists of the subsequent generation of daily irradiation values. To that end we have used a bootstrapping technique. The synthetic daily sequences have the same serial correlation structure as the observed data. The last step consists of the generation of 100 years of 1-min solar irradiance data out of the daily values based on the non-dimensionalization of the daily profiles by the clear sky envelope approach. The method has been applied for the location of Seville showing satisfactory results in terms of cumulative distribution functions (CDFs) of the synthetic data. We obtain an average monthly KSI (Kolmogorov-Smirnov test integral) index of 0.11 kWh/m^2 for GHI and 0.26 kWh/m^2 for DNI. The minimum KSI value is 0.07 kWh/m^2 for GHI and 0.15 kWh/m^2 for DNI obtained in January. The maximum KSI value is 0.19 kWh/m^2 for GHI and 0.34 kWh/m^2 for DNI obtained in June and August respectively.

Keywords

Typical meteorological year; high frequency; synthetic sequences; cloud transients; solar variability

1 Introduction

Concentrated solar power (CSP) and photovoltaic (PV) plants simulation tools typically use a 1-year meteorological series representative of “typical” conditions, defined as typical meteorological year (TMY) (Hall et al., 1987)). Accordingly, their results represent the “typical” yield values defined as specific values (with no confidence intervals), rather than “distributions of values” expected given the actual uncertainties of the model and inputs. As a result, it is not suited for designing systems to meet the worst-case conditions occurring at a location (Wilcox, S., & Marion, 2008), which is a common practice in feasibility analysis and is currently carried out through the definition of several probabilities of exceedance scenarios.

An approach of growing interest to take into account the inherent variability of the performance of solar plants is the consideration of a probability distribution for each input variable in the plant model for assessing both their uncertainty and variability (Ramirez et al., 2017). In particular, for generating the probability distributions of annual energy yield, simulation tools require as input a large number of plausible meteorological years (PMYs), whose generation is the objective of this work.

Solar-power financing is mainly based on the statistical characterization of the solar resource and its uncertainty. Current methods usually rely on the use of satellite-derived solar radiation time series (typically 15-20 years) locally adapted with coincident validated ground measurements at the project site. These series allow a deep understanding of the solar radiation dynamics over different temporal scales (intraday, day, season, year), and in particular the inter-annual variability which defines the probability of exceedance scenarios. These scenarios are modeled to evaluate a project's ability to meet the risk and return requirements of investors (Fernández-Peruchena et al., 2018). On the other hand, statistical approaches are taking a growing interest in CSP projects requiring high capital-intensive investments, through the implementation of probabilistic approaches addressing the variabilities in both energy yield and system costs (Eck et al., 2012; Ho and Kolb, 2010). In this new scheme, the characterization of the solar resource variability (at both short-term and long-term scales) is an essential issue, as the DNI is the most decisive variable in the CSP plant energy yield. The research community is addressing this issue and making substantial progress: several approaches have already been tested and presented in the literature (Nielsen et al., 2017). This new scheme relies on the concept of PMY, defined as a high-frequency yearly series of GHI, DNI and other relevant meteorological variables (temperature, relative humidity, wind speed), which are consistent with the corresponding monthly and annual series, thus preserving natural variability characteristics (Fernández-Peruchena et al., 2015). It is worth noting that, besides solar irradiance, also other meteorological variables are required for solar plants yield simulation, among which the wind speed and ambient temperature stand out, as well as additional variables of interest (circumsolar radiation, surface level radiation extinction, soiling, etc.). Site-specific features of these variables should be assessed for accurate PMY generation.

One approach for generating PMYs is based on sampling from a data bank of measured data. Usaola et al. (2014) presented a method based on Bootstrap Sampling, useful in the event of having a database with enough representative data. The bootstrap methodology has been also tested in (Ramírez et al., 2015) using 4 years of data, but without satisfactory results. Other approaches rely on the synthetic generation of PMYs. Röttinger et al. (2015) assumed the annual

DNI values to be normally distributed for defining a generation scheme, with a 1-σ width equivalent to the diagnosed data set uncertainty. In the multiscale stochastic model (MUS) scheme (Fernández-Peruchena et al., 2015), GHI annual values are generated by randomly sampling its Normal distribution (characterized by means of available GHI series at the site). From these annual series, each temporal scale generated (monthly, daily, intra-daily) reproduces the natural characteristics at that scale and also accounts for the boundary conditions imposed by the lower temporal resolution series previously generated (e.g., monthly series must reproduce annual values). DNI and other meteorological variables of interest for the plant simulation are generated from the corresponding GHI series.

This hierarchy of models is a valuable method of ensuring the consistency across time scales. This provided the basis for the multiple time scale synthetic generation of rainfall data in Piantadosi et al. (2009). As referred to in that paper, a number of other papers focusing on generating rainfall on different time scales start at a high frequency, say daily, and then progress to lower frequencies such as monthly and yearly. However, what was remarked was that if one followed that progression, even for the situation where the synthetic daily totals are statistically indistinguishable from the observed data, when one amalgamated those into monthly totals, the monthly means matched well but the variance of the synthetic monthly totals significantly underestimated those of the data. Thus, this alternative progression from lower to higher frequency is adopted.

There are a limited number of papers dealing specifically with the generation of daily solar radiation datasets, among which it is worth mentioning that of Aguiar et al. (1988). They generated daily sequences using only average monthly radiation and a library of Markov transition matrices for each month. The approach taken here is similar to that of Boland (2008), Magnano et al. (2010), and Boland (2010), and which was utilized in Grantham et al. (2018). In those papers, the daily solar irradiation data series was standardized by subtracting a Fourier series model which replicates the seasonality of the series, and dividing by an exponentially weighted moving average of the standard deviation of the series over the year. Then, a first order autoregressive AR (1) model was constructed for the standardized series and finally a white noise series is calculated from the differences between the data and the AR(1) model.

When it comes to downscaling high resolution time series, there has been many approaches based on Markov chains (Morf, 1998; Ngoko et al., 2014; Bright et al., 2015; Bright et al., 2017; Munkhammar et al., 2018), the partition of solar radiation into a deterministic and a stochastic component (Polo et al., 2011; Larrañeta et al., 2015; Larrañeta et al., 2018; Grantham et al., 2017) and the non-dimensionalization of the daily profiles (Fernández-Peruchena et al., 2015; Fernández-Peruchena et al., 2018; Larrañeta et al., 2018), but at the moment, there is still room to improve the reproduction of the solar variability in the high resolution temporal scale (Lohmann, G., 2018). In this paper, we rely on the methods recently published in Grantham et al. (2018) and Larrañeta et al. (2018) to present a 3-step methodology for generating 100 PMYs of coupled DNI and GHI at 1-min time resolution, using hourly input series of 10-15 years at the site that can be easily retrieved from satellite estimates. This work presents for the first time, to our knowledge, a globally applicable method for the synthetic generation of multiple annual solar radiation time series in high resolution (1-min) going beyond previous approaches that reached the monthly (Fernández-Peruchena et al., 2015) or daily resolution (Grantham et al., 2018). This methodology is oriented towards the new paradigm described on probabilistic approaches addressing the plant feasibility.

Fully stochastic analyses that take into account the uncertainty and variability in different parameters or factors affecting the energy yield of solar energy projects during their lifetime have been identified as a promising approach to overcome the limitations of traditional, deterministic feasibility analysis (Nielsen et al., 2017). Since variability and uncertainty in the solar irradiance is the first source of uncertainty in the estimation of the energy yield, the elaboration and assessment of *realistic multi-year generation* has been identified as one key research task for the creation of *Advanced Meteorological Data Sets for CSP Performance Simulations* (Ramírez et al., 2017)

The idea of the synthetic generation of long term series is not new: it was presented together with the concept of Plausible Meteorological Years (PMY) in Fernandez-Peruchena et al., 2015. That paper presented a general and functional methodology for the generation of such PMYs based in a number of sub-methodologies and models that perform adequately. However, the ultimate aim of that paper was to encourage other experts to contribute to an improved methodology with present or future methods for each of the steps that could lead to better results. In this context, the present work is based on the availability of long-term, high-precision series at the site, from which valuable information is extracted through novel synthetic generation techniques (Grantham et al., 2018).

The paper is organized as follows. Section 2 describes the data used for the testing the method performance. Section 3 describes the algorithm step by step. In Section 4, we present the results analysis and the discussion of the performance of the method, and in Section 5 the conclusion and future lines of work are drawn.

2 Meteorological database

In this work, an extensive database is used for testing the proposed method (Table 1). This database is composed of 1-min average values of DNI and GHI recorded during 14 consecutive years (2002–2015) for the location of Seville (Spain). The measurements were taken with a sampling and logging frequency of 0.2 Hz. An ISO first class Eppley NIP pyrheliometer mounted on a sun tracker Kipp & Zonen 2AP measured the DNI. A secondary standard Kipp and Zonen CMP21 measured the GHI. The devices are located at the meteorological station of the Group of Thermodynamics and Renewable Energy of the University of Seville and have been periodically calibrated, at least once every two years. Data used in this work have been subjected to quality-control procedures following the BSRN recommendations (McArthur, 2004). We have performed a gap filling procedure for the identification and replacement of missing or incorrect data in order to prepare a complete database (Moreno et al., 2016). Only 1% of the data has been filled for the entire training database.

Table 1. Location selected for the training the method.

	Latitude (°N)	Longitude (°W)	Altitude (m)	Climate	Period
Seville	37.4	6.0	12	Mediterranean	2002-2015

We use the observed 1-min GHI and DNI data obtained from the Australian Bureau of Meteorology website (Bureau of Meteorology, 2015) for the location of Adelaide, the capital of South Australia (Table 2) to generate the synthetic high temporal resolution synthetic time series for the location of Seville (step 3). GHI and DNI data were observed with a secondary standard Kipp and Zonen CM11 and a first class Carter Scott DN5 respectively

Table 2. Data used for the generation of the dimensionless database.

	Latitude (°N)	Longitude (°W)	Altitude (m)	Climate	Period
Adelaide	-34.9	-138.5	2	Hot Mediterranean	2003-2017

3 Methodology

In this section we describe the implemented algorithm for the multiyear synthetic generation step by step. Figure 1 shows the flow diagram of the implemented algorithm. The input to the algorithm is the available 10-15 years series at the site, either registered or modeled (referred to as *10-15 "observed" years* in Fig. 1). In step 1 we generate synthetic monthly data from the observed data. In step 2, both the output of the first step and the observed data are required as input. In the last step, we require as input an extensive normalized 1-min database from any location and the output of the second step.

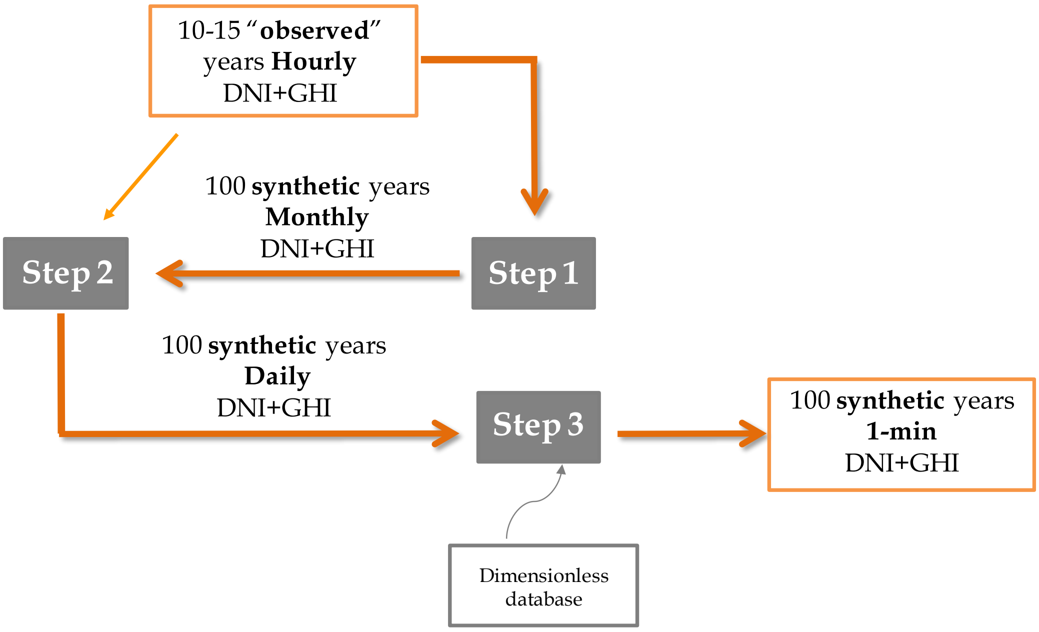


Fig. 1. Flow diagram of the implemented algorithm for the multiyear synthetic generation.

3.1 Step 1: From monthly observed to monthly synthetic

In the first step of the methodology, from the available initial series (hereafter *observed* series, even if they can also come from models), we generate 100 annual series at monthly scale, corresponding to the annual probability of exceedance from 1 to 100. The annual probability of exceedance is the probability that a given solar radiation total accumulated over one year will be exceeded in any year. For 100 years, the probability of exceedance is complementary to the percentile (P). For instance, a PoE10 corresponds to a P90. Given a random variable X with continuous and strictly monotonic probability density function $f(x)$, a quantile function Q_f assigns to each probability p attained by f the value x for which $P_r(X \leq x) = p$. Symbolically, the quantile function can be represented as follows:

$$Q(p) = F^{-1}(p) = \inf\{x; F(x) \geq p\}, 0 < p < 1$$

The k^{th} n-tile P_k is that value of x, say x_k, which corresponds to a cumulative frequency of $\mathrm{N} \cdot \mathrm{k}/\mathrm{n}$. If n = 4, the quantity is called a quartile, and if n = 100, it is called a percentile.

We use 10-15 years of observed data as input since it is a commonly available length of satellite derived time series, but the size of the sample may not be enough to statistically characterize the annual and monthly distribution of the solar radiation. The larger the sample size, the better the characterization of the annual and monthly distribution of the solar radiation. For smaller sample sizes than 10 years of observed data, the characterization of the distributions may not be statistically representative. We calculate the synthetic monthly values following the probability integral transform method to generate independent random monthly values. The procedure for generating a monthly GHI and DNI coupled pair is briefly described below:

i. Calculate the cumulative distribution function (CDF) of the observed GHI monthly series for each month. Since input data is obtained in the hourly resolution, we firstly integrate the observed hourly data into the monthly resolution.

The probability integral transform method is then applied (steps ii and iii):

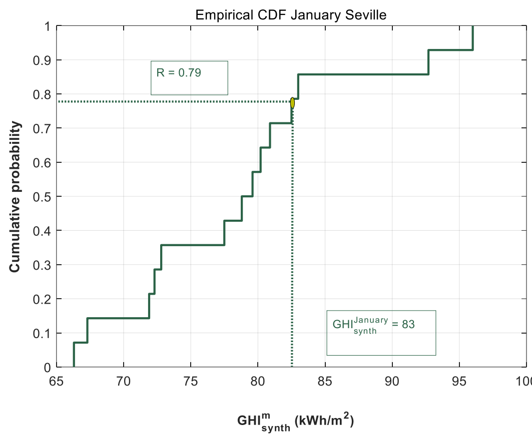
ii. Generate an independent random number *R* from a uniform distribution in the interval between 0 and 1.
iii. Identify the value whose cumulative probability corresponds to the value generated with the random number *R* to obtain the monthly synthetic GHI value (GHI_{synth}^{m}). Figure 2 illustrates this step.

To maintain the relation between the GHI and the DNI at the site, we include limitations for the calculation of the monthly synthetic DNI values (steps iv and v):

iv. Find the values of DNI in the observed monthly data of the month of the year under treatment in the range of the $GHI_{synth}^{m} \pm 5\%$.
v. Calculate the CDF of the identified possible values of DNI in the range $GHI_{synth}^{m} \pm 5\%$.

The probability integral transform method is then applied again for the DNI (steps vi and vii):

vi. Generate an independent random number Y from a uniform distribution in the interval between 0 and 1.
vii. Identify the monthly synthetic DNI value (DNI_{synth}^{m}) such as the value whose cumulative probability corresponds to the value generated with the random number Y.



i.

Fig. 2. Graphical reproduction of the stage iii of the procedure for the dynamic generation of synthetic monthly values. Example for the month of January for the location of Seville.

We repeat this procedure for each of the 12 months of the year, thereby obtaining the annual synthetic cumulative values by adding the monthly synthetic values. We calculate 10,000 annual synthetic values running therefore the procedure 120,000 times. The complete procedure of generating 10,000 synthetic annual values results in a low computational cost. In Fig. 3, we represent a scatter plot of the annual values of DNI versus the corresponding annual values of GHI of both the observed (dark blue squares) and the 10,000 synthetic data (light blue circles).

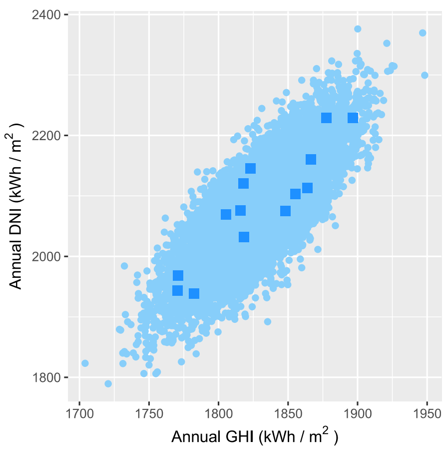


Fig. 3. Annual DNI versus their corresponding GHI values of both observed (dark blue squares) and 10,000 synthetic data (light blue circles) for the location of Seville.

Even if the procedure presented here provides synthetic monthly irradiation data within the limits of the observed values, it can generate annual irradiation values upper or lower than the extreme values of the observed annual series, as can be observed from Fig. 3.

To generate the 100 sets of annual synthetic values, a preliminary statistical analysis of annual solar irradiation probability density functions has been carried out. We have considered that the annual GHI and DNI series follow Normal and Weibull distributions, respectively (Peruchena et al. 2016). The assumption of normal distribution for annual GHI series is validated by the Kolmogorov–Smirnov and Shapiro–Wilk normality tests ($p > 0.05$), and the Weibull distribution assumption for annual DNI series is analyzed by a set of goodness of fit tests (p > 0.05) (for more details, the reader is referred to Peruchena et al. (2016)).

We calculate the annual DNI and GHI expected probabilities of exceedance (PoE^n_{DNI} and PoE^n_{GHI}, respectively) by fitting the observed annual values to a Weibull and a Normal distribution, respectively. Then, we seek for the lower absolute values of the differences of the synthetic annual coupled DNI and GHI values to the theoretical annual PoEs within the 10,000 generated synthetic values. In Fig. 4 we present a scatter plot of the annual values of DNI versus the annual values of GHI of the observed data (dark blue squares) and the 100 synthetic data corresponding to the PoEs from 1 to 100 (light blue circles).

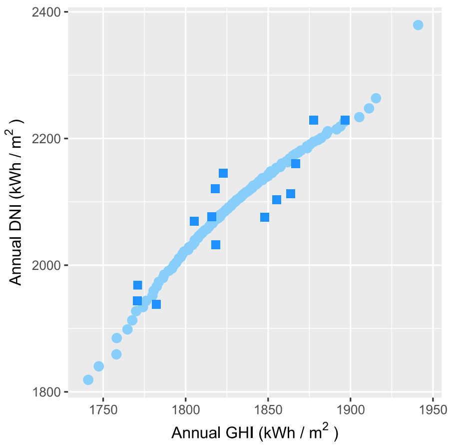


Fig. 4. Annual values of DNI versus the corresponding GHI values of the observed data (dark blue squares), along with the 100 synthetic values corresponding to the PoEs from 1 to 100 (light blue circles) for the location of Seville.

3.2 Step 2: From monthly synthetic to daily synthetic

The second step of the methodology is to generate daily synthetic solar irradiation values out of the monthly synthetic data generated in Step 1. The daily solar irradiation observed data series are standardized by subtracting a Fourier series model which replicates the seasonality of the series, and dividing by an exponentially weighted moving average of the standard deviation of the series over the year. Then, a first order autoregressive AR(1) model is constructed for the standardized series and finally a white noise series is calculated from the differences between the data and the AR(1) model. This pool of random error terms is bootstrapped in the reverse procedure to create any number of synthetic years before matching of monthly totals to the generated synthetic monthly totals from Step 1.

We aim to generate synthetic quartets giving information about the energy, variability and distribution of the daily solar irradiance profiles. We synthetically generate three daily indexes that can be used to characterize or classify different types of days (Moreno et al., 2017):

- Energy: We use the daily direct fraction index (Skartveit and Olseth, 1992) and the daily clearness index (Black et al., 1954) to characterize the daily energy of a given day following next equations:

$$K_b = \frac{DNI_d}{DNIcs_d} \ , \tag{1}$$

$$K_t = \frac{GHI_d}{H0_d} \ , \tag{2}$$

where DNI_d is the daily DNI, $DNIcs_d$ is the daily DNI under clear sky conditions, GHI_d is the daily GHI and $H0_d$ is the daily extra-terrestrial solar radiation.

- Variability: We use the Variability Index (VI) (Stein et al., 2012), defined as the ratio between the length of the DNI curve and the length of the maximum enveloping clear day curve.

$$VI = \frac{\sum_{k=2}^{n} \sqrt{(DNI_k - DNI_{k-1})^2 + \Delta t^2}}{\sum_{k=2}^{n} \sqrt{(DNIcs_k - DNIcs_{k-1})^2 + \Delta t^2}} \ . \tag{3}$$

$DNIcs$ is the hourly enveloping clear sky direct normal irradiance, Δt refers to an interval of one hour, and n is the number of 1-h intervals of the considered day.

- Distribution: We use the morning fraction index F_m defined as the ratio between the accumulated DNI in the first half of the day and the accumulated DNI for the whole day.

$$F_m = \frac{DNI_{d/2}}{DNI_d} \ . \tag{4}$$

$DNI_{d/2}$ is the DNI recorded from sunrise to solar noon and DNI_d is the daily DNI.

The procedure for the dynamic generation of the daily values is briefly described in 4 stages.

1. Generate a multitude of synthetic years of GHI_d, from using the time series of the observed data (we generate 1200 years). We identify the Fourier series component, which basically is the time varying mean, and then we identify the time varying standard deviation. By subtracting the mean and dividing the result by the standard deviation, we get a set of time varying standard scores, which are homoscedastic:

$$s_t = \frac{GHI_d - fs_t}{sd_t} \ . \tag{5}$$

2. In the second stage, we find a p-order autoregressive AR(p) model for $s_t = \beta_1 s_{t-1} + \beta_2 s_{t-2} + \cdots + \beta_p s_{t-p} + \varepsilon_t$. The white noise (WN) terms ε_t are put in separate bins for each month. Then, to créate the synthetic daily values for each month we first follow a similar procedure to that described in Figure 2 for the synthetic monthly totals. We generate a random number on the interval $(0,1)$, which will be a probability p, and then find the value of ε_t from the empirical cumulative distribution function for the month in question such that $P(WN \leq \varepsilon_t) = p$. The next step is to calculate s_t using the equation above. We substitute this into Eqn (5) and rearrange to obtain a synthetic value of GHI_d. This process is continued to generate a sequence that will make up a synthetic month of daily values. This can be repeated any number of times to generate a multitude of synthetic months of daily values. This is repeated for every month.
3. For each monthly synthetic total generated in step 1 (Section 3.1), select the sequence of synthetic daily totals that aggregates to that of a particular monthly total. Then we select a set of $F_m \ and \ VI$ variables for which the synthetic GHI_d most closely matches a observed GHI_d.

4. An extra dimension of stochasticity is added when selecting a value of DNI_d to pair with GHI_d. Instead of simply taking the direct normal value that matches the global in the historical record, we perturb that connection in a particular manner to select a statistically possible DNI value. First, for each month we construct a statistical model connecting DNI_d with GHI_d in the form of a logistic relationship. Then we use an exponentially smoothed moving variance to estimate the variance of DNI_d for each GHI_d. From that we can construct prediction intervals for values of DNI_d corresponding to each value of GHI_d . Then we select at random a particular value of DNI_d to pair with GHI_d.

3.3 Step 3: From daily synthetic to 1-min synthetic

The last step of the method consists on the generation of 1-min synthetic DNI and GHI data from the synthetic daily quartets of k_b, k_t, VI and F_m provided in Step 2. We use an improvement of the non-dimensional (ND) model (Larrañeta et al, 2018). The model consists of the normalization of the daily solar radiation profiles by the clear-sky envelope approach and the extraterrestrial solar radiation for DNI and GHI respectively, creating daily Dynamic Paths from observed solar radiation data. The method transforms each daily coupled DNI and GHI 1-min curve into a dimensionless curve where the time scale, the DNI scale and the GHI scale go from 0 to 1. The normalization of the time scale is accomplished fixing the day length to 1000 points by performing a linear interpolation proportional to the number of minutes of a given day. The DNI and the GHI are normalized calculating the direct fraction index and the cleanness index respectively. In Fig. 5, we show a daily solar radiation curve and the dimensionless daily shape for the same curve. We generate a database of dimensionless profiles from one location to be applied in any other location without any local adaptation. In this case, we use the observed 1-min GHI and DNI data obtained from Adelaide (Australia) to generate the synthetic high temporal resolution synthetic time series for the location of Seville. We generate dimensionless profiles of those days that fully respect the BSRN recommendations and present no gaps, generating a total 4,141 days from 15 years of measurements

For the synthetic generation of 1-min data, we implement the following steps in a daily basis:

1. Normalize the 1-min daily curves. We calculate the clear sky DNI envelopes and the extra-terrestrial solar irradiance to generate a database of dimensionless daily coupled GHI and DNI profiles for the location of Adelaide. Each day of the dimensionless database has been labelled with the calculated k_b, k_t, VI and F_m for that day
2. Search for the most similar day in terms of energy, variability and distribution. We use the daily synthetic quartets of k_b, k_t, VI and F_m synthetically generated in step 2 to find the most similar day in terms of Euclidean distance in the dimensionless database to the given day.
3. Generate synthetic coupled 1-min DNI and GHI series on a given day. We combine dimensionless daily DNI and GHI curves with the theoretical estimated envelope and extra-terrestrial profiles for the given day.

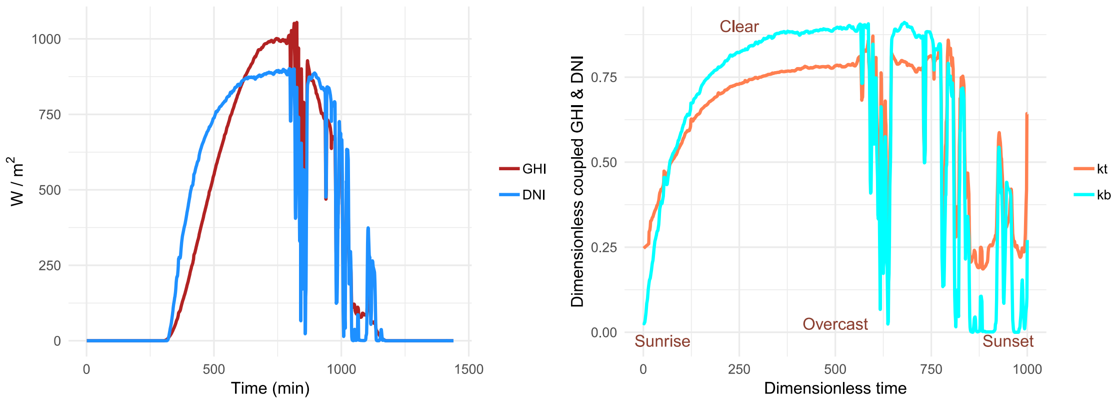


Fig. 5. Observed (a) and dimensionless daily shape (b) example.

4 Results

The results evaluation is performed step by step for the location of Seville, however, it is worth highlighting that the results from Step 3 involve the implementation of step 2 which in turn involves the implementation of Step1.

4.1 Step 1 evaluation

In Step 1 of the methodology we obtain 100 annual sets of 12 synthetic monthly values. Regarding the annual values, we assume the hypothesis that the GHI and DNI perform as a Normal and Weibull distribution respectively. In Fig. 6 we represent the CDF of the observed annual values of GHI (purple dots) and the observed annual values of DNI (red dots) together with the CDF of the synthetic annual values of GHI (orange line) and the synthetic annual values of DNI (blue line).

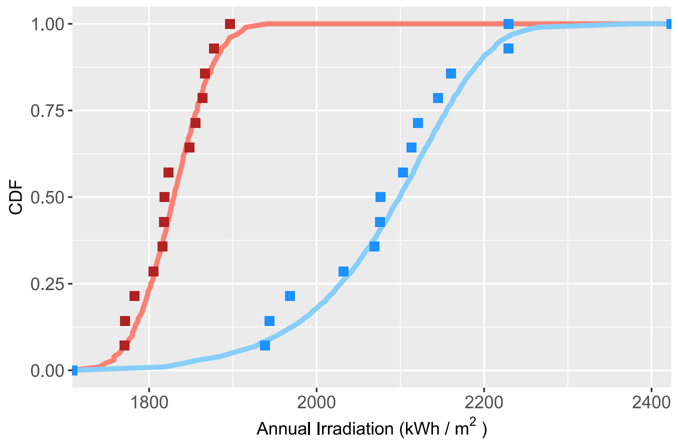


Fig. 6. CDFs of the observed annual values of GHI (purple dots) and the observed annual values of DNI (red dots) together with the CDF of the synthetic annual values of GHI (orange line) and the synthetic annual values of DNI (blue line) for the location of Seville.

We observe that the assumption can be validated since the CDFs of the observed annual values of GHI and DNI are closely fitted to the CDF of a Normal and a Weibull distribution, respectively. The fit is slightly offset for low annual values in both cases, for GHI and DNI.

In a monthly basis, we also intend to maintain the relation between the GHI and the DNI for a given location. There is a range of possible values of DNI for a given value of GHI depending on the atmospheric conditions that in turn has also a strong seasonal dependency. In Fig. 7 we show a scatter plot of the monthly values of DNI versus the corresponding GHI values of both observed (dark blue squares) and 100 synthetic datasets (light blue circles) broken down by months.

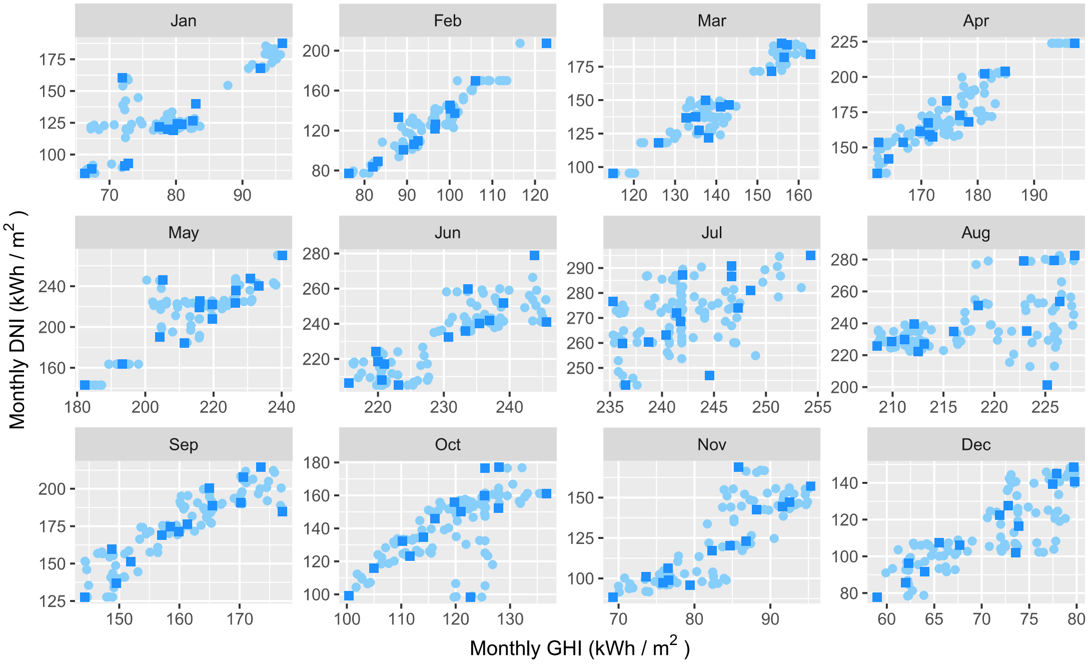


Fig. 7. Scatter plot of the monthly DNI values versus the corresponding GHI values GHI values of both observed (dark blue squares) and 100 synthetic datasets (light blue circles) broken down by months for the location of Seville.

We can observe that, for each month, synthetic data maintain the relation between DNI and GHI observed in observed dataset. For example, in March there is almost a linear relation with two well defined clusters while in August data is widely dispersed. The model is capable to reproduce both different performances.

We also intend to reproduce the performance of the radiometric variables separately. In Fig. 8 and 9 we reproduce the CDF of the observed monthly values of GHI and DNI (squares) together with the CDF of their synthetic generated values (lines).

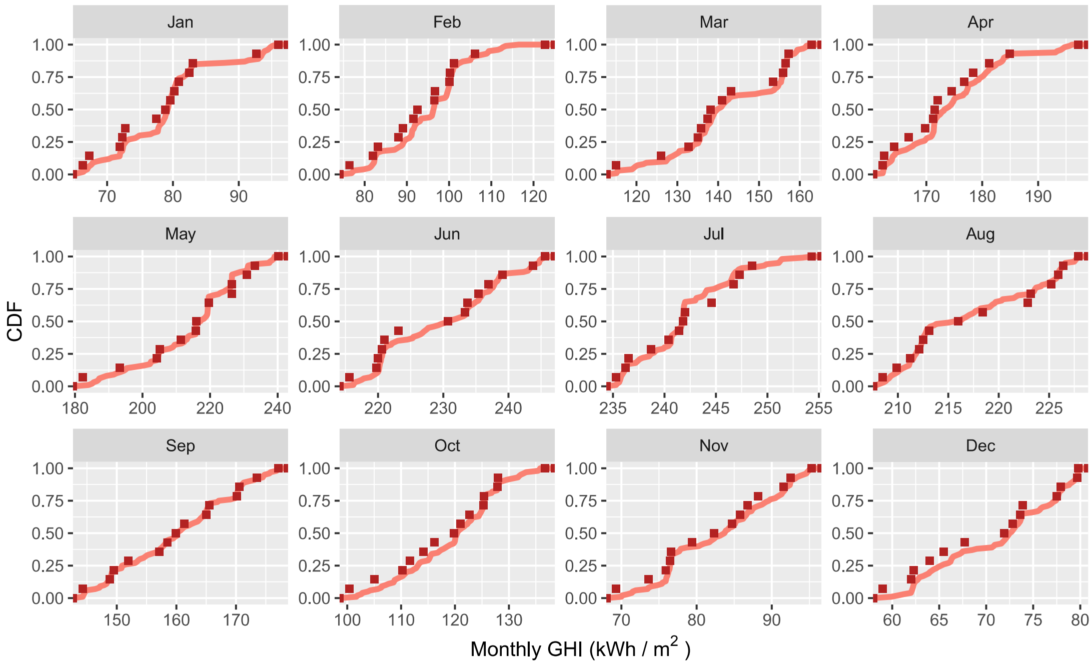


Fig. 8. Cumulative distribution functions of the observed (squares) and synthetic (lines) monthly values of GHI for the location of Seville.

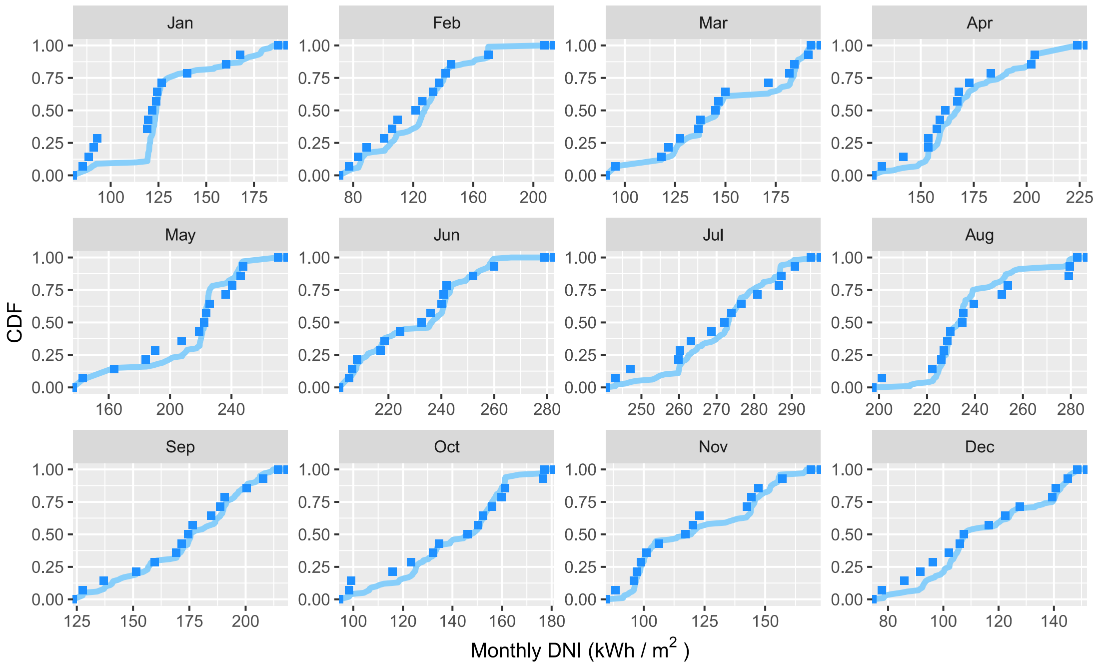


Fig. 9. Cumulative distribution functions of the observed (squares) and synthetic (lines) monthly values of DNI for the location of Seville.

It can be observed that the trend of the monthly GHI and DNI data is closely reproduced. It is worth highlighting the double trend found in January, identifiable in GHI and clearly defined in DNI and a lack of continuity that can also be observed in other months like October and November.

This method does not take into account the correlation between successive months because the synthetic monthly data is generated by independent random numbers. In Table 3 we present the P-values for the Spearman´s correlation coefficient between successive months (January, February), (February, March) of the observed and the synthetic monthly values for the location of Seville to test if any 2 months are correlated. The results suggest that, in the observed dataset, at the 0.05 significance level, there is a strong correlation between January-February and May-June for both GHI and DNI in the location of Seville while in the synthetic dataset there is no correlation between successive months. This issue, even if has no effect in the assessment of the long term annual probabilities of exceedance of solar plants performance simulations, should be addressed in future works.

Table 3. Spearman’s correlation between successive months at Seville

Month	*p*-value		*p*-value	
	DNI		GHI	
	Observed	Synthetic	Observed	Synthetic
January	0.040	0.433	0.016	0.433
February	0.542	0.799	0.605	0.799
March	0.605	0.582	0.660	0.582
April	1.000	0.686	0.417	0.686
May	0.035	0.679	0.170	0.679
June	0.976	0.940	0.215	0.940
July	0.295	0.206	0.988	0.206
August	0.128	0.116	0.341	0.116
September	0.302	0.686	0.483	0.686
October	0.988	0.811	0.310	0.811
November	0.940	0.213	0.892	0.213
December	1.000	0.948	0.820	0.948

4.2 Step 2 evaluation

In Step 2 of the methodology we obtain 100 annual sets of 365 synthetic daily quartets of k_b, k_t, VI and F_m. We again intend to reproduce the relation between GHI and DNI. Since there is a range of possible DNI daily values for a GHI daily value with a strong seasonal dependency, we confront the daily observed (light blue circles) and synthetic (dark blue squares) DNI and GHI for the location of Seville in Fig. 10.

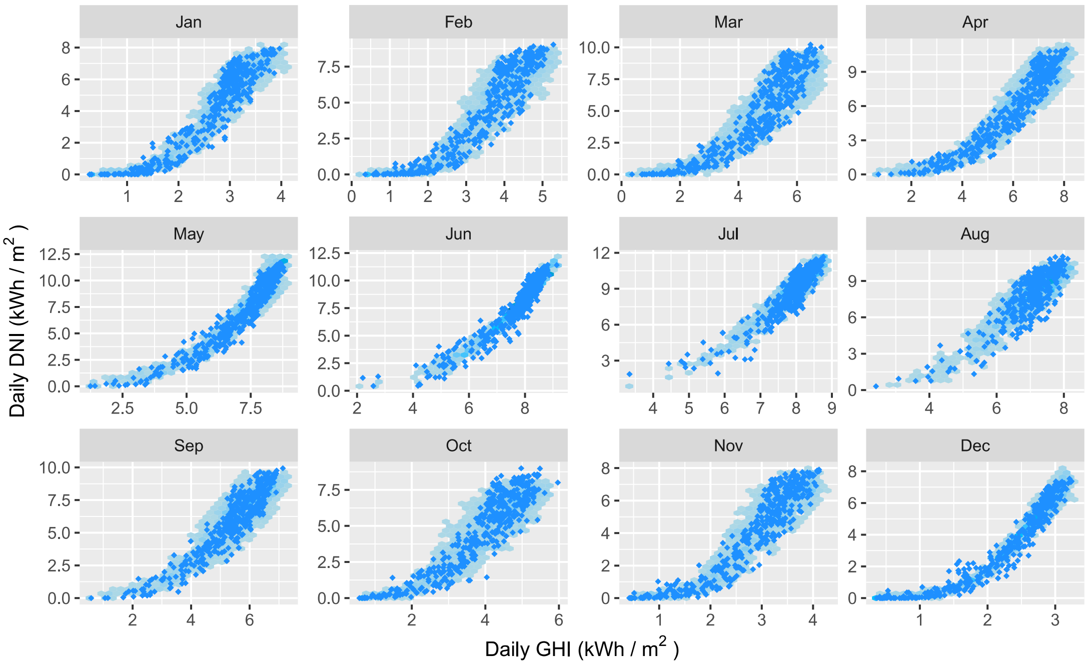


Fig. 10. Scatter plot between DNI and GHI daily values, both of observed (dark blue hexagons) and 100 synthetic (light blue hexagons) datasets broken down by months for the location of Seville.

We can observe the consistent relation between both variables on a daily basis. In May, June, July and December, there is a thinner cloud that is seemingly reproduced in the synthetic set. In other months like February, March, October and November the dispersion of the cloud is also reproduced in the synthetic set. To evaluate each radiometric variable separately, we calculate the CDFs of the daily observed (opaque) and synthetic (translucent) GHI in an annual basis (Fig. 11) and monthly basis (Fig. 12).

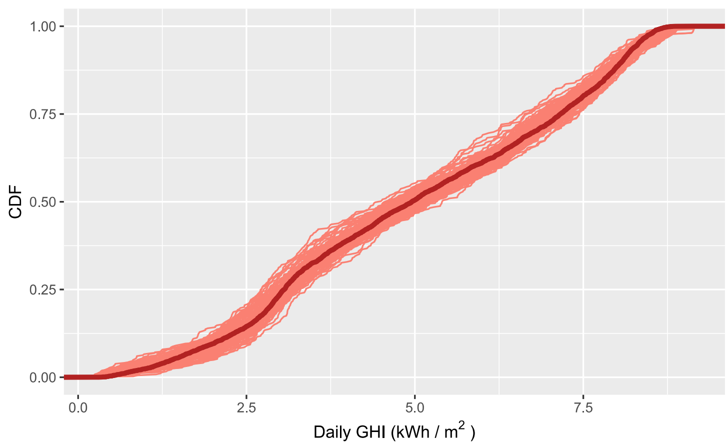


Fig. 11. Cumulative distribution functions of all observed daily values of GHI for the location of Seville (dark red), and the corresponding ones for each synthetic year (light red).

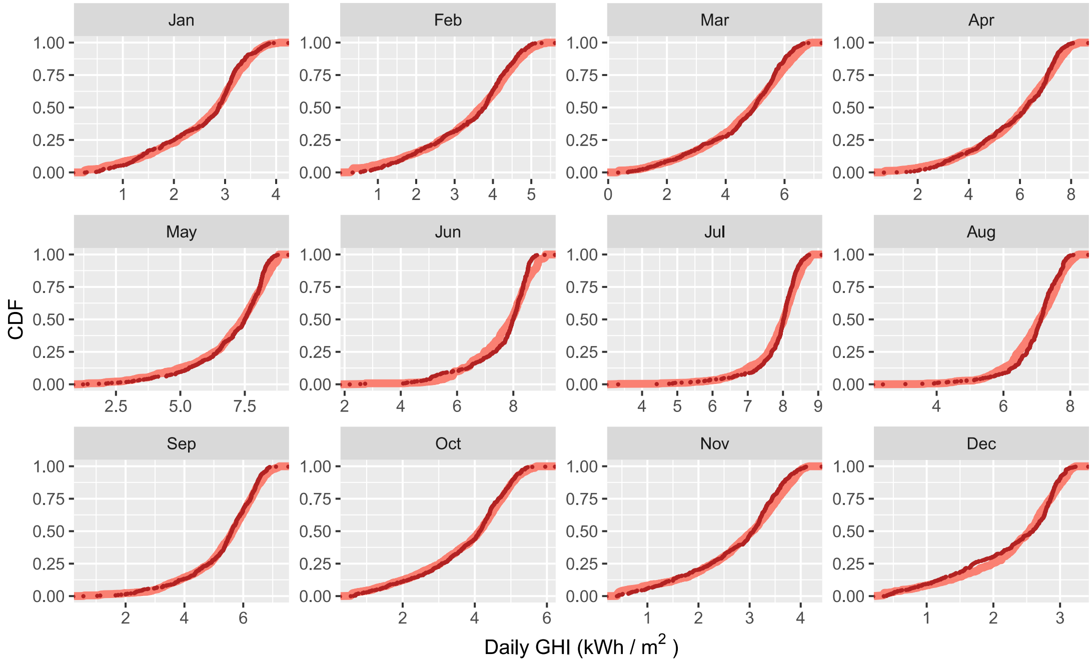


Fig. 12. Cumulative distribution functions of the observed (dark red) and synthetic (light red) daily values of GHI for the location of Seville broken down by months.

The CDFs of the synthetic data (translucent) show a wider range of scenarios that could be very useful to investors to understand the variability of the solar radiation and to evaluate its impact on the energy yield of solar plants. From fig. 12 we can assure that the monthly CDFs of the observed and synthetic daily values of GHI are almost overlapped for most of the months. It is interesting to point out that in months like February, March, October and November, that showed a disperse cloud point- see Fig. 10, the CDFs are totally superimposed while in months like May, June, July and December, when we could observe a thinner cloud point- see Fig. 10, the CDFs are slightly displaced. We also calculate the CDFs of the daily observed (dark blue) and synthetic (light blue) DNI in annual (Fig. 13) and monthly (Fig. 14) basis.

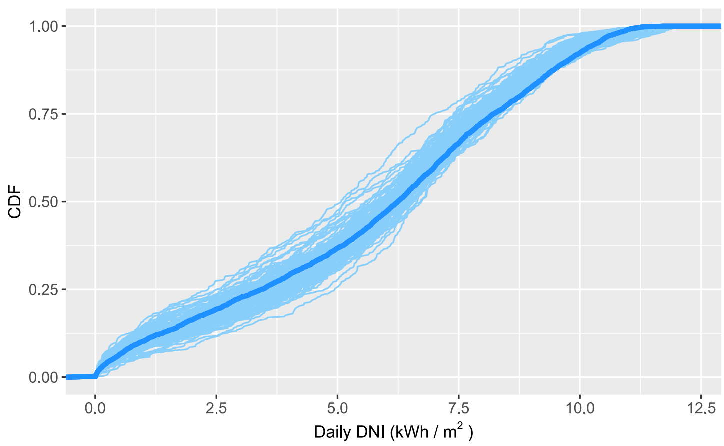


Fig. 13. Cumulative distribution functions of all observed daily values of DNI for the location of Seville (dark blue), and the corresponding ones for each synthetic year) (light blue).

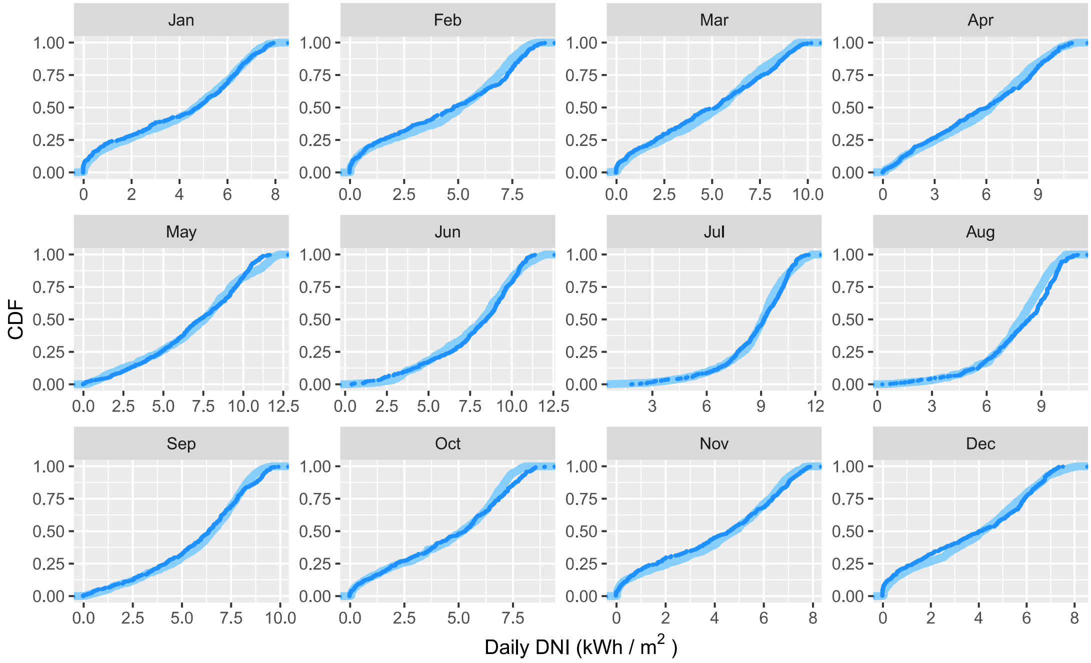


Fig. 14. Cumulative distribution functions of the observed (dark blue) and synthetic (light blue) daily values of DNI for the location of Seville broken down by months.

The annual CDFs of the daily synthetic DNI data (light blue) show again a wider range of scenarios than the TMY. We can observe a slight mismatch in the CDFs of the synthetic sets for daily values around 8.5 kWh/m^2, suggesting a scope to improve in future approaches focusing on clear sky situations. The monthly CDFs of the observed and synthetic daily values of DNI also show similarity in most of the months, perhaps in February, March and August we can observe a slight mismatch between CDFs. In Figs 15 and 16, we calculate the CDFs of the daily observed (opaque) and synthetic (translucent) VI and F_m broken down by months for the location of Seville.

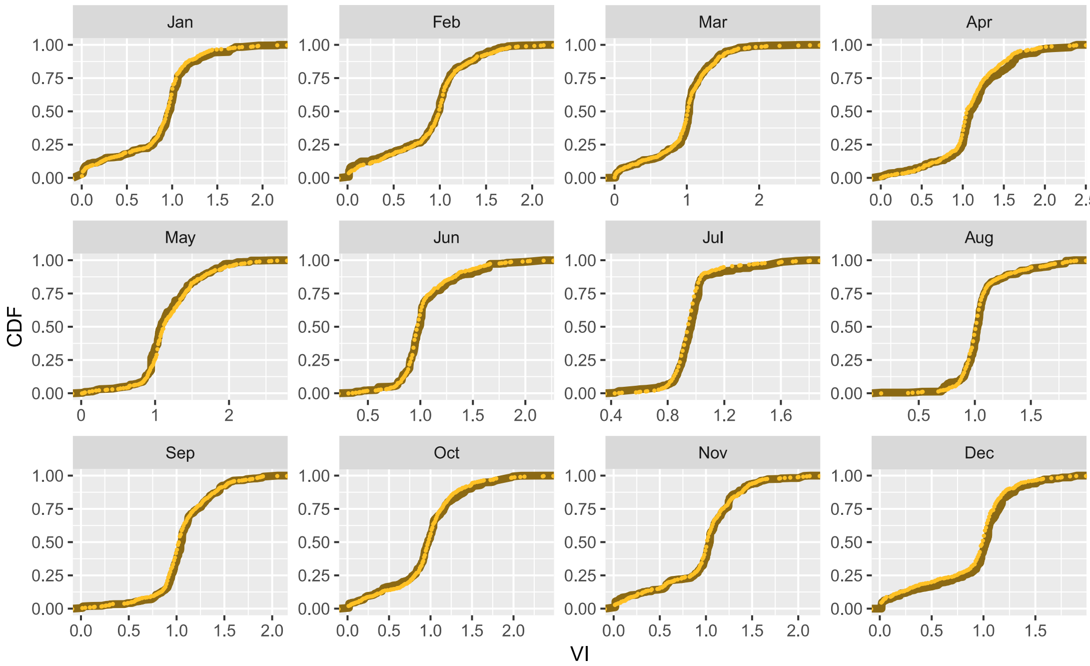


Fig. 15. Monthly CDFs of the observed daily VI values (opaque) together with the monthly CDF of the synthetic daily VI values (translucent) for the location of Seville

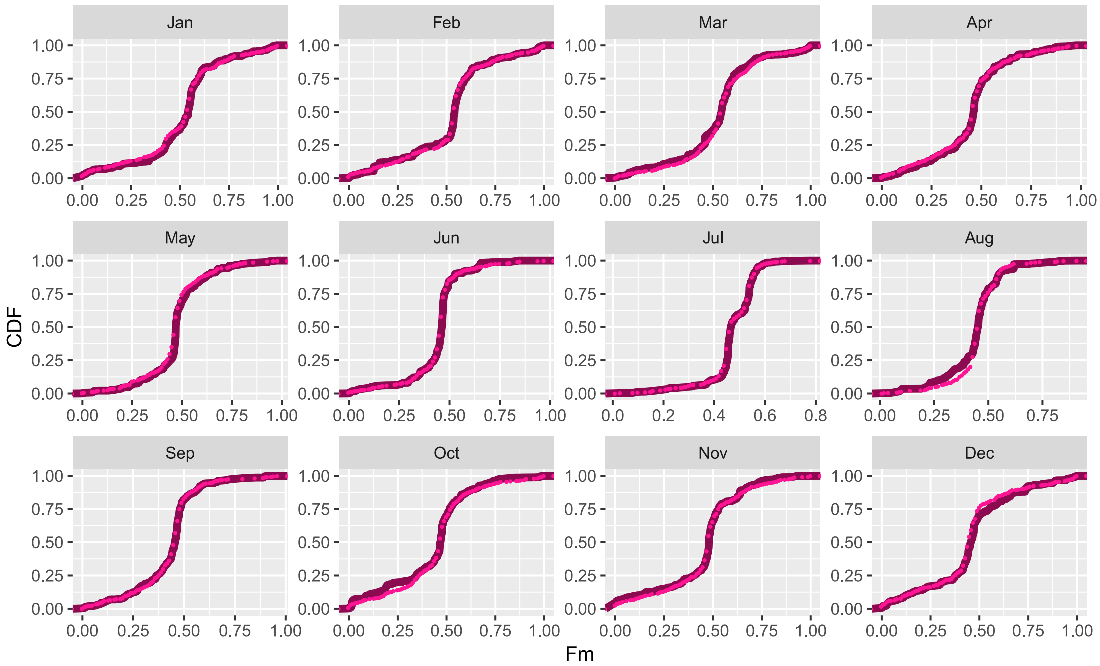


Fig. 16. Monthly CDFs of the observed daily F_m values (opaque) together with the monthly CDF of the synthetic daily F_m values (translucent) for the location of Seville

Figs. 15 and 16 show a great concordance between observed and synthetically generated variability and distribution. We can quantify the concordance between observed and synthetic sets CDFs by using the KSI (Kolmogorov-Smirnov test integral) index, that is defined as the integrated differences between the CDFs of the two data sets (Espinar et al., 2009). The unit of this index is the same for the corresponding magnitude. The higher the KSI values, the worse the model fit.

$$KSI = \int_{x_{min}}^{x_{max}} D_n dx, \quad (6)$$

where, x_{max} and x_{min} are the extreme values of the independent variable, and D_n are the differences between the CDFs of the observed and synthetic datasets. In table 4 we show the KSI values obtained for each synthetically generated variable in a monthly basis.

Table 4. KSI of the synthetic daily GHI, DNI *VI* and F_m for the location of Seville

Parameter	KSI (kWh/m^2)		KSI (-)	
Variable	GHI	DNI	VI	Fm
January	0.070	0.155	0.028	0.012
February	0.103	0.308	0.027	0.010
March	0.124	0.323	0.024	0.019
April	0.134	0.277	0.058	0.008
May	0.147	0.284	0.044	0.010
June	0.190	0.282	0.025	0.008
July	0.104	0.174	0.026	0.004
August	0.142	0.337	0.016	0.022
September	0.076	0.250	0.021	0.007
October	0.080	0.214	0.035	0.028
November	0.085	0.238	0.025	0.015
December	0.085	0.255	0.054	0.015

4.3 Step 3 evaluation

The qualitative assessment of the third step of the methodology, this is, the generation of 1-min coupled DNI+GHI data from daily synthetic quartets of indexes, is presented in the following. In Fig. 17, we present the annual CDFs of the 14 years of 1-min observed GHI (dark red) along with the corresponding CDFs of the 100 years of 1-min synthetic GHI (light red). The same results but for DNI are presented in Fig. 18.

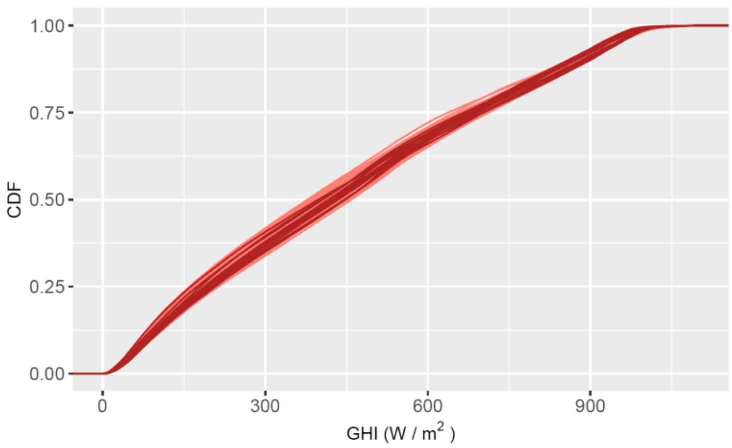


Fig. 17. Annual CDFs of the 1-min observed GHI (dark red) and synthetic (light red) 1-min GHI datasets.

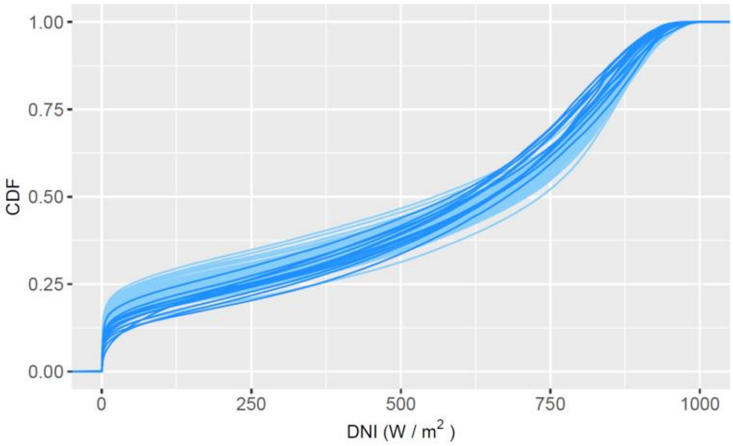


Fig. 18. . Annual CDFs of the 1-min observed (dark blue) and synthetic (light blue) 1-min DNI datasets

With this algorithm, we don´t intend to emulate only the distribution of the observed data, but also other scenarios whether extreme or not. From Figs 17 and 18, we can observe that we are capable of reproduce a larger range of scenarios including extreme years that would be useful for risk analysis. In Fig 18 we observe a misalignment between the CDFs of the synthetic sets and the CDFs of the observed data. The synthetic CDFs have a significantly increased probability at 0-250 W/m^2 resulting in a decrease in the cumulative probability around 750 W/m^2. The use of a dimensionless database from a location with a given climate (Adelaide) to reproduce the high-resolution synthetic data for a location with a different climate (Seville) may be the main cause of this deviations but also the slight deviation found in step 2 for clear sky days (Fig 13).In Figs 19 and 20 we perform the analysis month by month.

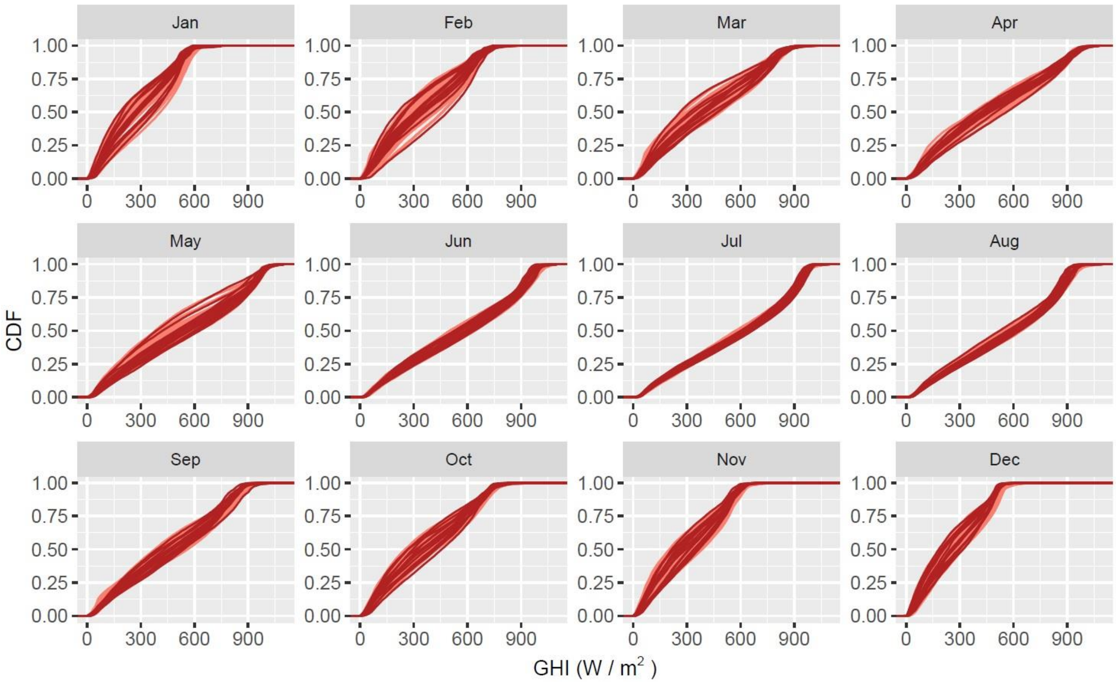


Fig. 19. CDFs of the 14 observed 1-min GHI annual sets (dark red) together with the monthly CDFs of the 100 synthetic 1-min GHI annual sets (light red) broken down by months for the location of Seville.

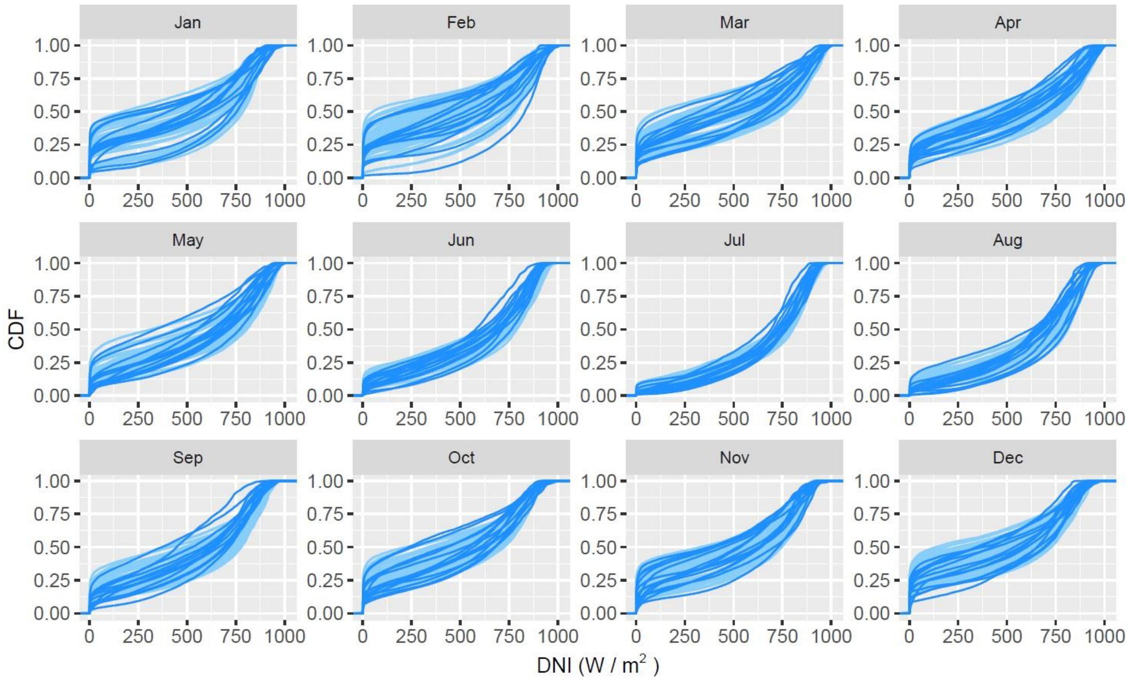


Fig. 20. CDFs of the 14 observed 1-min DNI annual sets (dark blue) together with the monthly CDFs of the 100 synthetic 1-min DNI annual sets (light blue) broken down by months for the location of Seville.

The shapes of the typical CDFs for the different months are reproduced both for GHI (Fig. 19) and DNI (Fig. 20). It is worth highlighting the high values and low dispersion of measured DNI CDFs of July, from which the CDFs reduce their maximum values and dispersion (June and August, and more markedly from May and September). Even if the typical values and variability of CDFs are well reproduced by the synthetic series, there are few extreme CDFs of some months (January and December), leading to a potential future improvement in the model

In figure 21 we present an example of the generated synthetic DNI and GHI time series by showing five consecutive days at the same time of the year for four different years.

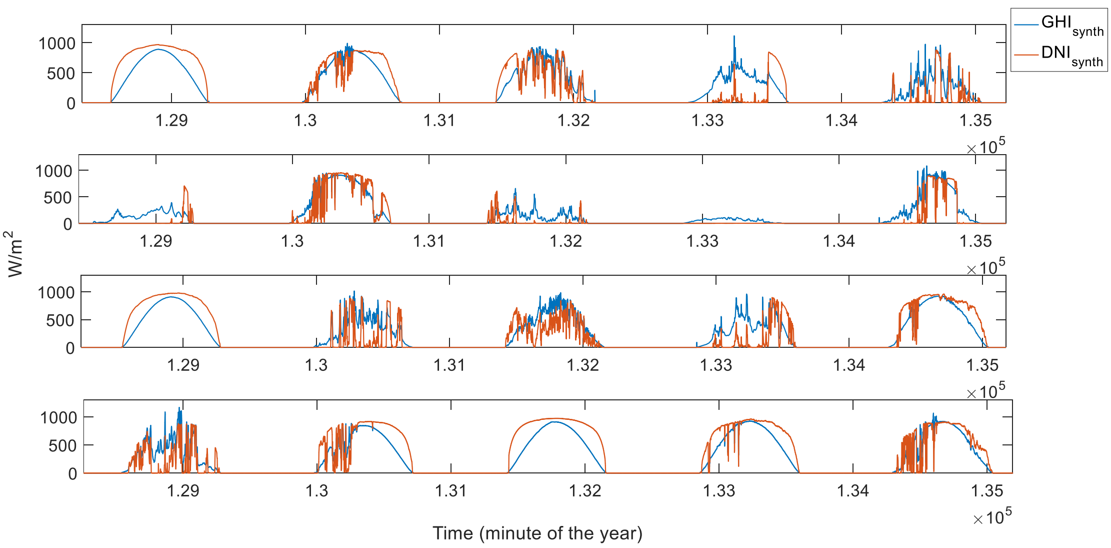


Fig. 21. Five consecutive synthetically generated daily coupled GHI and DNI profiles at the same time of the year for four different years.

The cumulative values should be preserved in order to maintain the consistency after each step of the procedure, however slight differences are found. In table 5 we present the root mean square error (RMSE) and the mean absolute error (MAE) between the outputs of each step of the method. Calculations are performed in the annual, monthly and daily resolution by integrating the synthetic data after each step of the procedure. We can observe a strong consistency in the results given by the low values of RMSE and MAE.

Table 5. RMSE and MAE (kWh/m^2) between the outputs of each step of the method.

RMSE (kWh/m^2)	Step 1-Step 2		Step 1-Step 3		Step 2-Step 3	
	GHI	DNI	GHI	DNI	GHI	DNI
Annual	5.7	6.1	13.4	2.7	11.6	6.3
Monthly	2.5	2.8	3.4	3.4	1.9	2.7
Daily	-	-	-	-	0.2	0.4
MAE (kWh/m^2)	GHI	DNI	GHI	DNI	GHI	DNI
Annual	4.5	3.7	11.7	2.3	11.1	4.1
Monthly	1.3	0.9	2.4	2.5	1.5	2.2
Daily	-	-	-	-	0.1	0.3

5 Conclusions and future improvements

In this work we present a method for the synthetic generation of 100 PMYs of coupled DNI and GHI at 1-min time resolution using hourly input series of 10-15 years at the site that can be easily retrieved from satellite estimates. The synthetic data can be used for the radiometric characterization of any site in which satellite-derived solar data is available and for risk evaluation of solar projects with the advantage of considering the interannual, seasonal and intra-day variability of the solar radiation.

The method is divided in three main steps in a chain where we start generating synthetic monthly, then daily and then 1-min values. For the generation of synthetic monthly values we have used a Monte Carlo method and we have assumed a normal distribution form the annual values of GHI and a Weibull distribution for the annual values of GHI. The correlations between months has not been addressed in this study. In a future iteration of this work we intend to develop tools to take care of this aspect. For the generation of synthetic daily values we have used a first order autoregressive model and a nonparametric bootstrapping technique. The relation between the daily GHI and DNI is kept and reproduced in the synthetic data. For the downscaling from synthetic daily to 1-min values, we have used an algorithm based on the normalization of the daily solar radiation profiles by the clear-sky envelope approach and the extraterrestrial solar radiation, creating dimensionless profiles of observed solar radiation data that can be used to generate synthetic 1-min data for any location and any day of the year.

Results show a great concordance between the synthetic and observed data in all the time scales. We reproduce the Normal and Weibull distributions of the annual values maintaining the particular relation between the DNI and GHI along each month of the year. We reproduce the CDF of the observed daily values in terms of the energy, variability and distribution. We have quantified the similarities in the CDFs by using the KSI in order to present comparable results for future approaches.

The presented method sets the base for the multiyear synthetic generation, however, there are several improvable issues in each step of the procedure. In step 1 we only generate synthetic monthly values within the thresholds of the maximum and minimum observed monthly values. Climate change trends should be included in order to generate extreme synthetic monthly values and the worst case scenario (PoE99), should be addressed from volcanic eruptions' assessment. Step 2 would require a larger observed database to gather all possible stochastic variations of the daily solar radiation. Regarding step 3, there is still room of improvement in the intra-day variability characterization. The VI index shows a significant dependence on the day of the year, location and time resolution. Novel indexes should be developed and advanced computational techniques could be used to synthetically generate the information lost when using aggregated hourly/daily values. In parallel, recent high-frequency generation techniques may be applied to achieve this third step

Acknowledgments

This research has been developed under the framework of the International Energy Agency Photovoltaic Power Systems Programme (IEA-PVPS) Task 16. "Solar resource for high penetration and large scale applications".

References

Aguiar, R.J., Collares-Pereira, M. and Conde, J.P., 1988. Simple procedure for generating sequences of daily radiation values using a library of Markov transition matrices. Solar Energy 40 (3), 269-280.

Black, J. N., C. W. Bonython, and J. A. Prescott, Solar radiation and the duration of sunshine, Quart. J. Roy. Met. So¢. 80, 231-235 (1954)

Boland J., 2008. Time series and statistical modelling of solar radiation, Recent Advances in Solar Radiation Modelling, Viorel Badescu (Ed.), Springer-Verlag, pp. 283-312.

Boland, J. 2010. Generation of synthetic sequences of electricity demand with applications, Handbook on "Uncertainty and Environmental Decision Making", in Springer's International Series in Operations Research and Management Science, F. S. Hillier (ed.), pp. 275-314.

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

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

Bureau of Meteorology, 2015. Bureau of Meteorology, 2015. One Minute Solar Data. <http://www.bom.gov.au/climate/data/oneminsolar/stations.shtml>.

Eck, M., Kretschmann, D., Feldhoff, J.F., Wittmann, M., 2012. Considering Uncertainties in Research by Probabilistic Modeling. Paper read at ASME 2012 6th International Conference on Energy Sustainability collocated with the ASME 2012 10th International Conference on Fuel Cell Science, Engineering and Technology.

Espinar, B., Ramírez, L., Drews, A., Beyer, H.G., Zarzalejo, L.F., Polo, J., Martín, L., 2009. Analysis of different comparison parameters applied to solar radiation data from satellite and German radiometric stations. Solar Energy 83 (1), 118–125.

Fernández-Peruchena, C. M., Gastón, M., Sánchez, M., García-Barberena, J., Blanco, M., & Bernardos, A. (2015). MUS: A multiscale stochastic model for generating plausible meteorological years designed for multiyear solar energy yield simulations. Solar Energy, 120, 244-256.

Fernández-Peruchena, C. M., Vignola, F., Gastón, M., Lara-Fanego, V., Ramírez, L., Zarzalejo, L., ... & Pulgar, J. (2018). Probabilistic assessment of concentrated solar power plants yield: The EVA methodology. Renewable and Sustainable Energy Reviews, 91, 802-811.

Grantham AP, Pudney PJ, Boland JW, 2018. Generating synthetic sequences of global horizontal irradiation, Solar Energy, vol. 162, pp. 500-509.

A. Grantham, P. Pudney, L. Ward, M. Belusko, J. Boland. Generating synthetic five-minute solar irradiance values from hourly observations Sol. Energy, 147 (2017), pp. 209-221

Hall, I. J., Prairie, R. R., Anderson, H. E., & Boes, E. C. (1978). Generation of a typical meteorological year (No. SAND-78-1096C; CONF-780639-1). Sandia Labs., Albuquerque, NM (USA).

Ho, C.K., Kolb, G.J., 2010. Incorporating uncertainty into probabilistic performance models of concentrating solar power plants. J. Sol. Energy Eng. 132 (3), 031012.

IEA, 2014. Technology Roadmap. Solar Thermal Electricity

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

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

Lohmann, G., 2018. Irradiance Variability Quantification and Small-Scale Averaging in Space and Time: A Short Review. Atmosphere MDPI 9, 1-22

Magnano L., Boland J. and Hyndman R., 2008. Generation of synthetic sequences of half-hourly temperature, Environmetrics, 19(8), pp. 818-835.

McArthur, L. B. J. 2004 “Baseline Surface Radiation Network (BSRN): Operations Manual (Version 2.1),”

Moreno-Tejera, Sara; Silva-Pérez, Manuel Antonio; Lillo-Bravo, Isidoro; Ramírez-Santigosa, Lourdes. 2016. Solar resource assessment in Seville, Spain. Statistical characterisation of solar radiation at different time resolutions. Solar Energy. 132: 430-441.

Moreno-Tejera, Sara; Silva-Pérez, Manuel Antonio; Ramírez-santigosa, Lourdes; Lillo-Bravo, Isidoro. 2017. Classification of days according to DNI profiles using clustering techniques. Solar Energy. 146: 319-333.

Morf, H., 1998. The stochastic two-state solar irradiance model (STSIM), Solar Energy 62, 101-112.

Munkhammar, J., Widén, J., 2018. An N-state Markov-chain mixture distribution model of the clear-sky index. Solar Energy 173, 487-495

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

Nielsen, K. P., Blanc, P., Vignola, F., Ramirez, L., Blanco, M., & Meyer, R. (2017). Discussion of currently used practices for:" Creation of Meteorological Data Sets for CSP/STE Performance Simulations (Doctoral dissertation, IEA SolarPACES).

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

Peruchena, C. M. F., Ramírez, L., Silva-Pérez, M. A., Lara, V., Bermejo, D., Gastón, M., ... & Gonzalez, R. (2016). A statistical characterization of the long-term solar resource: towards risk assessment for solar power projects. Solar Energy, 123, 29-39.

Peruchena, C. M, Gastón, M., Sánchez, M., García-Barberena, J., Blanco, M., Bernardos, A., 2015 MUS: a multiscale stochastic model for generating plausible meteorological years designed for multiyear solar energy yield simulations. Sol. Energy, 120 (2015), pp. 244-256

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

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

Piantadosi, J; Boland, J; Howlett, P, 2009. Generating synthetic rainfall on various timescales - daily, monthly and yearly, Environmental Modeling and Assessment, 14 (4), pages 431-438.

Ramirez, L., Nielsena, K. P., Vignola, F., Blanco, M., Blanc, P., Meyer, R., & Wilbert, S. (2017). Road Map for Creation of Advanced Meteorological Data Sets for CSP Performance Simulations (Doctoral dissertation, IEA SolarPACES).

Röttinger, N., Remann, F., Meyer, R., Telsnig, T., 2015. Calculation of CSP yields with probabilistic meteorological data sets: a case study in Brazil. Energy Procedia, 69, 2009-2018.

Skartveit, A., Olseth, J.A., 1992. The probability density and autocorrelation of short-term global and beam irradiance. Sol. Energy 49, 477–487.

Stein, J., Hansen, C., Reno, M., 2012. The variability index: a new and novel metric for quantifying irradiance and PV output variability. World Renew. Energy 1–7.

Usaola, J., 2014. Synthesis of hourly wind power series using the Moving Block Bootstrap method, in: 2014 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS). IEEE, pp. 1–6. doi:10.1109/PMAPS.2014.6960602

Wilcox, S., & Marion, W. (2008). User's manual for TMY3 data sets (pp. 4-5). Golden, CO: National Renewable Energy Laboratory.