**ABSTRACT**

Computationally generating irradiance data (known as synthetic solar irradiance) is an appealing approach when collecting irradiance information from observed historical measurements or derived from satellite estimates are not feasible. Because the solar photovoltaic energy is intermittent by nature due to variable weather conditions, a well-known solar irradiance data is required to consolidate a reliable performance evaluation of photovoltaic systems in terms of achieving accurate estimates of operational behavior, energy production and efficiency. In this article an autoregressive stochastic and an autoregressive bootstrap method are proposed to generate synthetic solar irradiance data for resolution greater than daily. Both methods ensure a statistically representativeness of the observed data by incorporating dynamic meteorological behavior while maintaining physical significance and capturing the intermittent nature of solar irradiance in the synthetic data sets. Validations through variability metrics, statistical distribution comparison, and energy production shows a mean average percentage error down to 2.74% in intraday irradiance fluctuations, up to 88% in statistical distribution goodness to fit, and close to 2.91% in the energy production. Synthetic cell temperature was also estimated, showing a better catch of the underlying physics and meteorological dynamics of the observed data, than a well-known and widely used physical model.

1. **INTRODUCTION**

The energy industry is migrating from the use of non-renewable fossil resources towards clean sources that imply a minimum environmental impact both in their generation and in their subsequent use. This change is observed in the evolution of global energy consumption by generation source, where renewable energies, specifically photovoltaic (PV) solar energy, have the highest growth rate (12%, approximately) compared to traditional sources, according to the 2020 Annual Energy Outlook by the US Energy Information Administration (EIA). In addition, entities such as the International Renewable Energy Agency (IRENA) validate this argument by stating that solar PV energy has positioned itself as the most widely used renewable energy in the last 15 years, while also highlighting the fact that solar projects now offer some of the cheapest sources of electricity in history.

However, the main weakness of solar PV energy is its intermittent and stochastic nature due to variable weather conditions (e.g., cloud cover and turbidity) caused by geographic factors, daily weather changes, and seasons, which result in a reduction in power generation.

Therefore, a precise characterization of the PV generation main input (i.e., solar irradiance) allows for the consolidation of a reliable performance evaluation of PV systems, which, in turn, guides decision making towards reaching the expected efficiency of the photovoltaic system [1]. Thus, having this climatic information at the design stages (i.e., pre-feasibility stage) is increasingly important to achieve accurate estimates of operational behavior and energy production.

The climatological information of a specific place required for the prefeasibility stage can be acquired mainly through two strategies: from observed historical data collected over many years and from satellite-derived estimates. However, the delays and high costs involved in measuring data in different locations and over long periods of time (i.e., space-time variations) are the major difficulties with these methods. Although the satellite strategy is suitable to capture these variations, the accuracy achieved is questionable due to the spatial resolution of the data it generates. These reasons give rise to a recent strategy that is becoming popular: computationally generating irradiance data series, known as synthetic solar irradiance.

Synthetic solar irradiance is a realistic time series that is statistically representative of actual observed data. Since the synthetic data is subject to the actual system production and weather conditions on a specific date, the consistency of the irradiance fluctuation dynamics with the meteorological state is assured. The literature highlights that synthetic solar irradiation data must mainly comply with capturing a variability equivalent to the real data and producing representative data of the observation data for the target location and time [2]. In addition, synthetic solar irradiation data should be generated with a computationally efficient algorithm and be easy to access and reproduce. Furthermore, one of the most relevant advantages of having a set of synthetic data is knowing the status of each data point and having the ability to statistically characterize stochastic behavior.

State-of-the-art.

In this article, two algorithms based on autoregressive models are proposed for the synthetic generation of solar irradiance data: with a stochastic method and with the bootstrap technique. The input of these algorithms are only the measurements of historical solar irradiance data. The proposed algorithms manage to estimate realistic values ​​for any temporal resolution greater than daily. In addition, the algorithms’ development is based on the characteristics and limitations of the commented models in the literature. For instance, synthetic solar irradiance data samples follow independent and identical probability distributions [2]; incorporate dynamic meteorological behavior while maintaining statistical and physical significance [3, 4]; and capture the intermittent nature of solar irradiance by managing to be generated with a temporal resolution of less than one hour [5].

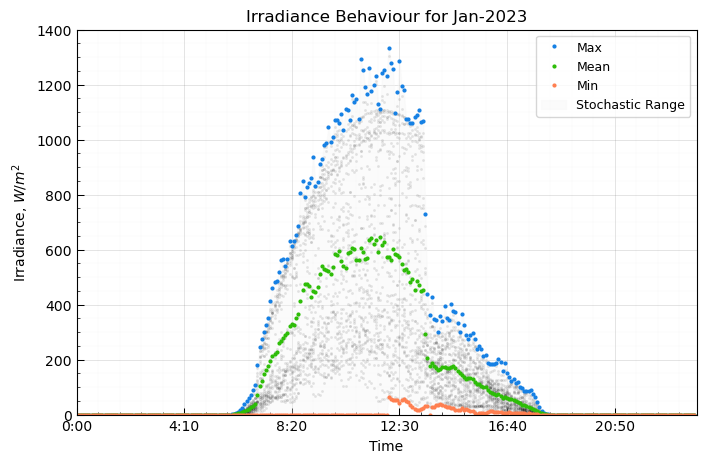
Furthermore, to conserve meteorological physics, two boundary conditions are imposed [3]:

* The overall potential of the solar resource must remain slightly constant. In other words, the daily and monthly average sum of the synthetic irradiance values must agree with the mathematical estimates.
* The dynamics of the fluctuations must be consistent with the state of the sky. That is, the frequency and amplitude of the fluctuations in the synthetic irradiance time series must be statistically representative with the measured conditions.

This paper is organized as follows: Section 2 describe used data set, location, theoretical methods, and validation approaches. Section 3 presents and discuss the results obtained. Finally, Section 5 presents the conclusions of the paper.

1. **METHODS SECTION**

Synthetic irradiance data is subject to weather conditions for a specific date and location, which are acquired by a meteorological monitoring system. In this case, the data acquisition system incorporated in the meteorological station of the University of Los Andes located in Bogotá D.C., Colombia. The initial data set is made up of measurements in five-minute resolutions during the years 2020 to early 2023. For greater clarity, an example is shown in Figure 1 including only the month of January 2023 and shows the measurements of daily irradiance. Maximum, average, and minimum irradiance statistics are also included. The gray fill indicates the range in which the algorithms will estimate synthetic irradiance values to maintain climatic physics.



**Fig. 1.** Maximum, average, and minimum behavior of daily irradiance for January 2023.

* 1. **Sky Categorization**

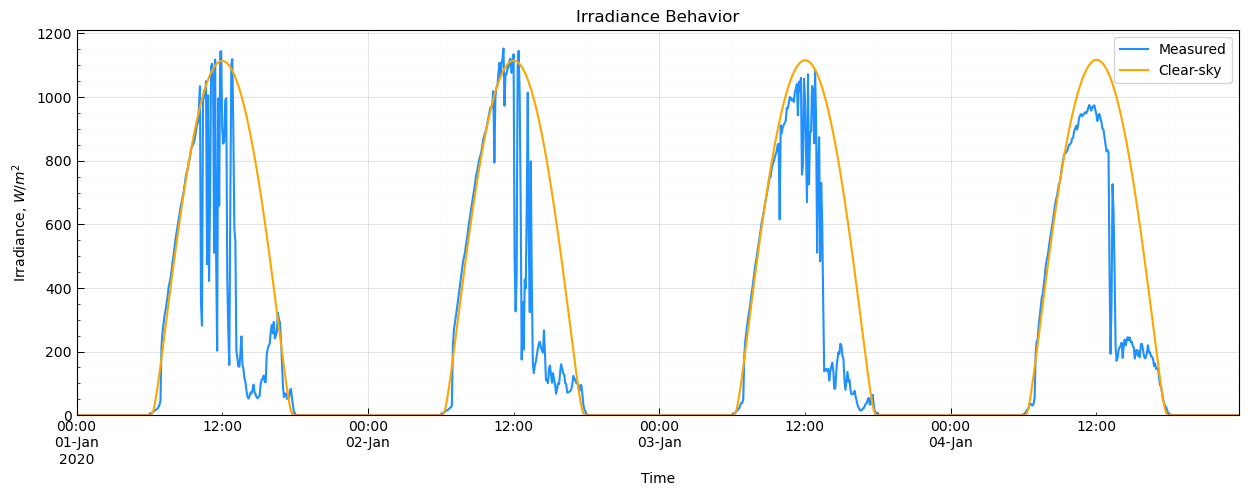
The climatic state and its dynamic nature are characterized by the metricclear-sky index , which is the irradiance in the plane of array normalized by the irradiance under clear sky conditions (i.e., ideal scenario) [14]. From this metric it is possible to estimate the meteorological conditions of each day of the month [4].

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

**Table 1.** Sky condition according to the value of . Adapted from [4].

|  |  |  |  |
| --- | --- | --- | --- |
| Range of | Denomination | Sky Condition | Description |
| [ 0 – 0.2 ] | SC1 | Fully covered | Thick clouds with little or no fluctuation |
| ( 0.2 – 0.4 ] | SC2 | Mostly covered | Alternate clouds and clear sky, some, or large fluctuations |
| ( 0.4 – 0.6 ] | SC3 | Partially covered | Thin clouds with some or large fluctuations and/or haze |
| ( 0.6 – 0.67 ] | SC4 | Mostly clear | No clouds but with some turbidity |
| ( 0.67 – 1 ] | SC5 | Totally clear | No clouds (clear sky) |

The values of are obtained with the pvlib-python library, while comes from measurements in situ. Figure 2 shows the behavior of these irradiances for the first week of January 2023. The values of are theoretically maximum. However, achieving this profile with little variability as seen in the orange curve in Figure 2 is not common. The instances in which exceeds can be caused by reflections in the reference cell. Therefore, it must be filtered.

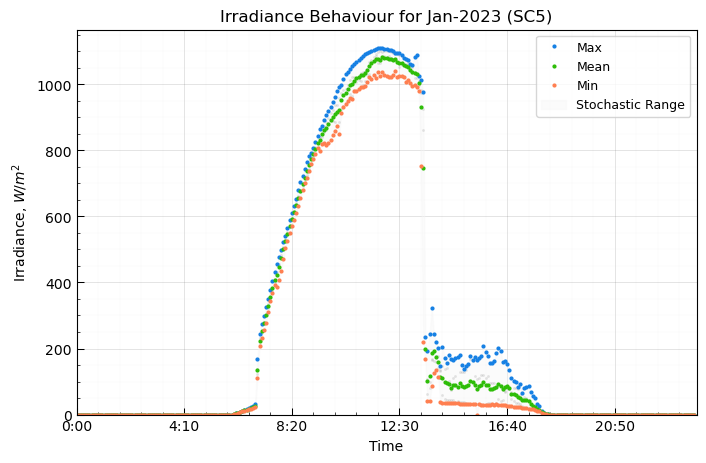
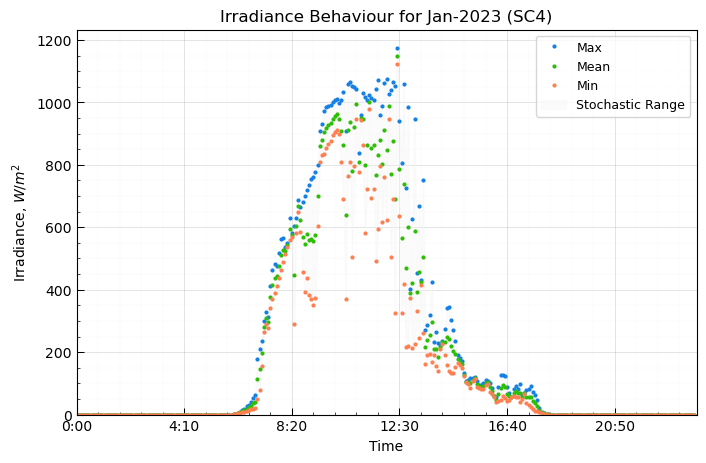
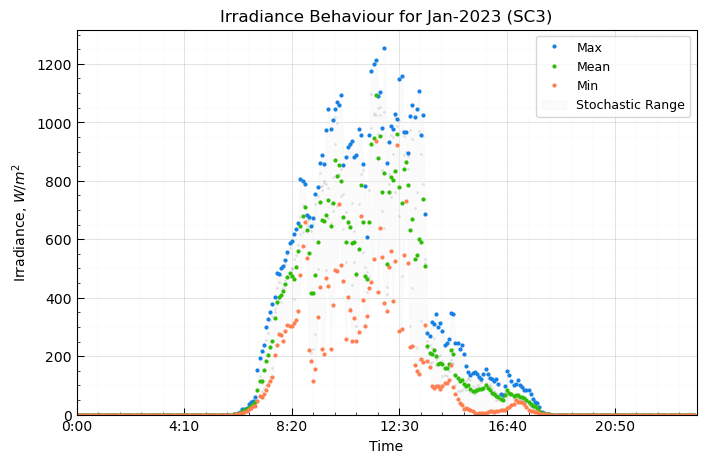
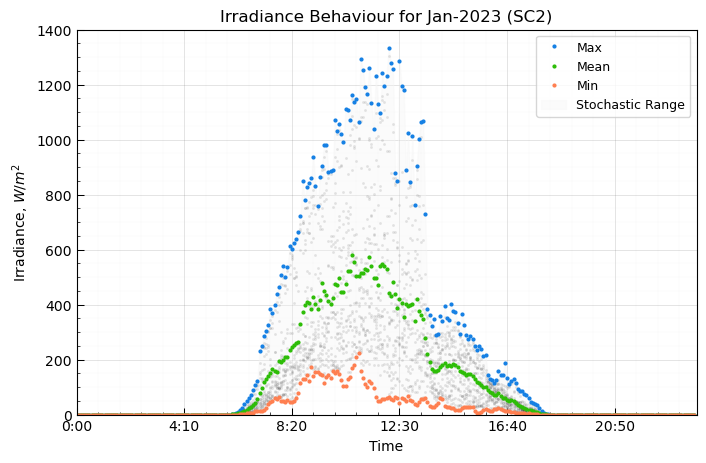


**Fig. 2.** Comparison between and behaviors for January 2023.

Thus, the metric is estimated to categorize the climatic behavior of each time stamp. Each value of is accompanied by a unique time stamp (e.g., 12:00:00 01/01/2023). Finally, the categorization of each day’s sky condition is obtained by averaging the value of of each time stamp for the corresponding daily data set. For each month, a set is made up of the discretized days according to the sky condition obtained to analyze each set individually. Figure 3 presents the irradiance measurements for each data set according to the categorization given by the metric for the example case of January 2023. Table 2 shows each sky condition set.

**Table 2.** Categorization of the sky condition of each day of January 2023.

|  |  |
| --- | --- |
| Sky Condition | Days |
| Fully covered | 7, 10, 12 |
| Mostly covered | 1, 2, 3, 4, 5, 6, 8, 9, 11, 13, 14, 17, 18, 19, 20, 21, 23, 24 |
| Partially covered | 15, 16, 22, 31 |
| Mostly clear | 25, 30 |
| Totally clear | 26, 27, 28, 29 |



**Fig. 3.** Daily irradiance for January 2023 categorized according to the sky condition given by .

Knowing the sky condition, the algorithms better capture the meteorological dynamics of each set of days and ensures that the synthetic irradiance value generated is physically consistent. For instance, it is observed that as the climatic conditions resemble a clear sky, the minimum and maximum irradiance values are closer to each other.

* 1. **Statistical Distributions**

The Shapiro-Wilk and Anderson-Darling tests are used on a data set to assess its fit to a normal or lognormal distribution. The data set corresponds to the irradiance values for a specific date and timestamp. For the example case (January 2023), three data sets for a time stamp were validated in each test: the entire data of the date, data belonging to the sky condition set mostly covered (SC2) of the specific date, and data belonging to the totally clear climate set (SC5) of the specific date. In addition, the statistical tests were accompanied by histograms and QQ-plots. The amount of data in each set is important to maintain statistical representativeness. The Gaussian or lognormal distribution is mandatory to proceed with the proposed methodology.

* 1. **Autoregressive Basis**

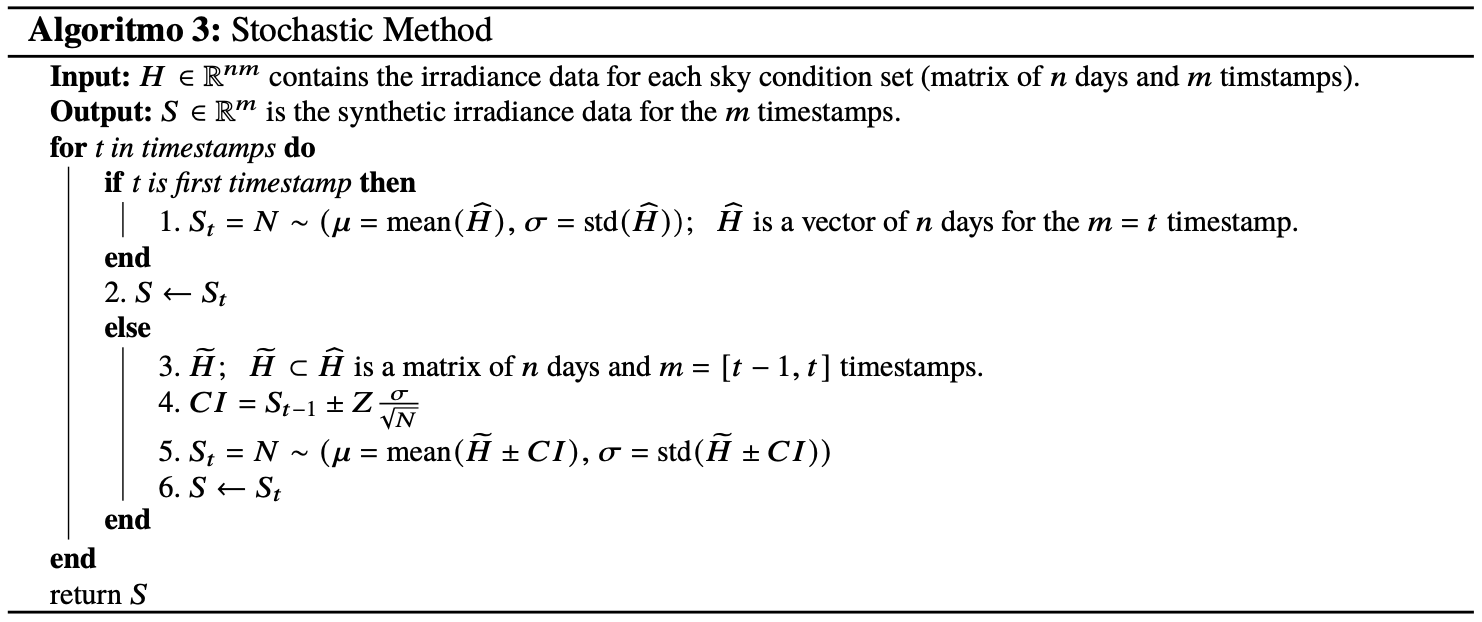
Autoregressive models (AR) are a good enough approach to reproduce irradiance time series. An AR is a random process where the value in each time step depends on the previous values [2]. For instance, Gaussian and time dependent AR models developed by Graham & Hollands and Aguiar & Collares-Pereira are considered state-of-the-art methodologies as they are implemented in well-known solar PV software such as HOMER Pro and PVsyst for the generation of synthetic hourly solar irradiance data [6, 7].

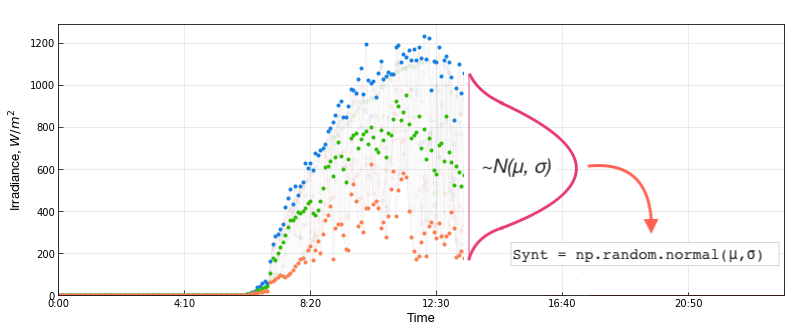
The proposed AR basis consists of establishing a confidence interval in which the synthetic irradiance will be generated at time step by considering the synthetic irradiance data previously generated at time step . The confidence interval is based on the Gaussian distribution standardized value for 95% accuracy. In this way, the most appropriate range is defined in which the synthetic irradiance value is most likely to lie for the next time step .

* 1. **Stochastic Method**

The stochastic method takes advantage of the statistical distribution of the set of values of a specific sky condition to generate the synthetic data. The operating principle consists of estimating a synthetic value of irradiance from the Gaussian distribution of the corresponding confidence interval for each time step. The procedure of the stochastic method is detailed in Algorithm 2. Also, Figure 5 illustrates the synthetic data estimation process. Although the Gaussian distribution is ideal, a lognormal statistical fit is also valid.

**Algorithm 2.** Stochastic method.



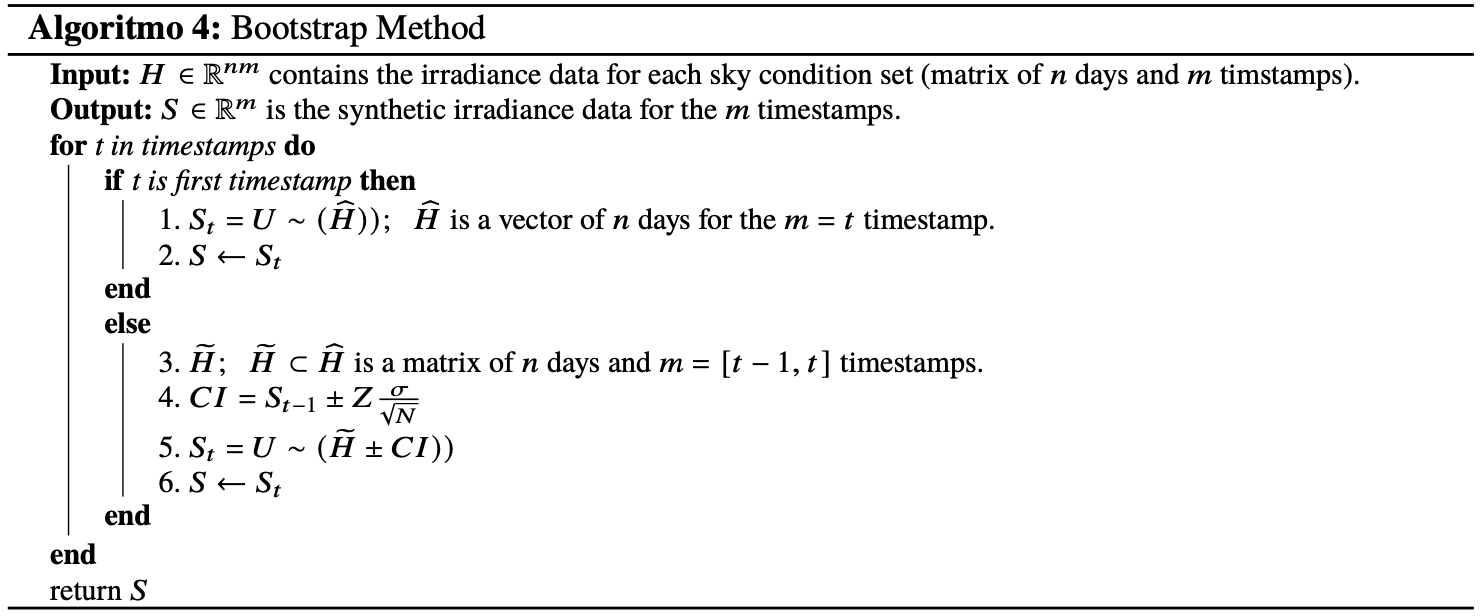


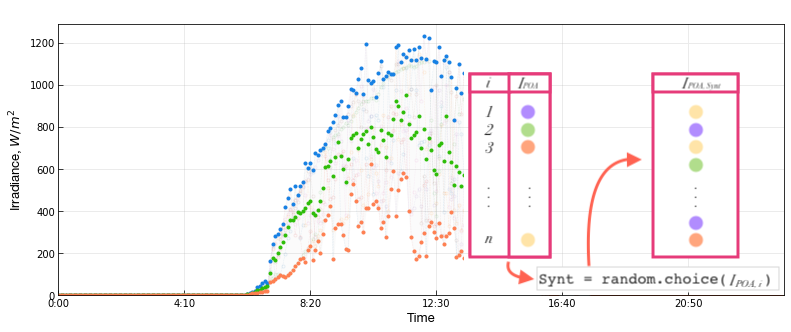
**Fig. 5.** Illustration of the generation of synthetic data with the stochastic method.

* 1. **Bootstrap Method**

The bootstrap method takes a value from the original data set of a specific date and timestamp and categorized sky condition to generate the synthetic irradiance data. The selection of the data point is random by following a uniform distribution, and with replacement, that is, the selected data point can be extracted again. The procedure of this method is detailed in Algorithm 3. In addition, Figure 6 illustrates the synthetic data estimation process with the bootstrap method.

**Algorithm 3.** Bootstrap method.





**Fig. 6.** Illustration of the generation of synthetic data with the bootstrap method.

* 1. **Validation**

Synthetic solar irradiance time series should be statistically representative of the measured observations, as they capture its inherent variability. Therefore, the goodness of fit between the synthetic and actual data distributions should be verified to confirm the credibility of the synthetic solar irradiance [2]. Three validation approaches are chosen: variability metrics, statistical distribution, and energy production. Each validation is performed for each date and sky condition separately.

Variability metrics shows how much the synthetic solar irradiance time series capture the dynamic behavior features of the observed data. Specifically, the standard deviation of increments (SDI), stability index (SI), integrated complementary cumulative distribution function (ICCDF), and variability index (VI) are evaluated. This metrics capture the increment variability between time steps. Each variability metric is evaluated over the synthetic and observed time series, and then compared by calculating the percentage error between them.

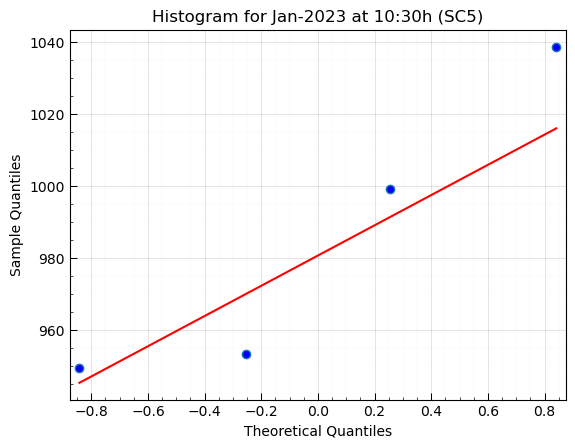
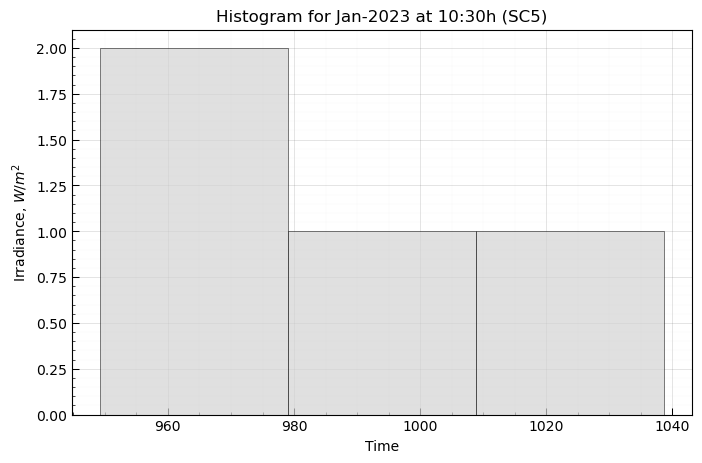
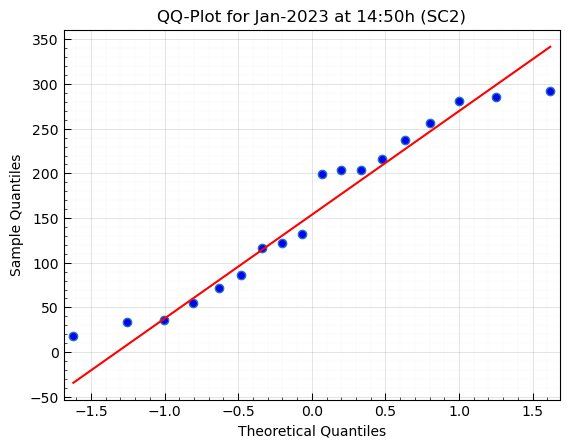
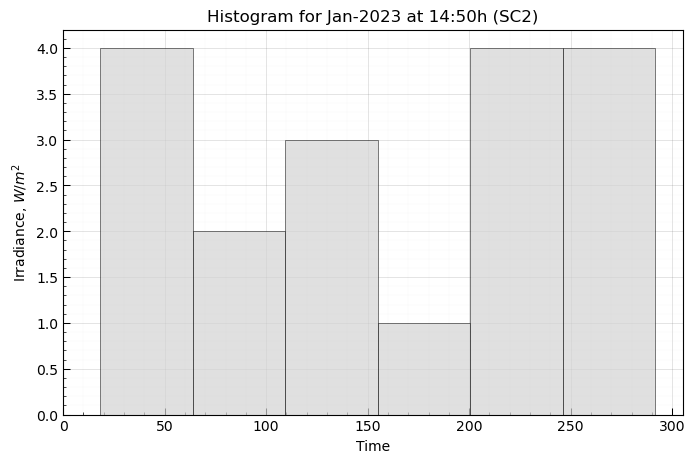
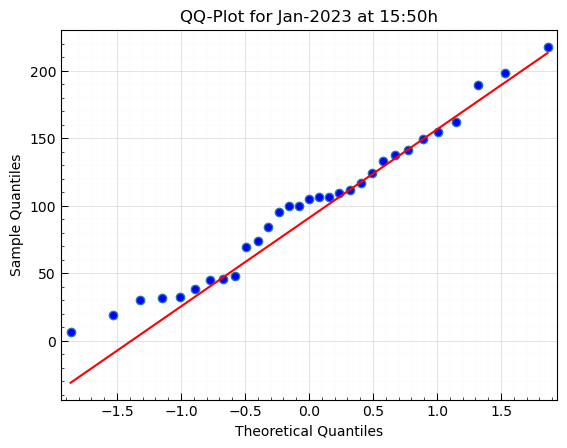
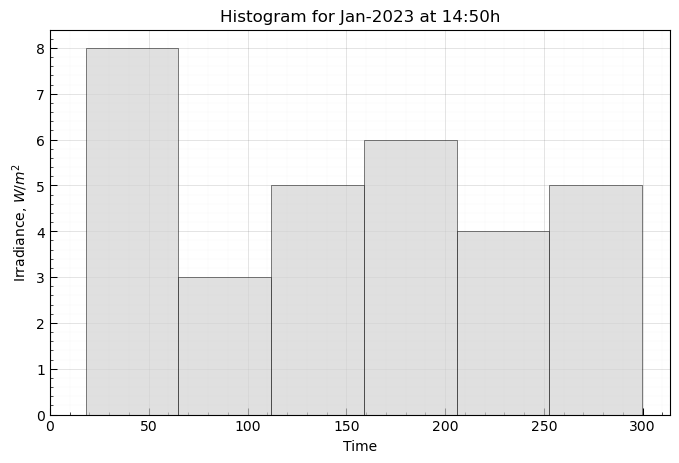
In addition, a comparison of statistical distribution determines whether synthetic irradiance data fits the observation data. For this, Kolmogorov-Smirnov test (KS), Kullback-Leibler divergence (KLD), and overlapping coefficient (OVC) are evaluated both in the synthetic and observed irradiance time series to determine the goodness of fit.

Finally, as the synthetic solar irradiance is a realistic representation of the observed irradiance time series, and it is typically used for a performance evaluation of PV systems, the energy generated should be also reliable. The energy estimation workflow is developed in pvlib-python. The direct current (DC) production is simulated with the California Energy Comission (CEC) model for the single diode equation, and the alternating current (AC) production is simulated with the Sandia’s Grid-Connected PV Inverter model [8]. As for the variability metrics, the energy is evaluated both for the synthetic and observed irradiance time series, and then compared by calculating the percentage error between them. The solar PV system architecture to account for the simulation of the energy generation is based on the 80 kWp University of Los Andes solar PV system located on the rooftop of the Santo Domingo building.

1. **RESULTS AND DISCUSSION**
   1. **Statistical Distributions**

The Shapiro-Wilk and Anderson-Darling tests were assessed in a data set corresponding to the irradiance values ​​for a specific date, timestamp, and sky condition. Three data sets were validated in each test for a specific date: the totality of the data, data belonging to the mostly covered climate set (SC2), and data belonging to the totally clear climate set (SC5). The results indicate that the most prominent fits in the evaluated data set is the Gaussian distribution, followed by the lognormal distribution. The most representative results are given when the complete data set is evaluated, given that there are more data points. However, although sets SC2 and CS5 notably contain less data, both statistical and visual tests confirm the fit between these two distributions.

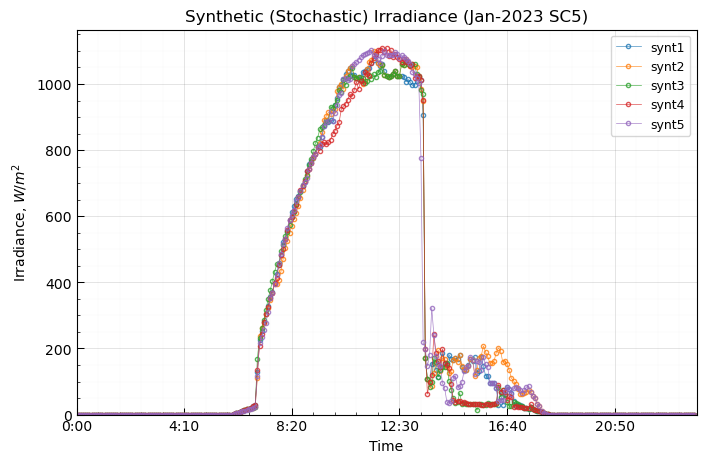
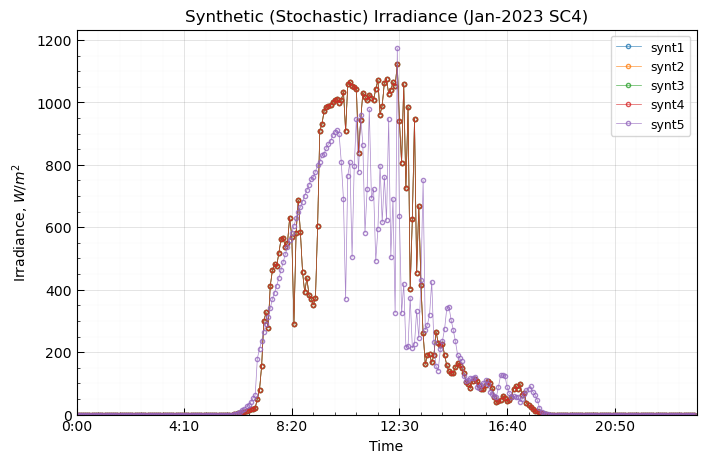
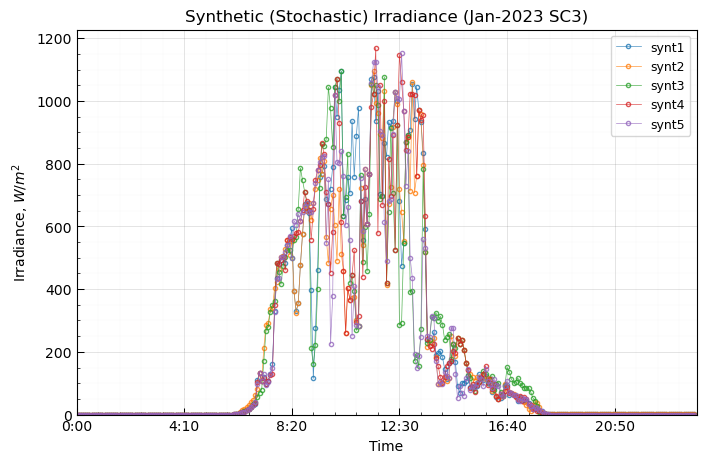
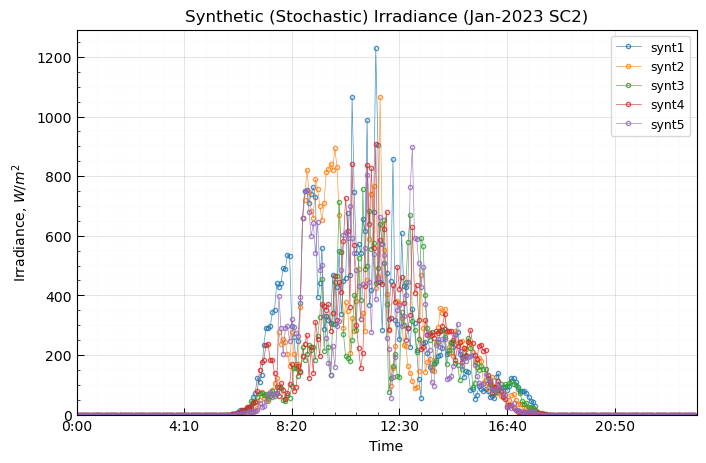
Figure 1 shows the histograms and QQ-plot for the date January 2023 at 14:50h timestamp for the three data sets; the Shapiro-Wilk and Anderson-Darling tests verify a better fit with the Gaussian distribution.



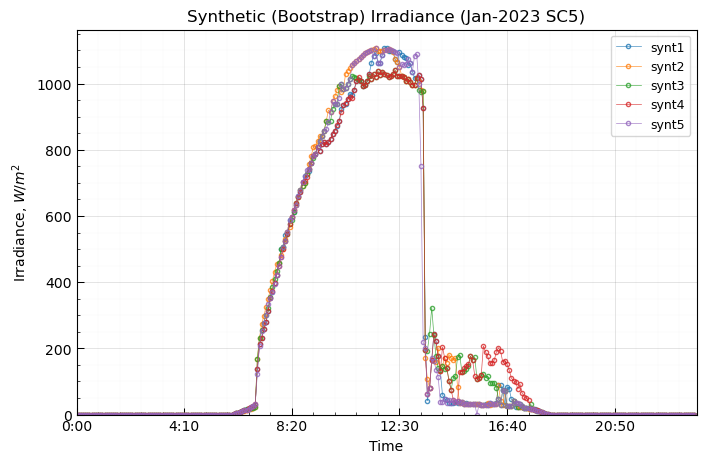
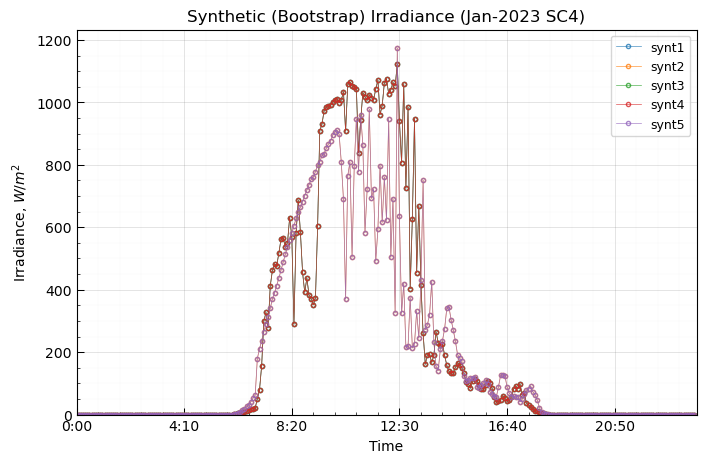
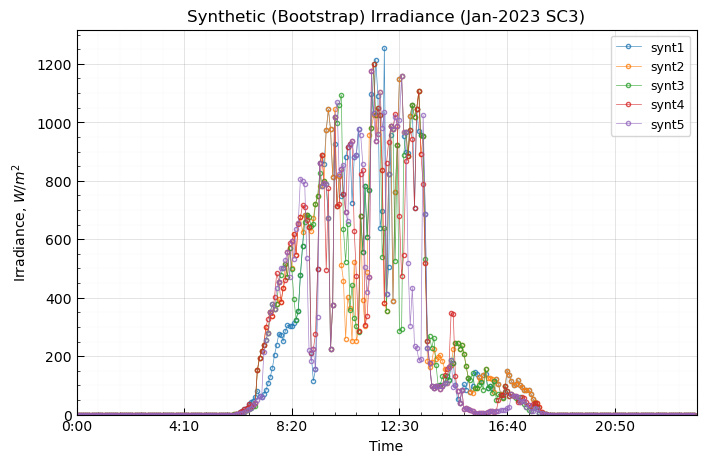
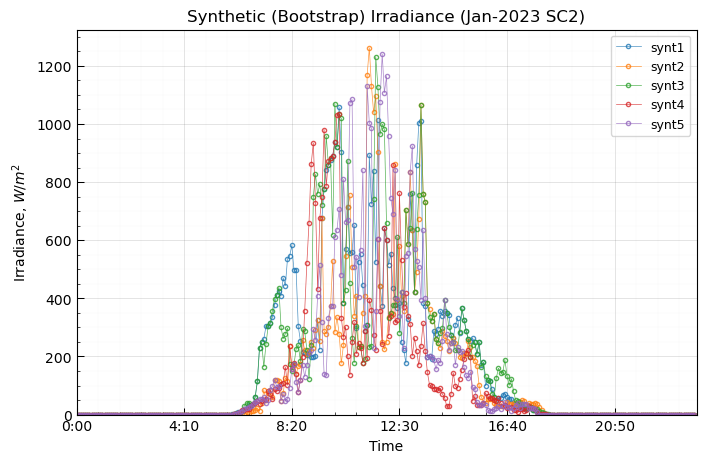
**Fig. 7.** Histogram and QQ-Plot to evaluate the Gaussian or lognormal fit of the data. Studies for Jan-2023 with the full dataset (top), SC2 dataset (intermediate), and SC5 dataset (bottom).

* 1. **Synthetic Irradiance**

Figures 8 and 9 show five synthetic irradiance time series generated by the autoregressive stochastic and bootstrap methods, respectively, for four sky conditions.Comparing Figures 8 and 9 with the real values ​​(Fig. 3), it is visually evident that the synthetic irradiance incorporates the physically significant dynamic behavior of climatic conditions, and that the five-minute resolution succeeds in capturing the intermittent nature of solar irradiance. For instance, observing the behavior between 10:00 a.m. and 2:00 p.m. for the mostly clear sky condition (SC2), the synthetic solar irradiance methods manage to capture the erratic behavior caused by frequent rains. Also, for the totally clear sky condition (SC5), a similar irradiance behavior is evident with an approximate maximum peak of 1 100 W/m2, in addition to the precipitous drop around 2:00 p.m. caused by shading of the reference cell.



**Fig. 8.** Synthetic irradiance for the month Jan-2023 categorized according to the climatic state given by the metric and generated by the autoregressive stochastic method.



**Fig. 9.** Daily synthetic irradiance for the month Jan-2023 categorized according to the climatic state given by the metric and generated by the autoregressive bootstrap method.

* 1. **Validation**

Five synthetic solar irradiance time series were generated for each month, year, and sky condition available in the observed data set (i.e., from 2020 to January 2023). The three validation approaches were evaluated by grouping each sky condition.

* + 1. **Variability Metrics**

Table 3 shows the mean average percentage error of the variability metrics for each sky condition. For all the metrics, the main trend is a decrease in the percentage error as the sky condition approaches a totally clear day (SC5). This is as expected because a fully covered (SC1) day has the biggest difference between maximum and minimum irradiance values for each timestamp, and as the condition approaches a clear sky condition this difference narrows remarkably. The mean and median of all sky conditions is also presented. The fact that these descriptive statistics are close to each other means that there are no outliers in the data set generated. In general, the mean average percentage error of all the sky conditions throughout the variability metrics varies between 3.74 to 7.40% for the autoregressive stochastic method, and between 2.74 to 7.50% for the autoregressive bootstrap method.

The biggest percentage error more often appears to be related to SC2. This is most likely because this sky condition (SC2) contains more days in its data set than SC1, and the fact of having a greater number of days possibly incurs in a greater variation.

Even though the threshold for the stability index (SI) was set at 100 W/m2, being this value stricter than what is suggested by the literature (500 W/m2) [9], no increase or decay was detected between time steps that exceeded the threshold both in synthetic and observed solar irradiance time series.

**Table 3.** Percentage error of the variability metrics for each sky condition of the complete synthetic solar irradiance data set generated from both methods.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Sky Condition** | **Stochastic** | | | | **Bootstrap** | | | |
| **SDI** | **SI** | **ICCDF** | **VI** | **SDI** | **SI** | **ICCDF** | **VI** |
| SC1 | 13.72 | 0 | 7.37 | 8.68 | 9.42 | 0 | 4.34 | 7.40 |
| SC2 | 9.9 | 0 | 6.67 | 11.79 | 14.43 | 0 | 7.46 | 13.61 |
| SC3 | 6.26 | 0 | 3.75 | 5.49 | 6.75 | 0 | 2.74 | 7.54 |
| SC4 | 2.02 | 0 | 0.43 | 2.12 | 0.21 | 0 | 0.10 | 0.47 |
| SC5 | 5.13 | 0 | 0.48 | 5.07 | 6.73 | 0 | 0.84 | 8.0 |
| **Mean** | 7.40 | 0 | 3.74 | 6.63 | 7.50 | 0 | 3.09 | 7.40 |
| **Median** | 6.26 | 0 | 3.75 | 5.49 | 6.75 | 0 | 2.74 | 7.54 |

* + 1. **Statistical Distribution**

Table 4 shows the comparison of statistical distribution metrics for each sky condition. For this validation approach consistent results regardless of sky condition is expected, which is the case.

For the two-sided Kolmogorov-Smirnov metric, the null hypothesis is that the distributions from synthetic and observed solar irradiance time series are identical; the alternative is that they are not identical. Therefore, a p-value bigger than the 0.05 threshold is expected to state that both data sets come from the same statistical distribution. The lowest mean average was obtained for SC2 (0.19) for the stochastic method, and for SC4 (0.24) for the bootstrap method, and both are still far from the threshold. Same as for the variability metrics, the mean and median of all sky conditions is also presented, and because they are close to each other, there are no outliers in the data set generated. These descriptive statistics are also close between all statistical distribution metrics.

The KLD metric estimate the average amount of information shared between synthetic and observed distributions. Therefore, the value of the KLD metric must be close to 0 to determine that the information conveyed by the two distributions is not significantly different. In other words, if the KLD metric is zero, then the synthetic data set distribution fits perfectly the true underlying distribution of observed data [9]. In average, the closest distribution fit was found for the mostly clear sky condition (SC4) with a value of 0.04 for the stochastic method, and 0.05 for the bootstrap method. In the same way, the biggest discrepancy was accounted for the mostly cover sky condition (SC2) with values of 0.36 and 0.5 for stochastic and bootstrap methods, respectively.

The OVC metric measures the agreement between two probability distributions, where a value of 1 is a perfect fit while a value of 0 is a totally disjointed densities [9]. Surprisingly, for both methods, the average value of the OVC metric is approximately 0.45, which is a lower result than expected. For individual OVC tests (i.e., not averaged as is the case), the values found for the OVC metric were around 0.78 on average. This could imply that this metric varies notably between months. Therefore, the same OVC metric was performed without averaging the values by sky condition nor by month, and the results showed a variation between 0.78 and 0.88 for 25 and 75th percentiles. This implies that the OVC metric is closely related to each month, year, and sky condition behavior, and is not appropriate to average them.

**Table 4.** Percentage error of the variability metrics for each sky condition of the complete synthetic solar irradiance data set generated from both methods.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Sky Condition** | **Stochastic** | | | **Bootstrap** | | |
| **KS** | **KLD** | **OVC** | **KS** | **KLD** | **OVC** |
| SC1 | 0.44 | 0.22 | 0.49 | 0.44 | 0.26 | 0.49 |
| SC2 | 0.19 | 0.36 | 0.56 | 0.26 | 0.50 | 0.57 |
| SC3 | 0.39 | 0.20 | 0.54 | 0.43 | 0.32 | 0.55 |
| SC4 | 0.24 | 0.04 | 0.26 | 0.24 | 0.05 | 0.26 |
| SC5 | 0.33 | 0.11 | 0.39 | 0.32 | 0.13 | 0.39 |
| **Mean** | 0.31 | 0.18 | 0.44 | 0.33 | 0.25 | 0.45 |
| **Median** | 0.33 | 0.20 | 0.49 | 0.32 | 0.26 | 0.49 |

* + 1. **Energy Production**

Table 5 shows the mean average percentage error of the energy production metric for each sky condition. As is presented, both synthetic solar irradiance generation methods demonstrate that they can generate an energy value equivalent to that estimated with the observed data in a simulated production workflow, which verifies that the synthetic time series does capture the physics and meteorological dynamics of the real observations.

Again, the mean and median of all sky conditions is also presented and is inferred that there are no outliers in the data set generated. On average, the biggest discrepancy was obtained for a mostly cover sky condition (SC2) with a percentage error of 8.08 and 6.90% for the stochastic and bootstrap method, respectively. In the same way, the lower discrepancy is for the mostly clear sky condition (SC4) with values of 0.14 and 0.19% for the stochastic and bootstrap method, respectively. Also, the mean average throughout all sky conditions is 3.34% for the stochastic method, and 2.91% for the bootstrap method. This implies that the last is a slightly better approach for generating synthetic solar irradiance time series, and the reason for the better proximity of the simulated energy production from the measured observations if because the bootstrap technique is based on selecting real data points instead of generating a totally synthetic one.

**Table 5.** Percentage error of the energy production metric for each sky condition of the complete synthetic solar irradiance data set generated from both methods.

|  |  |  |
| --- | --- | --- |
| **Sky Condition** | **Stochastic** | **Bootstrap** |
| SC1 | 4.88 | 4.35 |
| SC2 | 8.08 | 6.90 |
| SC3 | 3.20 | 2.61 |
| SC4 | 0.14 | 0.19 |
| SC5 | 0.41 | 0.51 |
| **Mean** | 3.34 | 2.91 |
| **Median** | 3.20 | 2.61 |

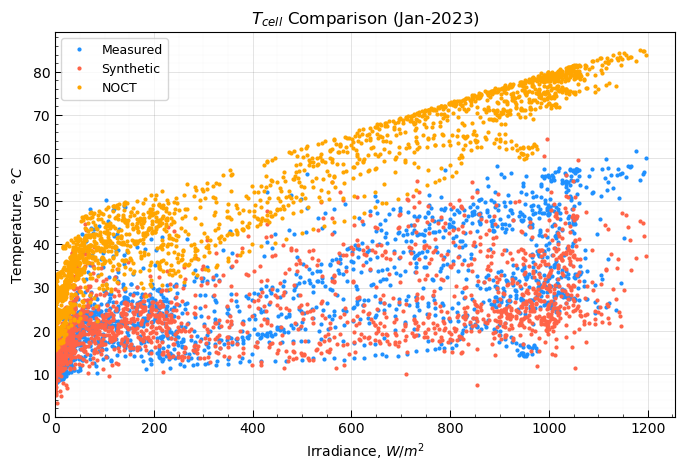
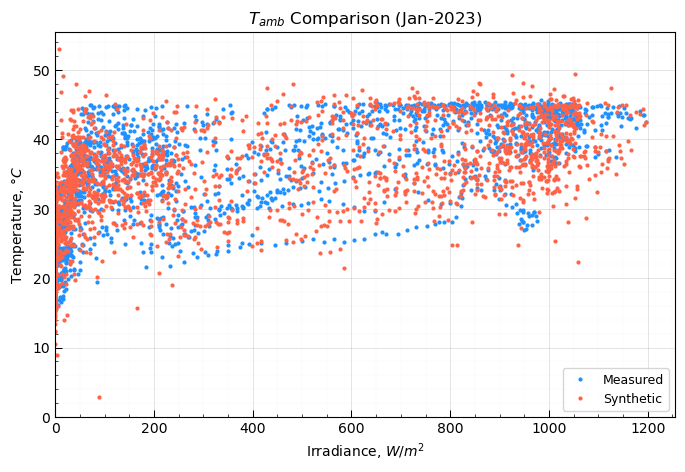
* 1. **Synthetic Ambient and Cell Temperature**

The autoregressive stochastic method was used to determine synthetic ambient and cell temperature, which are the second main resource for the solar PV systems, next to solar irradiance. However, in Algorithm 2, the confidence interval still filters the observed irradiance data set, but also filters the related ambient and cell temperature. Then, instead of generating a synthetic data point from the Gaussian distribution with mean and standard deviation of the irradiance filtered by the confidence interval, the synthetic ambient and temperature data point is generated from the Gaussian distribution but with the mean and standard deviation of the ambient and cell temperature, respectively.

Figure 10 shows the observed and synthetic ambient and cell temperature data sets for January 2023. Also, the well-known physical NOCT equation is also presented.

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

From a visual inspection it is observed that the synthetic data set maintains the physical consistency better and follows the meteorological dynamics of the real data, than the physical model with the NOCT equation.



**Fig. 10.** Comparison between ambient temperature (left) and cell temperature (right) between synthetic generated, observed, and physically calculated from NOCT equation data sets.

1. **CONCLUSIONS**

The Shapiro-Wilk and Anderson-Darling tests verified the Gaussian and lognormal distributions of the solar irradiance data for specific date and time stamps. From these distributions, a stochastic and a bootstrap method were proposed to generate synthetic solar irradiance time series for resolutions bigger than daily. Both methods are based on an autoregressive basis that establishes a 95% confidence interval in which the synthetic solar irradiance data point was generated. The confidence interval is defined for each time step and considers the previous synthetic data point generated.

The synthetic solar irradiance time series were validated through three approaches: variability metrics, comparison of statistical distribution, and energy production. Four variability metrics were evaluated (i.e., standard deviation of increments, stability index, integrated complementary cumulative distribution function, and variability index). Three goodness of fit metrics were performed (i.e., two-sized Kolmogorov-Smirnov test, Kullback-Leibler divergence, and overlapping coefficient). The energy production metric was based on a California Energy Comission model for the single diode equation and Sandia’s Grid-Connected PV Inverter model workflow, and the system architecture is based on the 80 kWp University of Los Andes solar PV system located on the rooftop of the Santo Domingo building.

The results showed that the mean average percentage error of all the sky conditions throughout the variability metrics varies between 3.74 to 7.40% for the autoregressive stochastic method, and between 2.74 to 7.50% for the autoregressive bootstrap method. For the comparison of statistical distributions, all the two-sided Kolmogorov-Smirnov metric evaluated approves the null hypothesis that the distributions from synthetic and observed solar irradiance time series are identical. In the same way, Kullback-Leibler divergence and overlapping coefficient showed that the synthetic data set distribution fits well the true underlying distribution of observed data, with values of 0.04 for the stochastic method and 0.05 for the bootstrap method for the first metric, and variations between 0.78 and 0.88 for the last.

The KLD metric estimate the average amount of information shared between synthetic and observed distributions. Therefore, the value of the KLD metric must be close to 0 to determine that the information conveyed by the two distributions is not significantly different. In other words, if the KLD metric is zero, then the synthetic data set distribution fits perfectly the true underlying distribution of observed data [9]. In average, the closest distribution fit was found for the mostly clear sky condition (SC4) with a value of 0.04 for the stochastic method, and 0.05 for the bootstrap method. In the same way, the biggest discrepancy was accounted for the mostly cover sky condition (SC2) with values of 0.36 and 0.5 for stochastic and bootstrap methods, respectively. In addition, the mean average percentage error for the energy production throughout all sky conditions was 3.34% for the stochastic method, and 2.91% for the bootstrap method, which implies that the last method is a slightly better approach for generating synthetic solar irradiance time series.

Finally, synthetic ambient and cell temperature were estimated following the autoregressive stochastic method. Results showed that this approach better captures the underlying physics and meteorological dynamics of the observed data, than the physical model with the NOCT equation.

**REFERENCES**

1. Triki-Lahiani et al. (2018). Fault Detection and Monitoring Systems for Photovoltaic Installations: A Review, Renewable and Sustainable Energy Reviews, Vol. 82 (2018), pp. 2680–2692, doi: 10.1016/j.rser.2017.09.101.
2. Joakim Munkhammar and Joakim Widén, "Established Mathematical Approaches for Synthetic Solar Irradiance Data Generation," in Synthetic Solar Irradiance: Modeling Solar Data, edited by Jamie M. Bright [AIP Publishing (online), Melville, New York, 2021], Chapter 3, pp. 3-1–3-34, available at: https://doi.org/10.1063/9780735421820\_003.
3. J. Polo, L.F. Zarzalejo, R. Marchante & A.A. Navarro (2011). A simple approach to the synthetic generation of solar irradiance time series with high temporal resolution. Solar Energy, Vol. 85 (2011), pp. 1164–1170, doi: 10.1016/j.solener.2011.03.011.
4. J.M. Bright, C.J. Smith, P.G. Taylor & R. Crook (2015). Stochastic generation of synthetic minutely irradiance time series derived from mean hourly weather observation data, Solar Energy, Vol. 115 (2015), pp. 229–242, doi: 10.1016/j.solener.2015.02.032.
5. M. Larrañeta, S. 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, Solar Energy, Vol. 122( 2015), pp.517–528, doi: 10.1016/j.solener.2015.09.030.
6. Graham VA, Hollands KGT. A method to generate synthetic hourly solar radiation globally. Solar Energy. 1990; 44:331-41.
7. Aguiar, R, & Collares-Pereira, M. TAG: A time-dependent, autoregressive, Gaussian model for generating synthetic hourly radiation. United States. https://doi.org/10.1016/0038-092X(92)90068-L.
8. William F. Holmgren, Clifford W. Hansen, and Mark A. Mikofski. “pvlib python: a python package for modeling solar energy systems.” Journal of Open Source Software, 3(29), 884, (2018). https://doi.org/10.21105/joss.00884.
9. Âzeddine Frimane and Jamie M. Bright, "Validation of Synthetic Solar Irradiance Data," in Synthetic Solar Irradiance: Modeling Solar Data, edited by Jamie M. Bright [AIP Publishing (online), Melville, New York, 2021], Chapter 4, pp. 4-1–4-44, available at: https://doi.org/10.1063/9780735421820\_004.