Generation of 441 typical meteorological year from INMET stations - Brazil

Rubinei D. Machado[1, 2], Gabriel Bravo[2], Allan Starke[2], Leonardo Lemos[2], Sergio Colle[2]

[1] Institute of Astronomy, Geophysics and Atmospheric Sciences/University of São Paulo, São Paulo (Brazil)

[2] Laboratory of Energy Conversion Engineering and Energy Technology, Department of Mechanical Engineering/ Federal University of Santa Catarina, Florianopolis (Brazil)

Abstract

Since 2005, the Brazilian National Institute of Meteorology (INMET) has developed and operated an extended automatic weather station network. There are about 564 stations currently in operation, each one providing hourly values of several meteorological parameters, including global solar irradiation (G). This long-term period (LT) of data is important to support research and development in the energy production sector, and to assist the development of the national energy matrix. However, due to interannual variability of meteorological variables, the use of typical meteorological year (TMY) data to represent a long-term dataset is crucial to ensure that risk analysis and simulation of solar plants are closer to reality. Thus, this paper presents a methodology to select the 12 representative typical meteorological months of each INMET station. The TMY for 441 cities obtained using Sandia method show that the process of constructing the TMY was well performed and the TMY dataset (including Brasilia's TMY) should be considered representative. This fact is supported by the analysis using the average KSI index over all stations and other statistical parameters. In the end, the evaluate indicated that TMY have a good statistical similarity when compared with original climate datasets.

Keywords: TMY, Energy Meteorology, multi-year analysis, building energy simulation, solar energy, KSI

1. Introduction

Due to the continental extension of Brazil, the country has extratropical, subtropical and tropical climate features (Garreaud et al., 2009). Heterogeneous topography and climatic fluctuations involving atmospheric systems in the temporal (interannual and interdecadal) and spatial scale directly impact the amount of energy produced from renewable sources in the country. Therefore, the climate characterization is extremely important to indicate favorable locations for energy applications and analyze the feasibility of such applications.

According to Hirsch (2017), the financing of complex systems involving renewable energy with high initial investment, such as solar concentrators, requires detailed risk analysis. Such analysis should consider all effects that may have an impact on the thermal and economic performance of the plant. To that end, designers use different software tools that allow a realistic estimation of energy production and which require meteorological data as input. This data can be composed from different sources, such as ground measurements, satellite images and atmospheric models. Moreover, only one year of data is usually used in the energy simulation software in order to reduce data volume and speed-up the simulation (Cebecauer and Suri, 2015). However, due to the interannual variability of the meteorological parameters, it is necessary to use a dataset that represents one typical meteorological year (TMY) to avoid the possible extreme variations contained in a particular year or in a long-term period (Wilcox and Marion, 2008).

In the literature, although there are many approaches available for the construction of TMY files, the commonly approach used to create such dataset is the Sandia method, proposed by (Hall *et al.*, 1978). This datasets usually containing 8760 hourly records of meteorological parameters, that represent climate conditions of a determined region (Sawaqed, Zurigat and Al-Hinai, 2005; Yilmaz and Ekmekci, 2017). Originally, this method considered hourly data (air temperature, dew point, wind speed and global horizontal solar irradiance) measured over a long period of time to generate a single year of data that represents in a stable way the climate conditions of a given location. In other words, TMY represents the occurrence and persistence of a given climate pattern for 12 typical meteorological months (TMM) of a given location.

In order to find the TMM, some procedures, such as the Finkelstein-Schafer (FS) statistic (Finkelstein and Schafer,

doi:10.18086/swc.2019.45.02 Available at http://proceedings.ises.org

1971), are applied over the hourly dataset. For each month of the dataset, the FS is used to find the minor distance between the monthly cumulative distribution function (CDF) of each meteorological parameter and the long-term CDF of the respective parameter, identifying twelve candidate months to compose the TMY. Another important procedure is the use of weighting factors (WF) on the meteorological parameters. According to Hall et al. (1978), the attribution of weights is used to define which meteorological variables will be of greater importance for the selection of TMM. Thus, the WF values should be based on the type application for which the TMY is generated, and so, for solar energy applications, the maximum weight should be assigned to global solar irradiance (Kalogirou, 2003). In addition, each change on the WF values or simply add a meteorological variable can generate a new TMY version. For example, the TMY2 (Marion and Urban, 1995) is a dataset of 239 TMY based on Sandia method, that includes direct normal solar irradiance to the variables list with same weight of the total horizontal solar radiation. The last version is an update from TMY2 and is called TMY3 (Wilcox and Marion, 2008). It's had a more than 1000 weather data files over USA and was generated by the National Renewable Energy Laboratory (NREL) from surface observations, models and satellite data of the National Solar Radiation Data Base (Sengupta *et al.*, 2018).

Other studies have been proposed to create new TMY, which propose new methodologies or modify the Sandia method in order to find the best weather dataset for different energy systems. For example, a TMY known as Design Reference Year (DRY) was proposed by Lund and Eidorff (1981) for Europe, and updated by Lund (1995). The DRY contains additional parameters such as diffuse horizontal irradiance (G_d), illuminance, longwave radiation and weather forecast data. Festa and Ratto (1993) modified the DRY to form a weather file to Ispra, Italy. They replaced the FS statistic with a Kolmogorov-Smirnov statistic, and used the relative frequency distribution to compare single month versus long-term frequency distribution of all the months. The study of Pissimanis et al. (1988) used the Sandia method to generate a TMY weather file to Athens. However, the authors modified the procedure to find five candidate years to form TMY by using the root mean square difference (RMSD) as primary selection criterion to hourly global horizontal irradiation. Later, also for Athens, Argiriou et al. (1999) produced seventeen weather files from different methodologies and compared the results to show that the best performing TMY was generated by a modified version of the Festa Ratto method, with an additional score system applied to the month with the minimum RMSD. Moreover, Cebecauer and Suri (2015) discussed the characteristics of TMY generation algorithms and conclude that simple methods to form TMY may not preserve the behavior of G and G_b when is considered specific solar energy technology. Cebecauer and Suri (2015) also suggest that to improve the financial and performance risks assessment for solar systems or electrical power output which should be consider a worst-case climate scenario. Due this, the authors defined a TMY construction considering a year with less-favorable solar resource (P90) and a year average climate (P50). In other words, P50, P90 or PXX meaning is a 50%, 90% or XX% of the probability of exceedance of G and/or G_b values, respectively. Therefore, P50 and p90 TMY should more appropriate for evaluate of CSP and CPV projects.

Thus, the overarching aim of this work is to present the TMY generation method for the INMET network distributed over Brazil. For this the TMY3 method was applied over 564 INMET stations, and the results were compared with the long-term average of G and G_b.

2. Data and method

2.1 Weather data

The meteorological data were obtained from The Brazilian National Institute of Meteorology (INMET), which is recognized as the National Weather Service and is linked to the World Meteorology Organization. Since 2005, the INMET has installed and operated about 564 Automatic Weather Stations (AWS) along Brazil (Moura, Tadeu and Fortes, 2016), as depicted on

Fig. 1. Although the AWS measure global horizontal irradiance and auxiliary meteorological variables (air temperature, relative humidity, wind direction and wind speed, precipitation and barometric pressure), for this work only the following meteorological parameters were considered: mean, maximum and minimum air temperature, maximum and mean wind speed, mean, maximum and minimum relative humidity, total daily global horizontal and direct normal irradiation. The time-series have hourly temporal resolution and about 13 years (2005-2018) of data. However, each station was installed in different dates, being operative during different periods of time, or being offline for a long time due to operational problems.

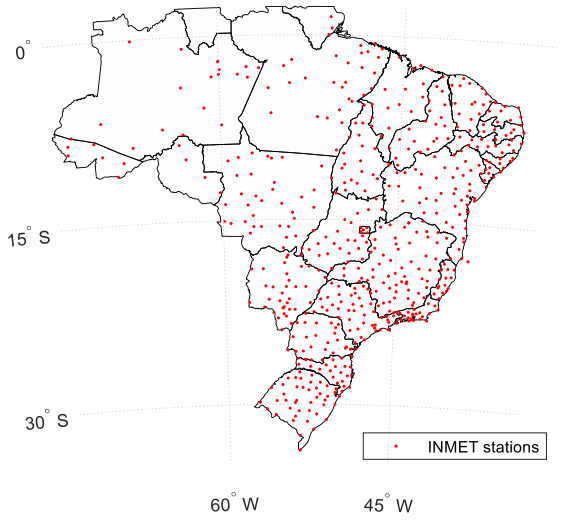


Fig. 1: Locations of the 564 INMET meteorological stations distributed over Brazil

Table 1 shows the specifications of the sensors installed in the AWS network and used to measure the meteorological parameters considered in this study. The AWS were controlled by the QML201 datalogger (Vaisala company) and data were measured at 5 seconds intervals, averaged every 1 minute and temporarily stored. Then hourly average values were calculated and recorded, except for G, for which hourly totals were stored.

Tab. 1: Specifications of sensors installed in the AWS of the INMET network

Parameter	Sensor	Company	Model	Accuracy	Range
G	CMP6	Kipp & Zonen	CMP 6	± 20 W/m²	0 – 2000 W/m²
Relative humidity	HUMICAP180	Vaisala	HMP155	± 1 %	0 – 100 %
Wind speed	Windsonic	Gill	1405-PK-021	± 2 %	0 – 60 m/s
Air temperature	PT-100	Vaisala	QMT103	±0.1 °C	- 50 – 60 °C

2.2 Direct normal irradiance estimate

As neither direct normal irradiance (G_b) or diffuse horizontal irradiance (G_d) were measured in the network, the BRL-Brazil model (Lemos et al., 2017) was used to estimate the diffuse horizontal irradiance (G_d) and thus obtain the G_b. The BRL-Brazil model is an adjustment of the BRL separation model for Brazilian data, developed using measured G, G_b and G_d data from INPE (Brazilian Institute for Space Research) weather stations. The model estimates the diffuse fraction of solar irradiance using a sigmoid function, which takes as inputs the clearness index at the evaluated hour, the solar altitude angle, the apparent solar time, the daily clearness index and a persistence factor, defined as the average between the clearness index in the preceding and the following hours. The model has been shown to deliver better irradiance estimates in Brazil than other hourly separation models widely mentioned in the technical literature.

2.3 Quality control and gap filling

In order to ensure data quality, automated tests were performed on time-series obtained from the INMET network. The quality control (QC) procedure employed for solar radiation is based in the quality checks proposed by Long and Dutton (2010). The procedure created by these authors is recommended by Baseline Surface Radiation Network (BSRN) (Driemel et al., 2018) and was developed to be an efficient procedure for regularly controlling operation of each automatic solar radiation station of the BSRN. Since G_b is estimated by G and G_d, all solar radiation data, in W/m², was discarded if G did not fulfill the following conditions:

$$GHI > -2 \quad \text{(eq. 1)}$$

$$GHI < 1.20\, E_{0n} \cos^{1.2} \theta_z + 50 \quad \text{(eq. 2)}$$

$$GHI < 1.50\, E_{0n} \cos^{1.2} \theta_z + 100 \quad \text{(eq. 3)}$$

where θ_z is the solar zenith angle, in degrees, and E_{0n} is the solar constant (1367 W/m²) adjusted for Earth-Sun distance along the year.

A quality control procedure was also applied on the auxiliary meteorological, where suspect variables were removed by applying the criteria proposed by Fiebrich et al. (2010). Temperature data (T, in °C) must lay inside a range that goes from -30°C to 50°C, and if two consecutive temperature values are equal, the values are discarded. Similar conditions apply to relative humidity (RH, in %) and wind speed (WS, in m/s). Those conditions are summarized by the following equations:

$$T > -30 \text{ and } T < 50\ ;\ T_t \neq T_{t+1} \quad \text{(eq. 4)}$$

$$RH > 3 \text{ and } RH < 103\ ;\ RH_t \neq RH_{t+6} \quad \text{(eq. 5)}$$

$$WS > 0 \text{ and } WS < 40\ ;\ WS_t \neq WS_{t+10} \quad \text{(eq. 6)}$$

After applying the quality control, the time-series of each station contains "gaps" – missing data – that need to be filled before creating the TMY. Therefore, these missing data were identified and divided in three groups, according to the length of the gaps. The first group contemplates gaps length between 1 and 3 hours, which were filled by a linear interpolation, as proposed by Wilcox and Marion (2008). The second group considered gaps between 3 and 24 hours, to preserve the diurnal cycle of the variables and fill larger gaps still contained within 24 hours. To this group, each hourly gap was fulfilled using the mean between the values of the previous and next day for that hour, as suggested by Liston and Elder (2006).

The last group contained gaps greater than 24 hours, which were filled using ERA5 reanalysis data from ECMWF (European Centre for Medium-Range Weather Forecasts). The ERA5 consists in numerical methods combined with historical observations to estimate the state of the atmosphere over the globe with fine spatial grid and high time resolution, 31 km and hourly, respectively (ECMWF, 2017). According to Urraca et al. (2018), ERA5 has shown great potential to estimate variables, such as global horizontal irradiance, and for this reason it was considered for filling of large gaps. The data were extracted from the grid point closest to the INMET station location, given by its latitude and longitude.

2.4 TMY development procedure

The TMY3 algorithm consists in selecting, from different years, 12 individual months, which are concatenated in order to generate a typical meteorological year. For example, in Brasilia station there are 14 years of data. Therefore, all 14 Januarys are examined and one of them is chosen as the typical January. The same process is done for the other months. Since the concatenation of typical months can cause abrupt discontinuities at month interfaces, a smoothing of 6 hours on each side is recommended. In this work, the method used for smoothing is a moving average. Below, the steps of the Sandia method are detailed using Brasilia station as example.

Step 1: Hourly data are used to produce daily parameters for the variables of interest. Short-term Cumulative Distribution Functions (CDFs) are then generated for the individual months, using the generated daily data. For example, in Brasilia there are 14 short-term CDFs generated for each of the 12 months. Additionally, long-terms CDFs are generated using the daily data from all years for the 12 months. Following the previous example, the long-term CDF for January is generated from the 14 Januarys, concatenated together. The CDF generation process is repeated for each variable used in the TMY algorithm. Therefore, 12 long-term CDFs are produced for each variable. Since there are 10 meteorological variables of interest, 120 long-term CDFs and 1680 short-term CDFs are generated in total.

Step 2: Each short-term CDF is compared against the long-term CDFs using the Finkelstein–Schafer (FS) statistic in order to find 5 candidate months for each of month of the year. This comparison is done using the following expression:

$$FS_x = \frac{1}{N}\sum_{i=1}^{N}\left|CDF_m(x_i) - CDF_{m,y}(x_i)\right| \qquad \text{(eq. 7)}$$

where x is the meteorological variable, N is the number of days of the month of interest, CDF_m is the long- term CDF for the month m and $CDF_{m,y}$ is the short-term CDF for the year y and month m. The weighted sums (WS) indicator used to select the five candidates' months is calculated using the equation that follows:

$$WS = \sum_{x=1}^{n} w_x FS_x \qquad \text{(eq. 8)}$$

where n is the number of variables, w_x is the weight assigned to each variable x, which are presented in Tab. 2. The weights are assigned in this work according to the NSRDB TMY2 and TMY3 (Wilcox and Marion, 2008), using the relative humidity instead of the dew point.

Tab. 2: The weighting parameters (w_x) used by Sandia method and in this work.

Sandia Method		This work	
Parameter	Weight	Parameter	Weight
Max dry bulb temperature	1/24	Max air temperature	1/20
Min dry bulb temperature	1/24	Min air temperature	1/20
Mean dry bulb temperature	2/24	Mean air temperature	2/20
Max dew point	1/24	Max relative humidity	1/20
Min dew point	1/24	Min relative humidity	1/20
Mean dew point	2/24	Mean relative humidity	2/20
Max wind speed	2/24	Max wind speed	1/20
Mean wind speed	2/24	Mean wind speed	1/20
Global horizontal irradiation	12/24	Global horizontal irradiation	5/20
Direct normal irradiation	Not used	Direct normal irradiation	5/20

For each of the 12 months, five candidate months are obtained. The candidates are the ones with the lowest weighted sums, representing the closeness to the long-term. The Fig. 2 shows the worst, the best, the chosen and the long-term CDF for total global horizontal irradiation of April. It is important to note that because of the weighs and the next steps of the algorithms, the best CDF (the one with the lowest FS value) is not necessary the one chosen to compose the final TMY.

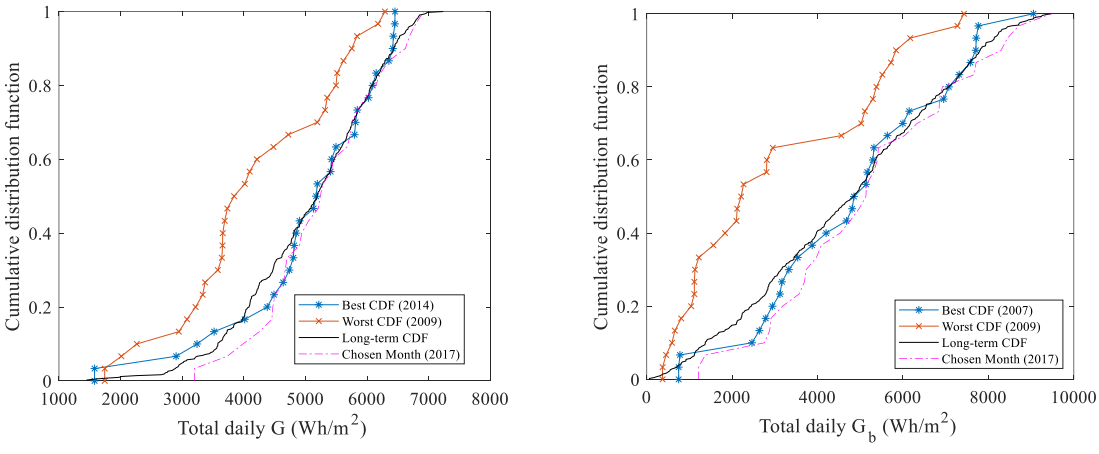


Fig. 2: Comparison of the short- and long-term CDF's for April months, for Brasilia, considering total daily G (a) and G_b (b)

Step 3 The absolute difference between the short-term (ST) mean and long-term (LT) mean are calculated in respect to temperature and total daily G for the five candidate months. The process is repeated considering not the mean, but the median, calculating another two differences. The candidates are then ranked in ascending order of their maximum

absolute difference. Note that four differences are calculated for each candidate, two for temperature and two for total daily G. Tab. 3 summarizes the absolute differences for the candidate years of April.

Tab. 3: Absolute differences and ranking for April

Candidate years	| Mean ST – Mean LT |		| Median ST – Median LT |		Maximum difference	Sorted maximum differences
	Temperature	G	Temperature	G		
10	0.2033	**28.11**	6.25×10^{-8}	9.2555	28.11	28.11
3	0.3275	**120.62**	0.6188	63.950	120.62	120.62
4	0.1854	157.98	0.3156	**401.84**	401.84	238.51
8	0.5196	235.72	0.5708	**291.14**	291.14	238.51
13	0.7625	**238.51**	0.8229	69.022	238.51	401.84

Step 4: The persistence of mean temperature and total daily G are evaluated by calculating the frequency and the length of runs above and below a fixed parameter for each candidate month. For temperature, determine the frequency (number of runs) and runs' length above the 67th percentile and below the 33rd percentile, which represent respectively consecutive warm and cool days. For G, only the frequency and runs' length below the 33rd percentile is evaluated, representing consecutive low-radiation days. Based on the persistence evaluation, a process of elimination occurs for each group of five candidates, in order to exclude months under extreme conditions. Tab. 4 shows the runs obtained for fist candidate year of Abril. The bold numbers in the temperature and total G indicate the values, below which the values under the 33rd percentile are found. In this case, runs above the 67th percentile for mean temperature were not found. The process excludes the months with most runs and the months with zero runs, if any is found. After the exclusion, the first month is selected from remaining months, ordered as the previous step. The selected month is the chosen month for the TMY. Repeating the process to the other set of five candidate, twelve months are selected and concatenated, in order to generate one typical meteorological year.

Tab. 4: Runs obtained for the mean temperature and total daily G for April's first candidate

Day		Mean temperature (°C)	Day		Total daily G (Wh/m²)
30	run length = 2	18,550	3		1577,889
29		18,550	23	run length = 2	2900,917
3		19,783	24		3240,722
24	run length = 2	19,958	8		3518,083
23		20,000	16		4014,028
28		20,327	25		4379,25
25	run length = 2	20,519	12		4489,25
26		20,773	22		**4638,056**
16		20,871	21		4738,806
4		20,942	9		4801,972
27		21,008	6		4816,667
12		**21,229**	27		4850,194

2.5 Statistical analysis

In order to evaluate if the generated TMY represent an annual typical behavior of G and G_b, two statistical analysis were performed. Firstly, the monthly averages calculated over the whole period (multi-years) were compared with monthly values obtained from TMY, which were done using the mean absolute error (MAE), mean bias error (MBE) and root mean square error (RMSE), defined as follows:

$$\text{MBE} = \frac{\sum_{i=1}^{n}(TMY_i - LT_i)}{n} \quad \text{(eq. 9)}$$

$$\text{MAE} = \frac{\sum_{i=1}^{n}|TMY_i - LT_i|}{n} \quad \text{(eq. 10)}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(TMY_i - LT_i)^2}{n}} \quad \text{(eq. 11)}$$

where TMY_i and LT_i are the TMY mean and long-term mean, respectively, of the variable of interest for month i and n is the number of months. For MBE with a positive value indicates overestimation and a negative value represent

an underestimation of the calculated values while MAE shows the absolute difference between TMY_i and LT_i. Thus, a value closest to zero is desirable because indicates a minor error. RMSE provides information about the global error by comparison between TMY_i and LT_i. A large value of RMSE indicates a wide deviation between TMY_i and LT_i while the best performance of the TMY_i regarding absolute deviation occurs for lower RMSE values.

The second statistical analysis uses the Kolmogorov-Smirnov (KS) to evaluate the TMY's representativeness. The KS, that is a non-parametric and distribution free test, was proposed by Massey (1951) to compare the maximum difference between two cumulative (or probability) density distribution. Several studies have evaluated TMY using the KS test. For example, Huld et al. (2018) compared air temperature from three different datasets (measured data, TMY and TMY's reanalysis). The integrated difference between the CDF of two datasets is called KSI (Kolmogorov-Smirnov Integrated) and is a measure of dissimilarity between these two distributions. Thus, KS and KSI were calculated for G and G_b to determinate the similitude between the CDF from the TMY and the values on the long-term CDF, considering the. The KS index is calculated as follows:

$$d_{KS} = \max(|CDF_{TMY}(xi) - CDF_{LT}(xi)|) \quad \text{(eq. 12)}$$

where, d_{KS} is the maximum value of the absolute difference between $CDF_{TMY}(xi)$ and $CDF_{LT}(xi)$ with i = 1…n and n is the number of observations. According to Massey (1951) for confidence level of 99% , the critical value CV is defined as:

$$CV = \frac{1.63}{\sqrt{n_e}} \quad \text{(eq. 13)}$$

the effective number of samples n_e is defined by Nielsen et al. (2017):

$$n_e = \frac{n_{LT}^y}{n_{LT}^y + 1} n_{TMY} \quad \text{(eq. 14)}$$

where n_{LT}^y is the number of years of the long-term and n_{TMY} is the number of samples of TMY. The KS ratio is expressed in percentage as the ratio of d_{KS} to its critical value as follows:

$$KS = \frac{d_{KS}}{CV} \times 100 \quad \text{(eq. 15)}$$

According to Espinar et al. (2009), if d_{KS} is lower than a critical value CV, then TMY and long-term datasets have a similar distribution and statistically agree. In other words, the null hypothesis stating that the two CDF are coming from the same distribution is accepted. On the other hand, the KSI quantifies the difference between the two CDFs over the whole data set, not just considering the maximum absolute difference. This indicator is defined as follows:

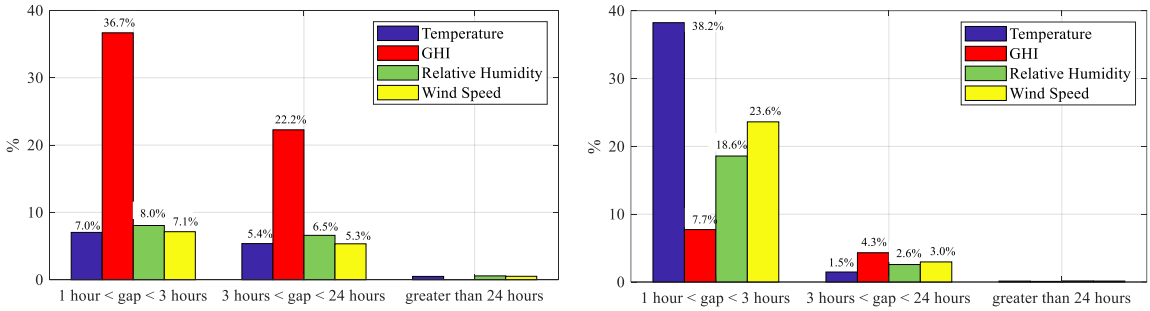
$$KSI = \frac{\int_{x_{min}}^{x_{max}} d_{KS}(x)dx}{CV(x_{max} - x_{min})} \quad \text{(eq. 16)}$$

3. Results

The first exploratory analysis of the 564 AWS time-series from INMET Brazil network showed that there were approximately 9% of missing data over the whole database that includes all meteorological variables. Fig. 3a shows the classification of gap lengths of the missing data found before applying the quality control procedure.

It was observed that approximately 60% of the missing data (9 % of all data) was found on G measurements, while the all auxiliary meteorological variables represented the remaining amount (40 %). The larger missing data on the G measurements can be explained by with higher sensitivity of pyranometer, regarding the presence of excessive dust in the air and that are deposited over the sensors, when compared to other sensors. It also can be noted that there were no gaps greater than 24 hours on G measurements due to replace by zero for nighttime gaps.

The quality control procedure removed an additional 7 % of data that were classified as erroneous or suspect data. Therefore, the dataset contained 16 % of missing data, respective to 9 % before the QC and 7 % of data removed by que QC. Fig. 3b shows that, after the QC was applied, gaps between 1 and 3 hours represented 88% of the missing data, which were filled using a linear interpolation method. Moreover, 11,4% of the gaps were filled with mean between the previous and next day, since its lengths were between 3 and 24 hours. Consequently, more than 99% of the gaps were filled with statistical methods. On the other hand, larger gaps (greater than 1 day), which represented approximately 0,5% of set of gaps, were then filled with ERA5 data. Therefore, all gaps were successively filled.

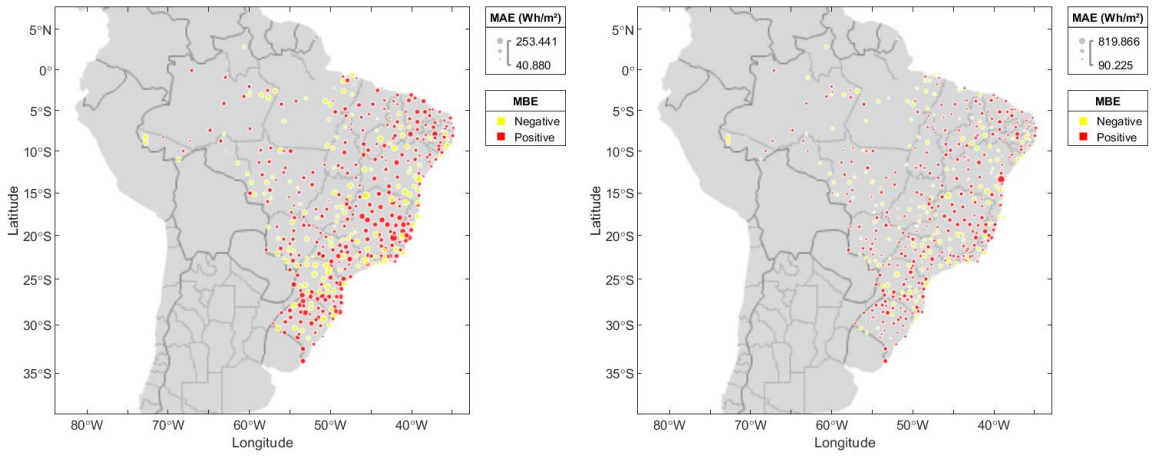


(a) **(b)**

Fig. 3: Length of missing data before (a) and after (b) data quality control for INMET – BRAZIL network

To generate the TMY using Sandia method for 564 weather stations from INMET Brazil network, the procedures described in 2.5 section were applied. The first step consists of selecting only station that have more than 5 year of data, as considered by the Sandia method. This results that 441 of 564 stations fulfill this requirement and could be used to generate the TMY files.

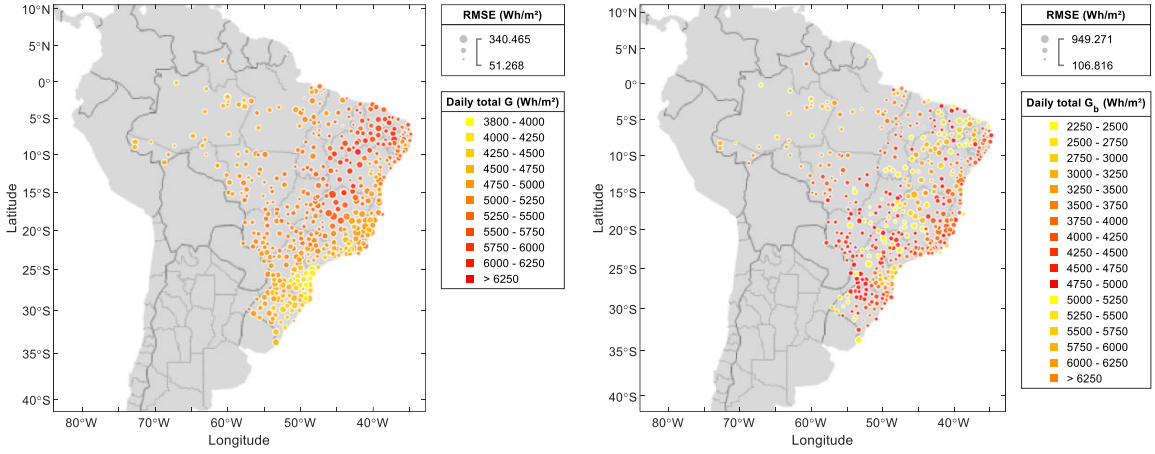
To assess the 441 generate TMY files the MAE, MBE and RMSE between the values of the monthly global horizontal and monthly direct normal irradiance obtained from the TMY and long-term data were calculated and depicted in Fig. 4 and Fig. 5. The individual biases calculated for each station indicate that there are more positive (overestimation) values for both solar variables. Regarding the magnitude of the MAE, 110 Wh/m² on average over all stations was recorded for G and 230 Wh/m² for G_b.



(a) **(b)**

Fig. 4: Mean bias error (MBE) for monthly G (a) and G_b (b) from TMY at each INMET Brazil station. The red (yellow) color indicate positive (negative) bias and the magnitude of the error is represented by size circle

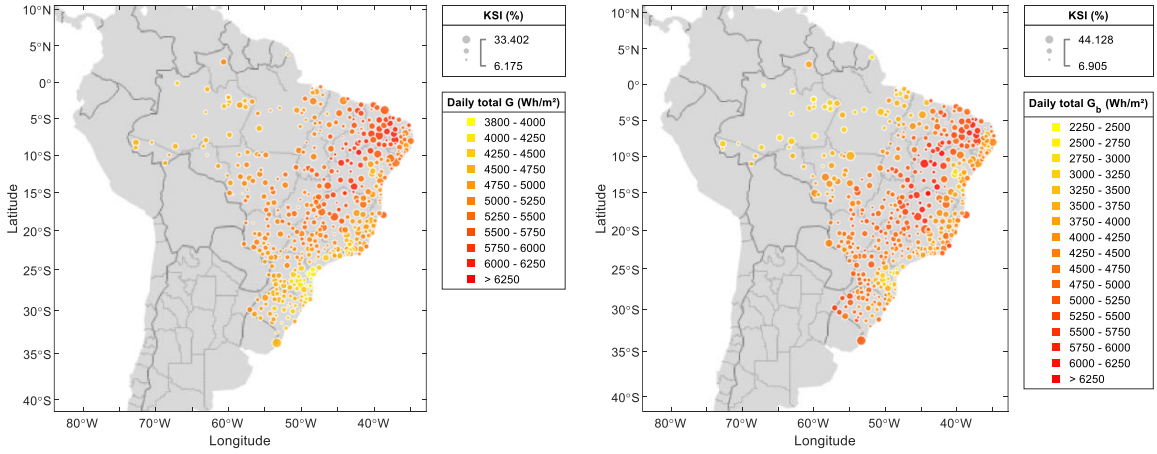
The results for RMSE between TMY and LT indicate that regions with high variability of climate tend to maximum RMSE. For example, the South and Southeastern of Brazil are influence by mesoscale and synoptic atmospheric systems and natural solar cycles (solar annual variability due four seasons well defined along whole year). The stations located in driest area of the Northeastern of Brazil where the total daily G is maximum, for instance, has a low RMSE. In other hand, stations operating over Northeastern of Brazil, but it's has a medium or high RMSE is due to the existence of dry and wet season of the region. According to Luiz et al. (2018) the dry season is due to the influence of Walker circulation cell while the wet season occur due to precipitation caused by Intertropical Convergence Zone present between February and April. The same can be seen for G_b (Fig. 5b), but an additional observation is that for all regions in generally there are smaller values of RMSE over Brazil when compared with G values. The reasons for this can be associated with the fact that the TMY monthly average for G_b is close to the long-term average when compared with TMY monthly average and LT for G.



(a) **(b)**

Fig. 5: Root mean square error (RMSE) for monthly G (a) and G_b (b) from TMY at each INMET Brazil station is represented by size circle. The irradiance is represented by gradient of colors

The results for KSI are displayed in **Erro! Fonte de referência não encontrada.** and provide information on the similitude between the CDF of the TMY and long term data of each INMET station. The smaller KSI values indicate that the TMY generated from the Sandia method is close to the long-term behavior, while larger values of KSI indicates that the TMY CDF's deviates of the long-term data. The magnitude of the KSI values found in this paper is similar to that presented by Espinar et al. (2009). KSI values was below 33% for G and 44% for G_b and the average magnitude over all stations was 14% (G) and 18% (Gb). Furthermore, higher G and G_b were found mainly in the northeastern part of Brazil where the larger KSI values are also concentrated. In this study, the higher distance between the CDF (higher KSI values) can be explained by the high year-on-year variations of the meteorological variables used to form the TMY. In other words, in regions where the weather data varied significantly from year to year were observed large KSI values. On the other hand, the south region of Brazil is where generally the smallest values of G, G_b and KSI can be found.



(a) (b)

Fig. 6: Kolmogorov-Smirnov Integrated (KSI) for monthly G (a) and G_b (b) from TMY at each INMET Brazil station is represented by size circle. The irradiance is represented by gradient of colors.

Finally, an analysis was made for Brasilia's TMY. In this work, the Sandia method was applied to 14 years of data from Brasilia and Fig. 7a presents monthly average values for G and Fig. 7b for G_b. The curves in each figure are the TMY (blue), the 14 years long-term average (green) and the monthly average to each year (black solid line). It is possible to note that both the TMY and the long-term curves reproduce the trend of the monthly variability shown by each year of the data, although for some months the TMY curve overestimates with relatively larger deviations from long-term average. However, the proximity between TMY and long-term values is not mandatory and depends on the cumulative probability distribution function obtained for the variable under analysis.

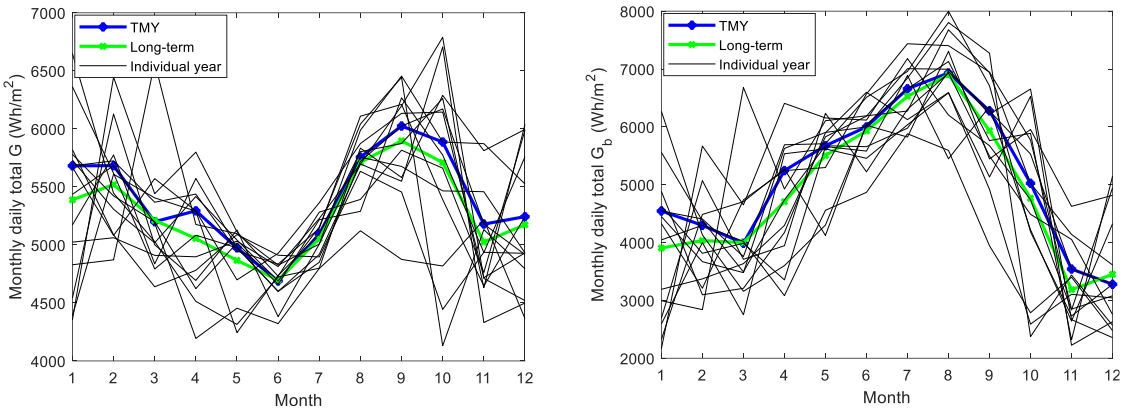


Fig. 7: Average monthly (a) global horizontal irradiance and (b) direct normal irradiance for Brasília city. The blue line represents the TMY monthly average, green line the long-term monthly average and the distribution of monthly averages of the multi-year series are described by black solid lines

Fig. 8 presents a comparison between the TMY and long-term CDFs for city of Brasilia. The selection of the most appropriate month (TMM) to form the TMY shows that there is a good agreement between TMY CDF and LT CDF for both solar variables with a minimum TMY deviation from the long-term CDF, mainly for smaller G and G_b values. This result was expected because the Sandia method-TMY3 was designed to preserve climate statistics especially for G and G_b where the maximum weights are assigned to these variables.

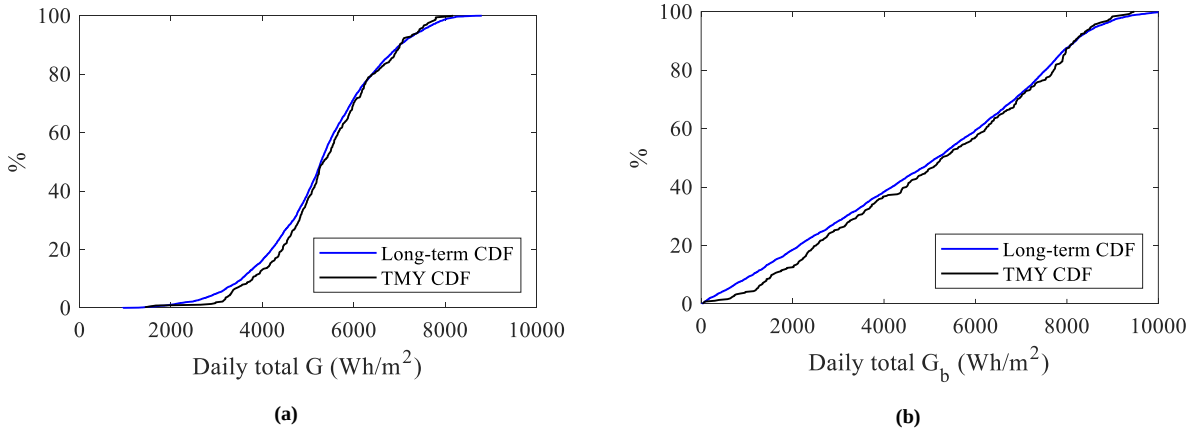


Fig. 8: TMY and long-term CDFs comparison for G (a) and G_b (b) for Brasilia station

4. Conclusions

The increased use of clean energy sources in Brazil indicates that there is still a great potential for solar energy applications in the country. In this context, in order to provide a weather files to use in computer simulation for design and research on buildings and energy systems we have presented the generation of the typical meteorological year using data from the largest weather station network in Brazil, operated by INMET. In developing TMY was considered Sandia Method that is widely adopted to establish typical weather files. Now, Brazil has a database with TMY for 441 cities, generated from measured data (including G_b estimated), to support solar studies on continental scale. The monthly profiles of the weather data were compared to the long-term weather data using statistical analysis such as the MAE, MBE, RMSE and KSI. The results show that the process of constructing the TMY was well performed and the TMY dataset (including Brasilia's TMY) should be considered representative. For example, the analysis of the average KSI index over all stations (14% for G and 18% for G_b) indicated that TMY have a good statistical similarity when compared with original climate datasets. Those interested may contact the authors for an electronic version of Brazil-TMY.

.

5. Acknowledgments

The authors would like the Brazilian National Institute of Meteorology (INMET) for providing the ground measurements. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. This work has been developed within the context of all activities related to the R&D project registered with National Agency of Petroleum, Natural Gas and Biofuels (ANP) under the code 19828-3 and sponsored by Petroleo Brasileiro (PETROBRAS).

6. References

Argiriou, A. *et al.* (1999) 'Comparison of methodologies for tmy generation using 20 years data for Athens, Greece', *Solar Energy*, 66(1), pp. 33–45. doi: 10.1016/S0038-092X(99)00012-2.

Cebecauer, T. and Suri, M. (2015) 'Typical Meteorological Year data : SolarGIS approach', 69, pp. 1958–1969. doi: 10.1016/j.egypro.2015.03.195.

Driemel, A. *et al.* (2018) 'Baseline Surface Radiation Network (BSRN): structure and data description (1992 – 2017)', *Earth Sysytem Science Data*, 10(August), pp. 1491–1501. doi: 10.5194/essd-10-1491-2018.

ECMWF (2017) *Copernicus Climate Change Service: ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate*. Available at: https://confluence.ecmwf.int/display/CKB/ERA5+data+documentation (Accessed: 20 July 2019).

Espinar, B. *et al.* (2009) 'Analysis of different comparison parameters applied to solar radiation data from satellite and German radiometric stations', *Solar Energy*, 83(1), pp. 118–125. doi: 10.1016/j.solener.2008.07.009.

Festa, R. and Ratto, C. F. (1993) 'Proposal of a numerical procedure to select Reference Years', *Solar Energy*, 50(1), pp. 9–17. doi: 10.1016/0038-092X(93)90003-7.

Fiebrich, C. A. *et al.* (2010) 'Quality assurance procedures for mesoscale meteorological data', *Journal of Atmospheric and Oceanic Technology*, 27(10), pp. 1565–1582. doi: 10.1175/2010JTECHA1433.1.

Finkelstein, J. M. and Schafer, E. R. (1971) 'Improved goodness-of-fit tests', *Biometrika*, 58, pp. 641–646. Available at: https://doi.org/10.1093/biomet/58.3.641.

Garreaud, R. D. *et al.* (2009) 'Present-day South American climate'. Elsevier B.V., 281, pp. 180–195. doi: 10.1016/j.palaeo.2007.10.032.

Hall, I. J. *et al.* (1978) *Generation of a Typical Meteorological Year for 26 SOLMET stations, SAND-78-1096C;* Albuquerque, New Mexico, USA.

Hirsch, T. (2017) *SolarPACES Guideline for Bankable STE Yield Assessment, IEA-SolarPACES*. Available at: http://www.solarpaces.org/wp-content/uploads/SolarPACES_Guideline_for_Bankable_STE_Yield_Assessment_-_Version_2017.pdf.

Huld, T. *et al.* (2018) 'Assembling Typical Meteorological Year Data Sets for Building Energy Performance Using Reanalysis and Satellite-Based Data', *Atmosphere*, 9(2), p. 53. doi: 10.3390/atmos9020053.

Kalogirou, S. A. (2003) 'Generation of typical meteorological year (TMY-2) for Nicosia, Cyprus', *Renewable Energy*, 28(15), pp. 2317–2334. doi: 10.1016/S0960-1481(03)00131-9.

Lemos, L. F. L. *et al.* (2017) 'Assessment of solar radiation components in Brazil using the BRL model', *Renewable Energy*,

108, pp. 569–580. doi: 10.1016/j.renene.2017.02.077.

Liston, G. E. and Elder, K. (2006) 'A Meteorological Distribution System for High-Resolution Terrestrial Modeling (MicroMet)', *Journal of Hydrometeorology*, 7(2), pp. 217–234. doi: 10.1175/JHM486.1.

Long, C. N. and Dutton, E. G. (2010) *BSRN Global Network recommended QC tests, V2.0*. Bremerhaven, PANGAEA. Available at: http://hdl.handle.net/10013/epic.38770.

Luiz, E. W. *et al.* (2018) 'Analysis of intra-day solar irradiance variability in different Brazilian climate zones', *Solar Energy*. Elsevier, 167(December 2017), pp. 210–219. doi: 10.1016/j.solener.2018.04.005.

Lund, H. (1995) *The Design Referene Year - Users Manual*. Lyngby, Denmark.

Lund, H. and Eidorff, S. (1981) *Selection methods for production of Test Reference Years: final report*. Lyngby, Denmark.

Marion, W. and Urban, K. (1995) *User's manual for TMY2s: Derived from the 1961-1990 National Solar Radiation Data Base*. Golden, CO (United States). doi: 10.2172/87130.

Massey, F. J. (1951) 'The Kolmogorov-Smirnov Test for Goodness of Fit', *Journal of the American Statistical Association*, 46(253), p. 68. doi: 10.2307/2280095.

Moura, A. D., Tadeu, L. and Fortes, G. (2016) 'The Brazilian National Institute of Meteorology (INMET) and its contributions to agrometeorology', *Agrometeoros*, 24(1), pp. 15–27.

Nielsen, K. P. *et al.* (2017) 'Discussion of currently used practices for: " Creation of Meteorological Data Sets for CSP/STE Performance Simulations "', *SolarPACES repository*, p. 103. Available at: http://solarpaces.org/images/BeyondTMY_Discussion_of_current_methods_v3_0.pdf.

Pissimanis, D. *et al.* (1988) 'The generation of a "typical meteorological year" for the city of Athens', *Solar Energy*, 40(5), pp. 405–411. doi: 10.1016/0038-092X(88)90095-3.

Sawaqed, N. M., Zurigat, Y. H. and Al-Hinai, H. (2005) 'A step-by-step application of Sandia method in developing typical meteorological years for different locations in Oman', *International Journal of Energy Research*, 29(8), pp. 723–737. doi: 10.1002/er.1078.

Sengupta, M. *et al.* (2018) 'The National Solar Radiation Data Base (NSRDB)', *Renewable and Sustainable Energy Reviews*. Elsevier Ltd, 89(January 2018), pp. 51–60. doi: 10.1016/j.rser.2018.03.003.

Urraca, R. *et al.* (2018) 'Evaluation of global horizontal irradiance estimates from ERA5 and COSMO-REA6 reanalyses using ground and satellite-based data', *Solar Energy*. Elsevier, 164(March), pp. 339–354. doi: 10.1016/j.solener.2018.02.059.

Wilcox, S. and Marion, W. (2008) *Users manual for TMY3 data sets, National Renewable Energy Laboratory*. doi: NREL/TP-581-43156.

Yilmaz, S. and Ekmekci, I. (2017) 'The Generation of Typical Meteorological Year and Climatic Database of Turkey for the Energy Analysis of Buildings', *Journal of Environmental Science and Engineering*, 6, pp. 370–376. doi: 10.17265/2162-5298/2017.07.005.