Renewable Energy 200 (2022) 344–359

Contents lists available at ScienceDirect

Renewable Energy

journal homepage: www.elsevier.com/locate/renene

Spatio-temporal dynamic clustering modeling for solar irradiance resource assessment

Patricia Maldonado-Salguero [a], María Carmen Bueso-Sánchez [a], Ángel Molina-García [b,*], Juan Miguel Sánchez-Lozano [c]

[a] Department of Applied Mathematics and Statistics, Universidad Politécnica de Cartagena, 30202 Cartagena, Spain
[b] Department of Automatics, Electrical Engineering and Electronic Technology, Universidad Politécnica de Cartagena, 30202 Cartagena, Spain
[c] University Centre of Defence at the Spanish Air Force Academy, 30720 San Javier, Spain

ARTICLE INFO

Keywords:
Clustering
Functional data analysis
Global horizontal irradiance
Solar resource
Spatio-temporal variability

ABSTRACT

Nowadays, with the development of international policies and agreements to promote the integration of renewable energy sources, mainly solar and wind, modeling the solar resource by including the spatio-temporal variability is crucial to determine future PV power plant locations and estimate potential power generation performances. However, contributions involving long-term periods and different time windows to explore such potential solar resource variability are generally scarce. Under this framework, the present paper proposes a methodology focused on characterizing and clustering the spatio-temporal solar resource variability through the global horizontal irradiance analysis. Hierarchical clustering technique is firstly used to classify the spatial data. Different time windows — from short-term to long-term data — can be subsequently evaluated by using various sources of information. The Spanish territory is selected as case study, considering 22-year period data (1999–2020) and 1,936,917 observations from online satellite database. Spatial variability and geographical clustering differences are discussed and compared depending on the selected time windows, identifying relevant spatial variations for some specific months. Additionally, some years present more variability as well, in line with the sunspot peak of the solar cycles. The proposed approach gives an alternative comprehensive spatio-temporal clustering and characterization of GHI evolution, providing a suitable methodology to help the current European sustainable energy transition.

1. Introduction

The exponential increase in the installation of renewable energy power plants around the world will make it possible to generate a high percentage of electricity demand in the coming years. An energy dependence on renewable sources has the weakness of resource fluctuations, which can vary both spatially and temporally. Among the different renewables, the estimation and stability of photovoltaic (PV) power generation is subject to such variability. It is then necessary to know in detail the potential solar irradiation variations for the design and efficient operation of a storage system or, for example, the upcoming generation of green-hydrogen through solar energy. Actually, alternative solutions based on storing surpluses for future power demand [1], or the accumulation of excess energy in green-hydrogen and subsequently to distribute them require a detailed knowledge of the variability of this renewable potential on both spatial and temporal scale. Therefore, long-term variability analysis of the solar resource is crucial for an optimal development of the corresponding conversion systems [2]. In addition, it is pointed out that the spatio-temporal variability of solar energy is important in various areas related to human life.

With regard to solar energy analysis, the spatio-temporal variability modeling and forecasting has a particular importance [3]. Short-term stochastic variation conducted by cloud motion were analyzed in many studies [4–6]. In a similar way, there are natural events that modify irradiance values for long-term variability: volcanic eruptions that reduce annual averages [7,8], water vapor, weather patterns such as El Niño or La Niña, wildfires or aerosols which are also included as a cause of variations in radiation trends [9,10] in addition to human activity. Long-term spatio-temporal variability was studied in the last decade [2,11], concluding that solar irradiance was influenced by dimming and brightening trends [9,12,13]. In addition, a variability

* Corresponding author.
E-mail addresses: patricia.maldonado@edu.upct.es (P. Maldonado-Salguero), mcarmen.bueso@upct.es (M.C. Bueso-Sánchez), angel.molina@upct.es (Á. Molina-García), juanmi.sanchez@cud.upct.es (J.M. Sánchez-Lozano).

https://doi.org/10.1016/j.renene.2022.09.113
Received 17 May 2022; Received in revised form 23 August 2022; Accepted 26 September 2022
Available online 29 September 2022

is also induced by the solar irradiance due to the sun cycles [2,14–17]. Solar radiation fluctuations depend on the number of sunspots, which vary in a cyclic pattern from minimum to maximum roughly every 11.2 years [14]. This fact is due to the sun's magnetic field flips on average every 11 years. The sun goes back to its original state every two solar cycles, after flipping its poles twice. When solar activity decreases due to the number of sunspots, the rate of cloud cover increases considerably, causing a solar radiation decreasing on the earth's surface [18]. Understanding and forecasting the behavior of solar cycles remains as unknown. It is important, however, to know the potential and availability if the solar resource is to be used efficiently and to choose locations optimally [18].

In this paper, a spatio-temporal variability of solar irradiance is analyzed and a dynamic clustering model of the global horizontal irradiance (GHI) is proposed and evaluated. Indeed, and according to the specific literature previously discussed, most of contributions present a static clustering modeling for GHI values, neglecting potential spatio-temporal variations of GHI data and, consequently, clustering pattern evolution. Regions with similar solar potential are thus unified, excluding any possible GHI temporal variability. To overcome this drawback, three-stage method based on clustering algorithms are applied and evaluated. Each stage of this study is analyzed in different time windows. The first stage performs a hierarchical cluster analysis of the GHI data of the whole studied area. The second stage estimates the number of optimized clusters based on a threshold irradiance values which limit possible differences among the different clusters. Finally, in the third stage, temporal variation of solar irradiation values are analyzed and evaluated. In this case, monthly clusters are compared to annual average clusters to detect variations among the different clusters along the subsequent months. All this work was developed including most of the 23rd solar cycle, the total 24th solar cycle and partial 25th solar cycle. The Spanish territory is then considered as a case study area, where there is an important difference in latitude and longitude if the Canary Islands are included as well. Moreover, it is a territory where the solar resource has relevant values, as well as a diverse territory from an orographic point of view. Pérez-Burgos et al. [19] considered useful for future solar energy applications the representative of an annual solar radiation behavior; being July the month with direct radiation maximum values for all locations — 750 W/m^2 at noon for inner sites and some lower values for coast zones. Indeed, solar irradiation in Spain has been a subject of interest in other works, focused on studying the spatial variation of solar radiation [20], and inter-annual variability [21–23]. The spatio-temporal variability of renewables in the Iberian Peninsula was also analyzed by Gutiérrez et al. [24], estimating the potential power generation from solar irradiation and PV energy yield. Rodríguez-Benítez et al. [25] evaluated the intra-day variability of solar resource, their associated weather patterns and their impact on solar production. Nevertheless, all contributions propose mapping solar radiation solutions with averaged GHI values that can address some uncertainties to provide an accurate solar resource assessment [26–28].

The rest of the paper is structured as follows: Section 2 describes the proposed methodology; Section 3 presents the case study; results and discussion are given in Section 4 and, finally, conclusions are given in Section 5.

2. Methodology

The aim of this paper is to propose and assess a spatio-temporal characterization of solar resource variability in a certain area. To overcome previous contributions and give a significant relevance to the temporal GHI variations, potential relations among specific areas and their GHI values are determined and analyzed through clustering algorithm techniques. The proposed methodology is schematized in Fig. 1 and each stage is following described in detail.

2.1. Data gathering

Both geographical information of latitude and longitude as well as GHI values of the selected area are gathered and structured in a database. The frequency of observations should be at least daily, and a large number of records should be available to analyze properly the spatio-temporal variability. Meteorological data can be collected from on-the-ground weather stations or online satellite data, which can be downloaded in a variety of sources and formats. A previous satellite-based and ground-measured GHI comparison was carried out by the authors and it can be found in [29]. Nevertheless, and in order to organize and manipulate temporal data, it is necessary to consider a set of tidy data principles which are able to be extended to temporal data as was previously described by other researchers [30]. The proposed methodology can be carried out by using such GHI values or the corresponding normalized GHI values; being k_t = GHI/GHI$_{cs}$, where GHI$_{cs}$ is a clear-sky model.

2.2. Data preparation and filtering

In general, a data-set preparation and filtering is required to create a tidy data frame. Raw data commonly contain irrelevant information and/or outliers that can distort the entire analysis. To avoid this situation, tidy data gives data in a nice format, with no inconsistencies or other issues. The principles of tidy data was enumerated by Wickham [31], and widely applied to different data structures [32]. In summary, data processing needs to transform raw data until they become understandable knowledge [33]. In this case, this conversion operation process is carried out as follows:

Missing values. A preliminary filtering process to avoid locations and areas with a high number of missing values is applied. Choosing the right technique to handle the missing values is a choice that depends on the problem domain, the data's domain and the goal of the study [34]. In this case, all locations with more than 1% of missing data in the analyzed period are not considered for subsequent analysis. Considering a higher percentage of missing data would not be a problem for a long-term analysis of the whole period, but when selecting seasonal time windows, it could mean a relevant loss of observations.

Moving averages. Daily averaged differences among the monthly mean values are commonly found in GHI time-series due to weather fluctuations and oscillations [35]. These observations deviated from average values can be outliers and address abnormal conclusions. To avoid this undesired influence on the results, a smoothing–filtering process are applied on the GHI time series. In this case, the Weight Moving Averages (WMA) is selected for smoothing technique purposes, see Eq. (1). It is commonly used with time series data to smooth the noisy data by avoiding short-term fluctuations and highlight long-term trends [36].

$$Y_t = \frac{\sum_{i=1}^{N} W_i \cdot X_{t-i+1}}{\sum_{i=1}^{N} W_i}, \tag{1}$$

where the weights W_i correspond to values 1 or 0, depending on data is or not available at $t-i+1$, respectively. The smoothed time series (Y_t) is obtained from the original time series (X_t), which is mean-weighted (W_i) depending on the window size (N). This window size (N) is then defined as the number of points averaged at each time instant (t). Note that the larger the window size is, the smoother the curve [37]; while smaller window sizes produce noisier smoothed outputs. The missing values of the data-set are also filled by using this technique.

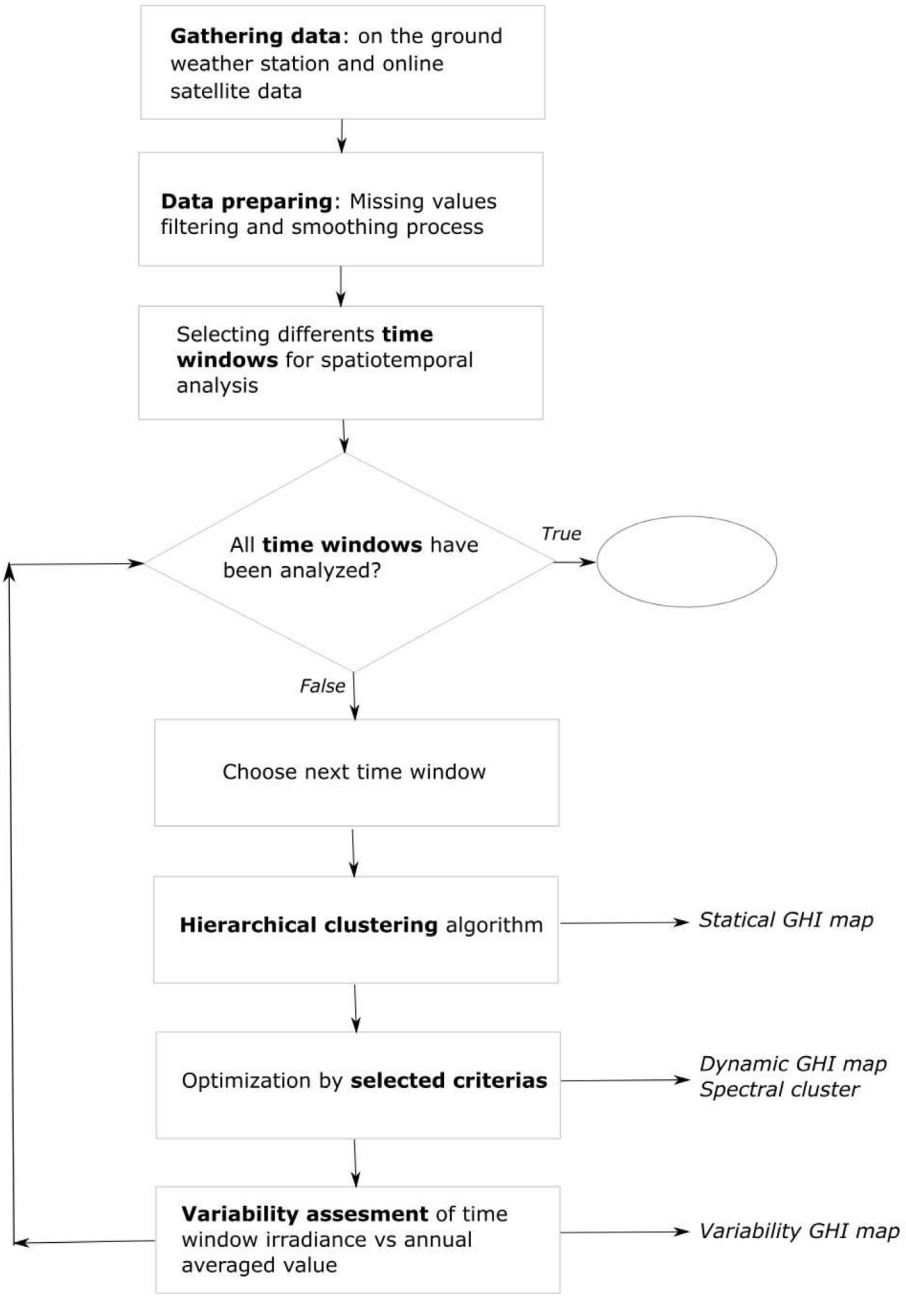

Fig. 1. Proposed methodology. General scheme.

2.3. Dynamic clustering algorithm

Clustering techniques allow us to simplify the spatial analysis and characterize the data through a set of groups and their corresponding most likely patterns [38]. Therefore, a dynamic clustering approach based on the spatio-temporal GHI evolution is proposed to characterize and evaluate the solar irradiance resource.

Firstly, the goal is to characterize the global studied area by dividing into a certain number of spatial groups with similar GHI data. Hierarchical clustering technique [39] is selected to gather all spatial data into such selected number of groups. In hierarchical clustering solutions, clusters are identified by dividing the patterns according to two possible iterative ways: (i) top-down or (ii) bottom-up process [40]. The top-down process is also known as divisive hierarchical clustering. All data are initially included in a general group, which is breaks up into smaller clusters until such clusters fulfill certain conditions or until each object constitutes a cluster on its own. The bottom-up process, named as agglomerative hierarchical clustering [41], starts with atomic single objects which are sequentially merged into larger and larger groups until all single objects are finally included in a single cluster or certain conditions are fulfilled. The primary goal of a clustering analysis is then to minimize the variability on objects belonging to the same group, while maximizing the differences among objects in different clusters. The similarity or difference in the observations is defined by a distance function. The squared Euclidean distance is used in this work, defined between two points, $p = (x_{p1}, \ldots, x_{pm})'$ and $q = (x_{q1}, \ldots, x_{qm})'$, in an m-dimensional space as follows:

$$d_e(p,q) = \sqrt{\sum_{k=1}^{m}(x_{pk} - x_{qk})^2}\,, \tag{2}$$

where $X_1, X_2, \ldots, X_m$ denote the observed variables and x_{pk} is the value for the kth variable for the pth individual. The methods applied to

quantify the similarity between pairs of clusters in this work are the following [42]:

- Complete or maximum: The largest distance between a point in the first cluster and another point in the second cluster is determined.
- Single or minimum: The smallest distance between a point in the first cluster and another point in the second cluster is determined.
- Average: The averaged distances among the points of a cluster and the points of another cluster is determined.
- Ward: In each iteration, the pair of clusters whose merger leads to the smallest increase in the total intra-cluster variance is identified.

Both divisive and agglomerative hierarchical clustering techniques use a binary tree-based data structure called the 'dendrogram'. From this dendrogram, it is possible to deduce — and automatically choose — the suitable number of clusters by splitting the tree at different levels [43]. Different clustering solutions are then provided for the same data-set, avoiding executing the clustering algorithm again. In this paper, hierarchical clustering is aimed to detect and identify automatically groups of areas with similar levels of solar irradiance, while extracting the averaged GHI value of such areas. Initial clusters are then determined and geographic areas with similar GHI values are classified according to the selected time window. These time windows can include an analysis with all data. Therefore, annual, seasonal or monthly time windows depend on the data filtering that is executed prior to running the clustering algorithm.

From these initial clusters and their corresponding averaged GHI patterns of each clusters, the following step is to optimize the number of clusters. This process aims to remove some redundant and/or duplicity clusters. The criteria focus on minimizing averaged GHI values between a cluster and the corresponding near clusters to estimate how different the cluster is with respect to such clusters. The results corresponding to these optimization criteria are a certain number of predefined clusters and their averaged GHI values. An iterative regrouping process is proposed: (i) Clusters are ranked in decreasing order according to their averaged GHI values; (ii) these averaged GHI values are compared among them to detect some similarities; (iii) if the difference between two clusters exceeds the optimization criteria, the clusters will remain in two different groups; (iv) if the difference between two clusters is less than the optimization criteria, both clusters are unified, becoming a single equivalent cluster with a new averaged GHI value. The process is iterated until there are no more pairs of clusters to compare between them.

The spatio-temporal variability can be studied as a function of the time window, whether annual, seasonal, monthly, by solar cycle, etc. In the same way, differences between periods can be analyzed, since the number of optimal groups can be different among them, providing similar or differentiated periods. Finally, the final step aims to determine the variability between the monthly (or seasonal) averaged irradiance versus the annual corresponding averaged value. For this purpose, a third algorithm is proposed to evaluate and represent graphically differences among the annual averaged value and the corresponding averaged values of each month. This algorithm includes the following steps: (i) For each selected location, a single annual averaged value of total irradiance is calculated from all available data; (ii) all data regarding temporal information are filtered by month (or season) and the corresponding monthly (or seasonal) irradiance averaged value are then calculated for each location; (iii) the percentage variation between the monthly (or seasonal) and the annual mean value is calculated for each location; and (iv) a hierarchical clustering algorithm is run to gather the locations with similar variability with respect to the annual averaged values. The process is iterated from (ii) to (iv) for each month (or season) accordingly.

3. Case study

As a case study, the proposed methodology is applied to analyze the long-term spatio-temporal GHI characterization on the Spanish territory. Spain is selected by considering that it is a territory with significant solar resource as well as orographic diversity. In addition, a wide range of difference between latitude and longitude is included when considering the Canary Islands as well. Two databases were considered according to the information from the case study, in order to contrast which of them is more reliable. The first data-set is the collection of GHI data with daily frequency for 16 years (2004–2019) from on-ground meteorological stations of the Spanish territory. Data were downloaded from a total of 135 stations through the Spanish SIAR platform supported by the Spanish Ministry of Agriculture [44]. The second data-set gathers daily averaged GHI values for 22 years (1999–2020) from 241 locations within the Spanish territory. In general, it is considered that a minimum period of 10–15 years of records should be selected to analyze the temporal variability of the solar resource [2]. These data were obtained from the NASA Langley Research Center (LaRC) POWER Project funded through the NASA Earth Science/Applied Science Program [45]. This solar data is based upon satellite observations from which surface isolation values are inferred.

Both databases have different locations and sample times. All data corresponding to both databases were downloaded in `csv` format, with one file per location. In both cases, the variables required for the study are the following: longitude, latitude, date of observation (day, month and year) and the mean GHI value observed for each day. Table 1 summarizes the contents of both databases. In addition, it is required to download the number of sunspots in the period of study. Monthly smoothed total sunspot numbers are provided from the World Data Center SILSO, Royal Observatory of Belgium, Brussels [46]. In total, 761,059 records corresponding to 135 on-ground meteorological stations and 1,936,917 observations corresponding to 241 locations measured by satellite have been managed in this study. The meteorological stations available to collect data do not cover the whole territory of the case study. In addition, not all on-ground stations had 22-year data available to the users. Therefore, our contribution is focused on satellite online data and, thus, on-ground station data are not included due to such stations do not cover homogeneously the case study region and, moreover, they present relevant time series of missed data. With regard to the satellite online data downloaded from the Canary and Balearic Islands, they are included to emphasize the magnitude of the possible temporal variations obtained. The Canary Islands, whose latitude and longitude differ considerably from the rest of the locations studied, would belong to a different cluster than the peninsular points. However, the graphical results are focused on the Iberian peninsula to facilitate the visualization of the graphic scales. The locations evaluated and the locations presented in this paper are showed in Fig. 2. Fig. 3 shows examples of GHI time series for a location in the north of the Spanish peninsula from the satellite data. Fig. 4 depicts GHI data for another location in the south. The blue time series represents GHI, the red curve represents the 7-day weight moving average; and the green time series represents the 30-day weight moving average. GHI time series are smoothed by the Weight Moving Averages (WMA). Regarding the use of the normalized GHI values (k_t = GHI/GHI_{cs}, where GHI_{cs} is a clear-sky model), the number of missing data by considering the total of 1,936,917 observations in 241 locations was 673,232 (34.76%); considering as 'missing data' such values that the corresponding model data cannot be computed or out of model availability ranges. Therefore, we assumed that GHI_{cs} provided a relevant amount of missing data to be included k_t in this analysis and, at the same time, to ensure a remarkable quality from the results. The dynamic clustering analysis was then determined by using the GHI values of the online satellite data provided by the POWER NASA as summarized in Table 1 and Fig. 2. Nevertheless, a long-term analysis was determined by using k_t. With this aim, both GHI and GHIcs variables were smoothed by using WMA 7

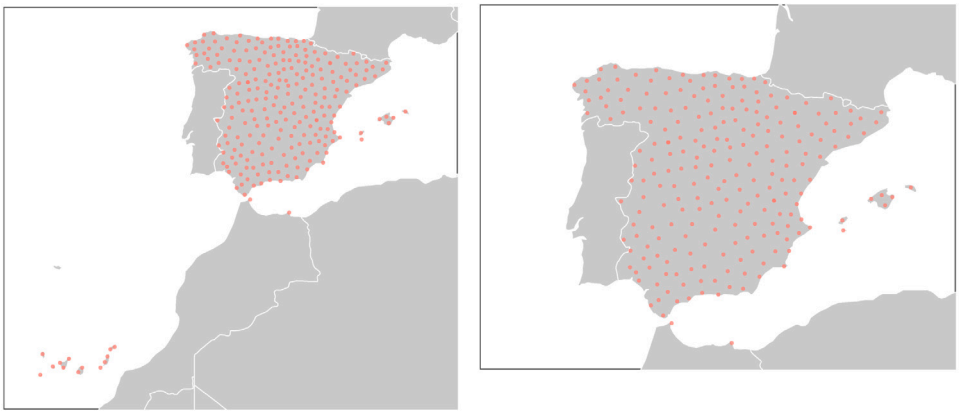

Fig. 2. Locations evaluated (left) vs locations presented (right) with online satellite data.

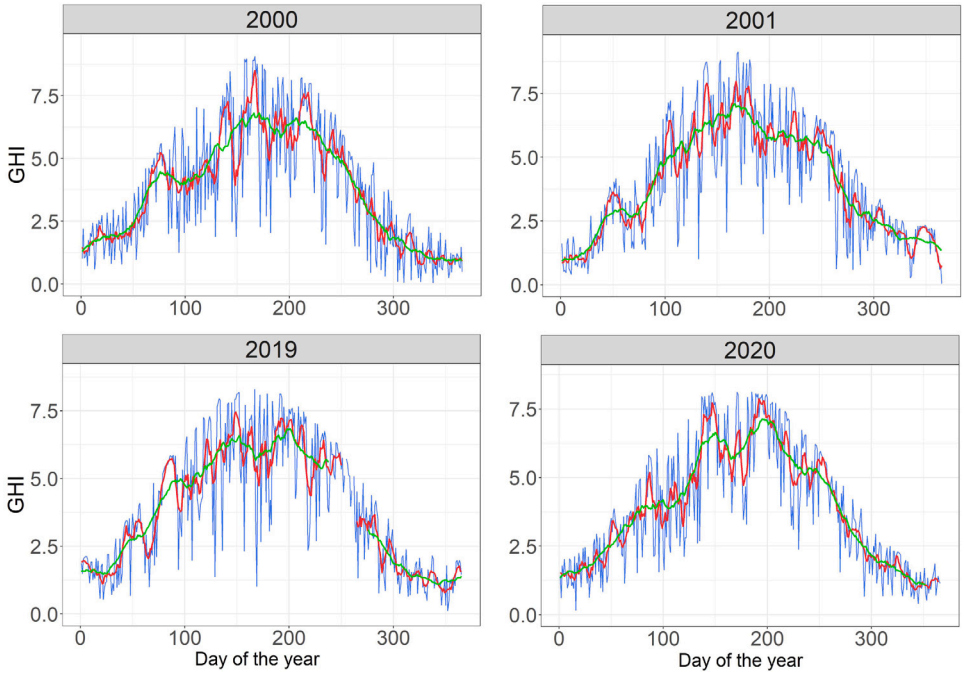

Fig. 3. Examples of GHI (kWh/m^{2}d) time series (blue curve), 7-days WMA (red curve) and 30-days WMA (green curve) in the location (43.23831, −8.15609) from online satellite data.

Table 1
Dataset summary: on-ground vs. online satellite data comparison.

Database	SIAR [44]	POWER NASA [45]
Number of locations	135	241
Frequency	Daily	Daily
Period	2004–2019	1999–2020
Number of records	761,059	1,936,917

days — $N = 7$ in Eq. (1). A hierarchical clustering was performed using Ward's method to normalized GHI (k_t = GHI/GHI$_{cs}$). The clustering algorithm classified a total of 1,263,685 observations, during the period 1999–2020 for 241 locations. The mean k_t of each pool was calculated accordingly, as well as their mean GHI values. From the corresponding clustering results, they did not provide significant differences in comparison to the clusters given by the methodology considering the GHI mean values.

4. Results and discussion

4.1. Database and tidy data process

From the raw data obtained by the on-ground meteorological stations, a relevant weakness of such data is observed after downloading for the long-term spatio-temporal evaluation of the GHI variation. Spatially-wise, the selected locations for the meteorological station database were subject to the geographical positions of these stations, not covering homogeneously the whole studied area. However, the points chosen for the online satellite database were uniformly distributed over the peninsular and insular territory. Temporally-wise, not all meteorological stations have data for 22 years, reducing the downloaded data up to 16 years. Consequently, both databases are related to the duration of one or two solar cycles, respectively.

After a preliminary checking of GHI data, a total of 5.63% of missing data are determined for the 135 analyzed on-ground stations. Subsequently, by applying the criterion that stations with a missing data higher than 1% are not included, only 96 on-ground stations

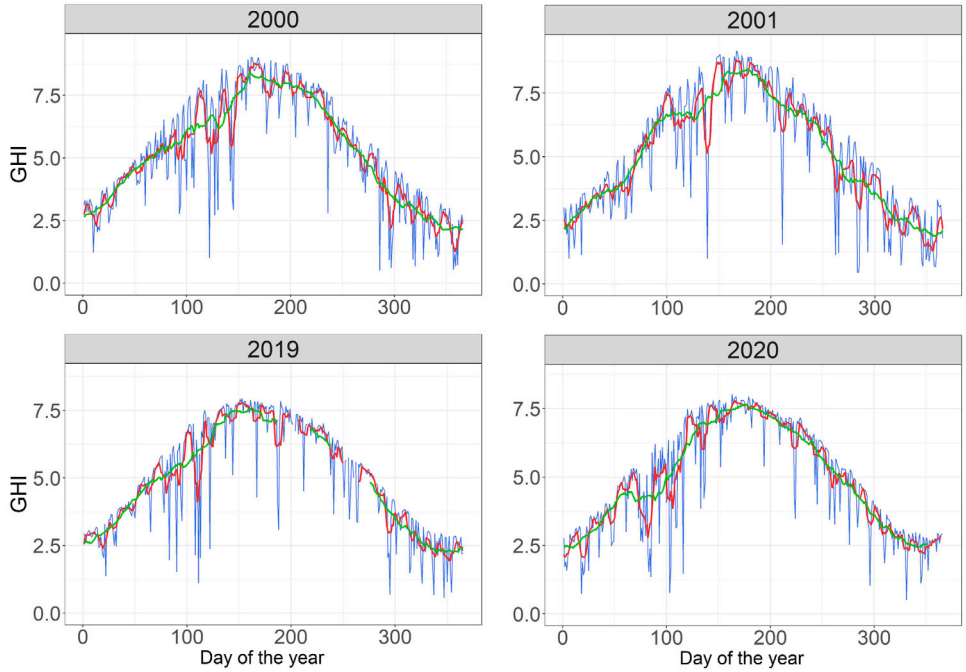

Fig. 4. Examples of GHI (kWh/m^{2}d) time series (blue curve), 7-days WMA (red curve) and 30-days WMA (green curve) in the location (36.93081,−2.5008) from online satellite data.

remain available. For the online satellite database, the missing data represent less than 1% of the total data in all points. Consequently, all the locations for online satellite database initially selected are able to be included in this analysis. Due to the reliability of the satellite data, the GHI time series curves are smoothed by the Weight Moving Averages (WMA) using a 7-day window size. Therefore, the on-ground meteorological stations databases are discarded due to their high value of missing data and low territorial definition. These databases strongly depend on the geographical location of the stations, as was also affirmed by Gutierrez et al. [24].

4.2. Dynamic clustering algorithm for spatio-temporal GHI variability assessment

After the tidy data process, a dynamic clustering algorithm is then applied to estimate the most likely GHI patterns and their temporal evolution and modifications. The aim of this process is thus to analyze and characterize the long-term spatio-temporal variability of GHI by using different time windows: (i) Global evaluation: the input is all the daily observations for the 22 years, and a single 'static' map is provided as output — in line with most of previous contributions as was discussed in Section 1; (ii) Monthly global evaluation: the input to the algorithm are the observations for the 22 years divided into each month. The output is then 12 maps corresponding to each specific month; and (iii) Monthly evaluation: the input to the algorithm is, in this case, the monthly observations corresponding to for each month, being the output 264 maps with dynamic GHI characterization.

Firstly, a hierarchical cluster is selected to obtain the initial classification of the zones with similar GHI values. Initially, all the data downloaded for the period 1999–2020 are used to obtain the global map shown in Fig. 5. Note that using all values for 22 years and choosing 6 hierarchies, the hierarchical cluster algorithm classifies uniformly the selected case study area by means of 5 peninsular groups and 1 group belongs exclusively to the Canary Islands. A long-term difference between latitude and longitude and GHI mean cluster values is concluded. The averaged estimated values for each cluster are summarized in Table 2. This 'static' characterization is avoided by reducing the time window and applying the hierarchical cluster algorithm on each defined time interval. Subsequently, monthly time periods are selected to analyze in detail the seasonal variability. In this case, the following points can be affirmed: (i) The clusters change drastically according to the month and year, losing the uniformity showed in the clusters of Fig. 5; consequently (ii) clusters shared between the peninsula and the Canary Islands in several months, locations in the south that are grouped with those in the north are detected; (iii) some months present significant similarity among most of the points of the peninsula, not being necessary the six groups defined above as they were redundant. Due to this detected variability, it is then proposed to carry out a dynamic hierarchical clustering process in order to analyze the spatio-temporal variability in a more detailed framework. This process is suitable to be performed with a large number of data and thus, to model the spatial GHI variation, as affirmed in [24].

Table 2
Global hierarchical cluster. Averaged GHI cluster values (1999–2020 period).

Cluster	Color	Averaged GHI cluster (kWh/m^{2}d)
1-Cantabric	Dark green	3.706
2-North	Light green	4.060
3-Center	Yellow	4.470
4-South East	Orange	4.648
5-South West	Red	4.920
6-Canary Islands	Out of map	5.599

From the previous results, it can be observed that depending on the selected time window, the clustering groups changed spatially as a function of this specific time window. In addition, the number of clusters do not provide a suitable and reliable characterization. By improving the previous approach, different clustering criteria are tested and their results are subsequently compared. The defined criteria are the following:

- Criteria A: Minimum GHI difference between consecutive clusters 0.50 kWh/m^{2}d.
- Criteria B: Minimum GHI difference between consecutive clusters 0.35 kWh/m^{2}d.
- Criteria C: Minimum GHI difference between consecutive clusters 0.25 kWh/m^{2}d.

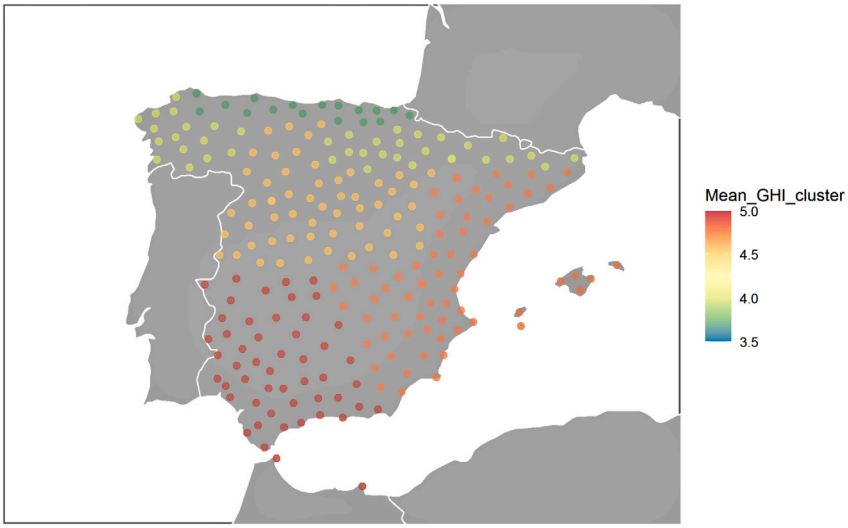

Fig. 5. Global hierarchical cluster.

Table 3
Optimization criteria results. Averaged GHI values for the clusters (kWh/m^{2}d). Period 1999–2020.

Criteria	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Criteria A (0.50)	3.955	4.671	5.599			
Criteria B (0.35)	3.706	4.060	4.671	5.599		
Criteria C (0.25)	3.706	4.060	4.565	4.920	5.599	
Criteria D (0.15)	3.706	4.060	4.470	4.648	4.920	5.599

- Criteria D: Minimum GHI difference between consecutive clusters 0.15 kWh/m^{2}d.

Table 3 summarizes the results obtained by applying such criteria in the entire 1999–2020 period. Note that the higher the optimization criterion, the lower the number of groups determined by the algorithm. For all tested cases, they give $n-1$ clusters in the peninsula area and 1 cluster for the Canary Islands. This last cluster provides the highest averaged GHI value, remaining constant for all tested cases. In addition, cluster 1 and 2 are constant from optimization criteria lower than 0.35 kWh/m^{2}d. These clusters correspond to the Cantabrian and northern Spanish peninsular areas. The southwest peninsular area constitutes a new cluster for optimization criteria lower than 0.25 kWh/m^{2}d. However, in order to discretize the eastern area of the central zone, it is necessary to refine the optimally criteria up to 0.15 kWh/m^{2}d. Further information can be found in the Annex A, see Figs. 11–14.

By selecting a time window corresponding to all data for the period 1999–2020 filtered by each month, i.e., a global monthly analysis, it is observed that the number and locations of the groupings vary depending on the months and on the optimization criteria. Clusters are then observed in horizontal variable bands, whereby the latitude for a long-term monthly analysis influences considerably on the final classification. Nevertheless, slight variations in the number of groups is observed for each optimization criterion. Indeed, between 2 and 3 groups are obtained for the case of Criteria A (0.5). Criteria B (0.35) and C (0.25) give between 3 and 4 groups. Criteria D (0.15) gives between 4 and 6 groups. This fact means that, in a long-term monthly analysis for every criteria, there are no significant variations in the number of clusters for the corresponding months, which could conduct to a uniform behavior of averaged GHI values. Nevertheless, the averaged GHI varies in intensity and the clusters change locations for each month. In all cases, the peninsula area shares a cluster with the Canary Island area. The number of estimated clusters depending on the different criteria is shown in Table 4.

Criteria A (0.5) is selected by considering previous results and due to the similarity between the number of clusters obtained for the long-term period simulations. In fact, this criterion gives the least number of clusters and highlights potential variations for subsequent analysis. By choosing a monthly time window, i.e., selecting only the daily values for each month and for each year, the spatial and temporal variation is studied, considering that each cluster differs from the next identified cluster by at least 0.5 kWh/m^{2}d. The numbers of clusters estimated are summarized in Fig. 6. These results show the significant variability of the GHI intensity evolution among the different years and months of the case study period. The most stable month, in terms of GHI variability is December, varying for all years between 2 and 3 clusters. The most sensitive months regarding GHI intensity variation is March. February, April, May, June and July present between 2 and 5 clusters depending on the corresponding month. Furthermore, it can be concluded that the year in which there was more variability in the intensity of the GHI was 2001, identifying 5 clusters for 3 months, and 4 clusters were characterized for other 3 months. Note that the year 2001 coincided with the sunspot peak of solar cycle 23. The year with the most homogeneous variation in GHI intensity was 2020, coinciding with the beginning of solar cycle 25. Annex B gives additional results, showing in a dynamic map the spatio-temporal variability of GHI.

An additional conclusion can be derived from the previous analysis. A spatial GHI monthly variation of clusters were also detected, in parallel to the analyzed seasonal and annual variability. This GHI behavior was hidden if a 22-year global analysis is carried out. Indeed, it was also neglected when a monthly analysis is determined based on all data simultaneously. In both cases, the clusters appeared in practically horizontal bands, while in this spatial analysis there are points distant in latitude corresponding to the same cluster. The maximum monthly averaged values corresponding to the clusters of maximum GHI values per month are plotted in Fig. 7. This graphical representation leaves out the spatial variable, since the groups of locations including on each cluster change depending on the month. Note that the highest GHI values are achieved for 1999, 2000 and 2001 years. These values decrease with respect to the following years until 2007 and 2008; when they are increased again. From 2014 to 2020, the curves are below those of the other years. This fact also indicates a variability in the magnitude of the GHI globally per year, regardless of whether a certain

Table 4
Number of clusters for the optimization criteria. Monthly results (1999–2020 data).

Month	Criteria A (0.5)	Criteria B (0.35)	Criteria C (0.25)	Criteria D (0.15)
January	3	3	4	6
February	2	3	4	5
March	3	3	4	5
April	2	3	4	4
May	2	3	4	5
June	3	4	4	5
July	3	3	3	4
August	3	3	4	5
September	2	3	4	5
October	2	3	4	5
November	3	3	4	5
December	2	3	3	5

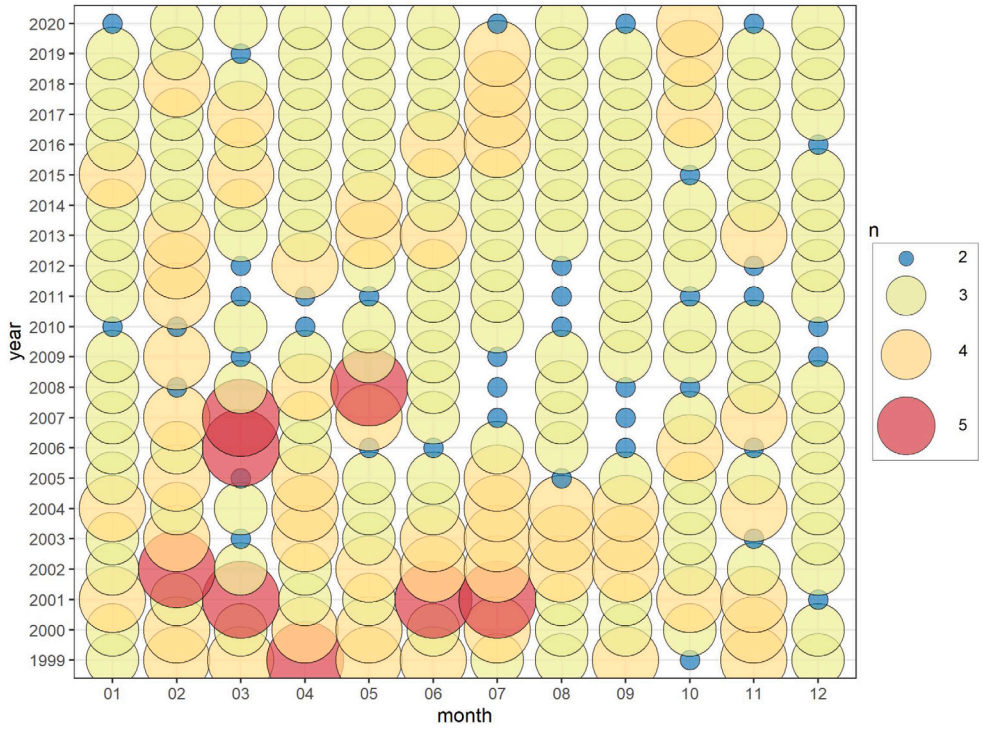

Fig. 6. Spectral cluster. Results.

region is not fixed. Indeed, and regarding the maximum GHI values in comparison to Fig. 8 in which the smoothed monthly sunspot number is plotted [46], note that in the period when the sunspot number increases the maximum GHI values are higher; while the opposite situation occurs when the sunspot number decreases. In addition, Fig. 9 presents these results for each year, in order to make easier the comprehension of this variability.

With the aim of comparing long-term vs seasonal space-temporal variability, variations among the monthly values obtained and their corresponding averaged annual values are analyzed. In this way, Fig. 10 summarizes these comparisons, with a range of [−50%, +50%] and excluding such higher or lower values out of range. As a complementary information, Table 5 gives the numerical data, determined as the differences among the monthly averaged values of the clusters vs their annual averaged value of each year. As can be seen, all monthly values differ from their corresponding averaged annual values in the Iberian peninsula. The months with high deviation from the annual averaged value are June and July, with a difference in the whole peninsula of more than 50% above. December differs by more than 50% below. Regarding January and November, the difference of more than 50% below is exceeded in most of the peninsula, except in the southern areas being slightly below 50%. Only the months of March and September are similar to the averaged annual value, though there are some peninsular areas that differ by more than 10% for these months. In addition, this analysis also shows that only the months of April and September share the same cluster as the peninsula and the Canary Islands. The rest of the months, the Canary Islands have their own different cluster, with the following averaged variations compared to the annual averaged values: January (−34.88%), February (−19.5%), March (2.54%), May (30.03%), June (30.09%), July (25.96%), August (20.69%), October (−12.07%), November (32.36%) and December (−40.31%). A more uniform distribution of monthly GHI vs the monthly averaged values is observed in the Canary Islands, while the differences are more noticeable in the peninsula and vary depending on the month and location.

5. Conclusion

In this paper, an exhaustive spatio-temporal analysis of the variability of the global horizontal irradiation is performed, paying special attention to the sensitivity of the results obtained according to different time windows. A total of 761,059 observations corresponding to 135 on-ground meteorological stations and 1,936,917 observations of satellite data from 241 locations throughout the Spanish territory

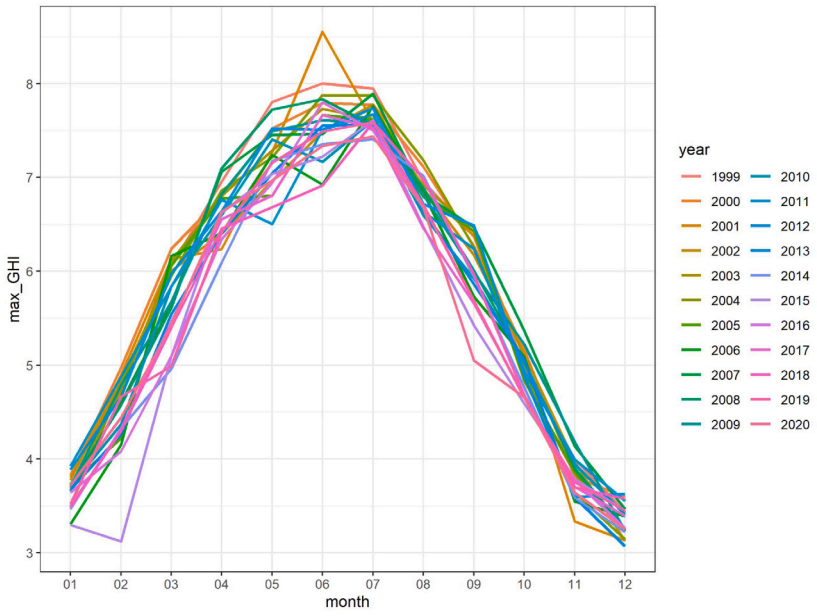

Fig. 7. Maximum GHI group of clusters.

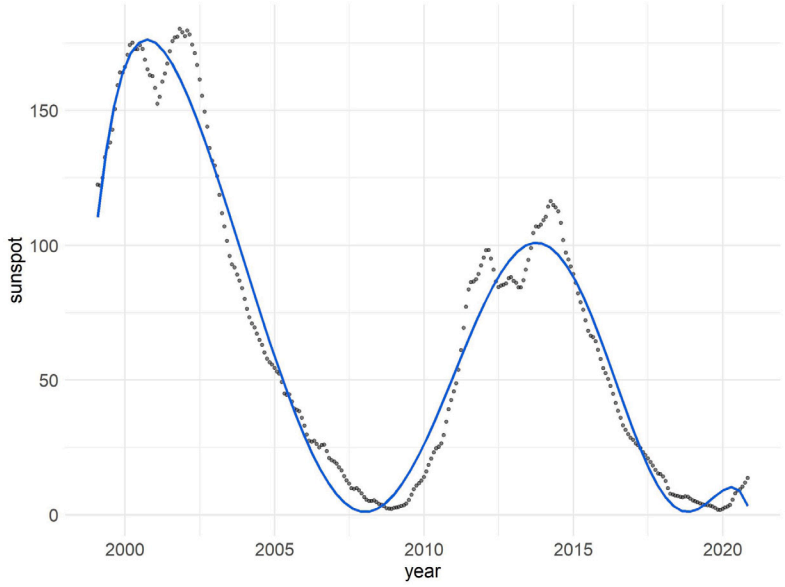

Fig. 8. Number of sunspots [46].

Table 5
Monthly cluster variability. Numeric results.

Month	Min variation (%)	Max variation (%)
January	−44.41	−63.11
February	−26.41	−40.09
March	−2.840	−11.08
April	12.88	26.27
May	33.24	48.82
June	50.45	62.29
July	50.03	71.44
August	31.32	49.13
September	−0.39	17.36
October	−20.23	−30.82
November	−42.18	−56.79
December	−50.48	−66.48

were analyzed. From the proposed spatio-temporal dynamic clustering modeling for solar irradiance resource assessment, it is confirmed that the results obtained highly depend in any case on the selected time window. Therefore, and according to different time windows, the organization of the number of optimal clusters for the entire Spanish territory varies between 2 and 5 per month, considering an optimization criterion of 0.5 kWh/m^{2}d of potential differences between the averaged irradiation values for two consecutive clusters. Similarly, depending on the time window, a clear spatial variation of the grouping is detected.

On the other hand, a cyclical behavior with a positive relationship between the maximum mean values of irradiation of the clusters and the increasing trend of the number of sunspots is observed. It is also noted that the year with the highest number of sunspots in this study (2001) presented a greater heterogeneity in the number of monthly groups. In addition, the year with the lowest number of sunspots (2020) was the most homogeneous year in the number of monthly groups.

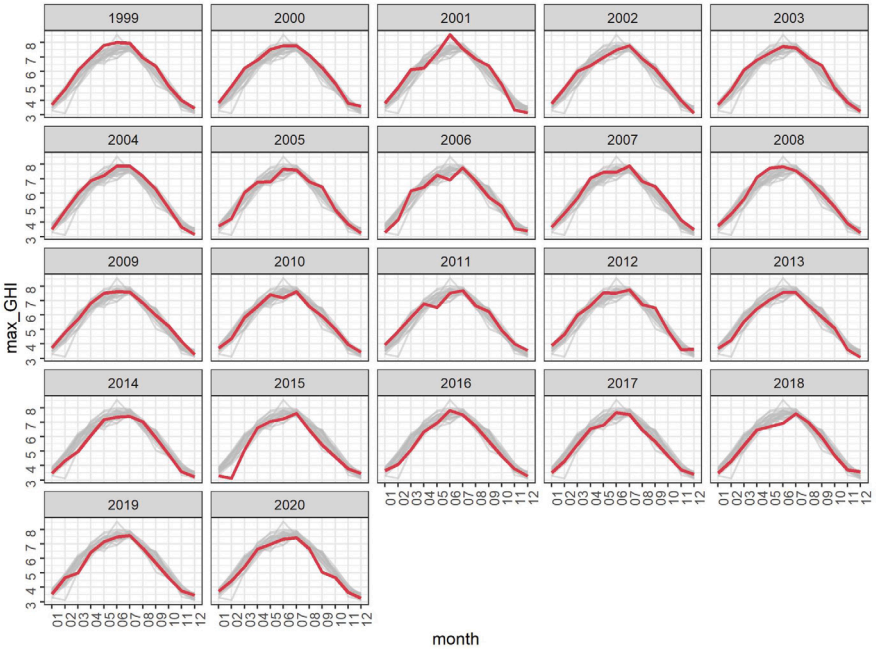

Fig. 9. Maximum GHI group of clusters by years.

Fig. 10. Monthly cluster variability distribution map.

It can be concluded that the annual mean value of global horizontal irradiation for 22 years is similar to the mean irradiation value reached in the months of March and September. The months of December, June and July vary significantly for all the clusters by more than 50% with respect to the mean value. This fact occurs both in values calculated in the long term and in shorter time windows. Therefore, this spatio-temporal analysis proposal is suitable for long-term irradiance characterization and assessment. Moreover, it is necessary to detect clustering variations along the years which can have a significant impact on potential use of solar resource.

CRediT authorship contribution statement

Patricia Maldonado-Salguero: Methodology, Case example, Data curation, Validation. **María Carmen Bueso-Sánchez:** Conceptualization, Investigation. **Ángel Molina-García:** Writing – original draft, Writing – review & editing. **Juan Miguel Sánchez-Lozano:** Methodology, Writing – original draft, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

These data were obtained from the NASA Langley Research Center (LaRC) POWER Project funded through the NASA Earth Science/Applied Science Program, United States. The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

This work was partially funded by the research project PID2020–112754GB–I00, financially supported by the Ministerio de Ciencia e Innovación (Spain).

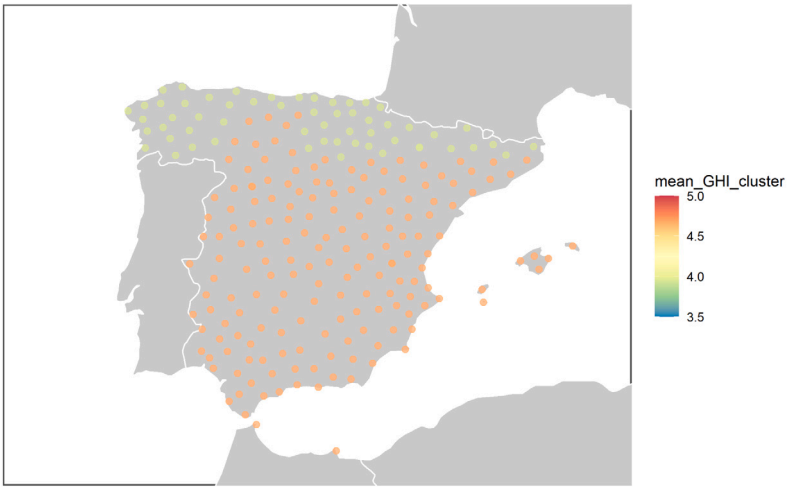

Fig. 11. Mean value of GHI selecting the optimization criteria A (0.5).

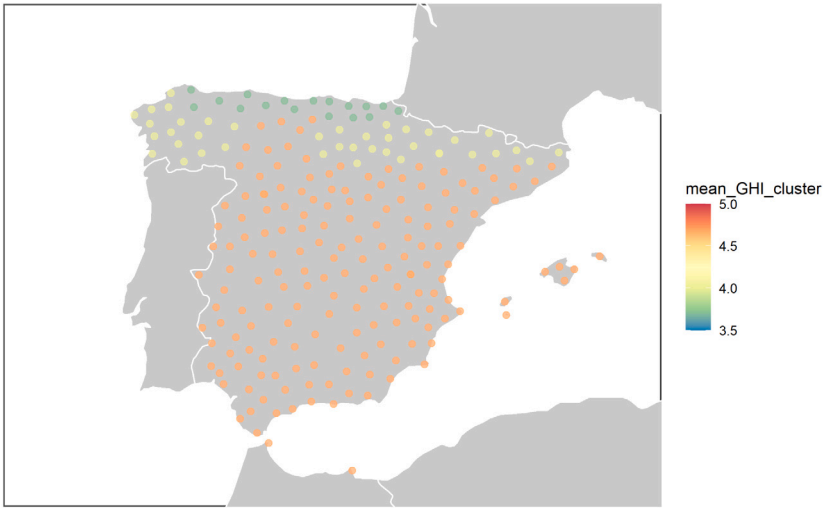

Fig. 12. Mean value of GHI selecting the optimization criteria B (0.35).

Annex A

From the results obtained in the initial hierarchical clustering phase, it is observed that the spatial clustering changes depending on the selected time window. The first hierarchical clustering step does not define an optimal number of clusters. Improving the previous approach, different clustering criteria are proposed and tested. These criteria are defined as the minimum difference between the average GHI value between two consecutive clusters. The defined criteria are as follows:

- Criteria A: 0.50 kWh/m^{2}d.
- Criteria B: 0.35 kWh/m^{2}d.
- Criteria C: 0.25 kWh/m^{2}d.
- Criteria D: 0.15 kWh/m^{2}d.

Fig. 11 shows the results obtained after running the optimization algorithm for criteria A, Fig. 12 for criteria B, Fig. 13 for criteria C and Fig. 14 for criteria D respectively. The time window represented by these figures is the entire data set, for the period 1999–2020. After analyzing these results, it is observed that the higher the optimization criterion, the lower the number of groups. For this reason, criterion A is chosen for the rest of the spatio-temporal analysis of this work. In fact, it represents a smaller number of groups, showing the spatio-temporal variations in a more accentuated way than other criteria which would be useful for detailed studies

Annex B

Figs. 15 to 24 summarize the annual dynamic GHI maps, where monthly GHI characterizations are estimated for the whole case study period 1999–2020. In all cases, the optimization criteria A is used; i.e., the difference between the averaged GHI values of consecutive clusters is greater than 0.5 kWh/m^{2}d. Annual trend changes in the number of sunspots and their peaks are determined according to Fig. 8. These annual maps thus provide complementary information of the results depicted in Fig. 6. The number of clusters varies between 2 and 5 groups, subsequently, a clear spatio-temporal variation of the GHI characterization is then concluded. Note that the year 2001 represented in Fig. 17 is where the clusters have more differences, varying according to the months between 2 and 5 clusters. The year 2020 represented in Fig. 24 is the most homogeneous year, where the number of clusters varies between 2 and 3, except for the month of October. Fig. 8 shows that the year 2001 has the highest number of sunspots in the analyzed period, and the year 2020 has the lowest number of sunspots.

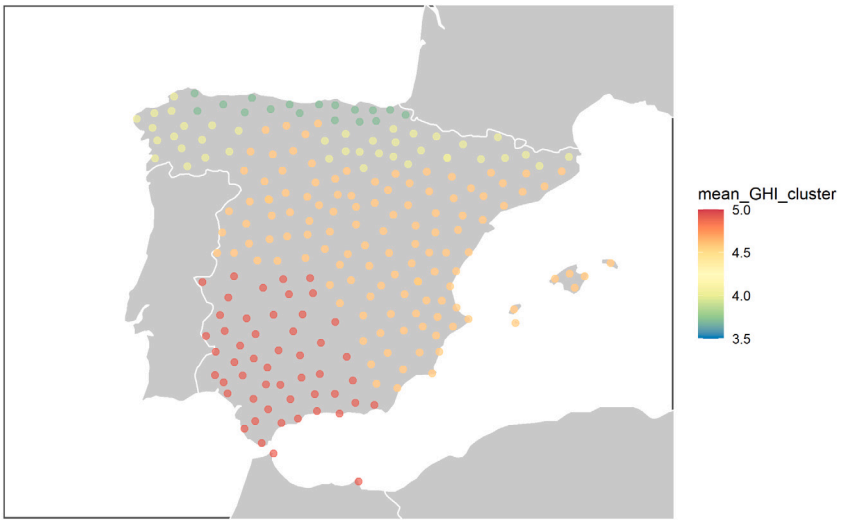

Fig. 13. Mean value of GHI selecting the optimization criteria C (0.25).

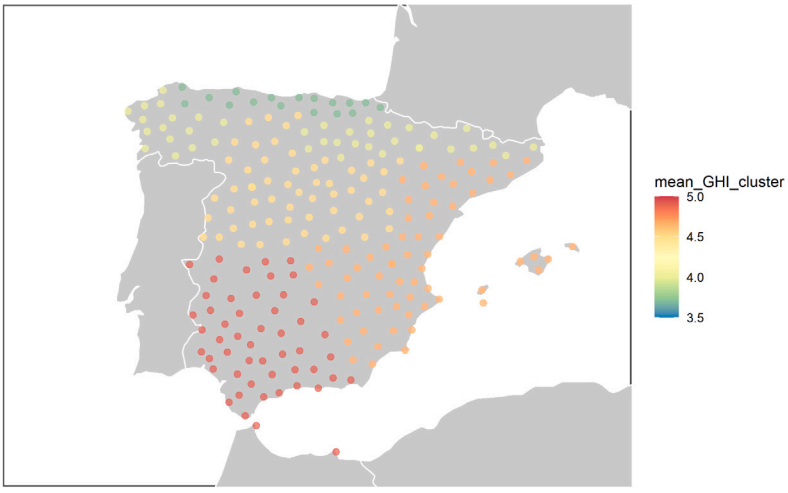

Fig. 14. Mean value of GHI selecting the optimization criteria D (0.15).

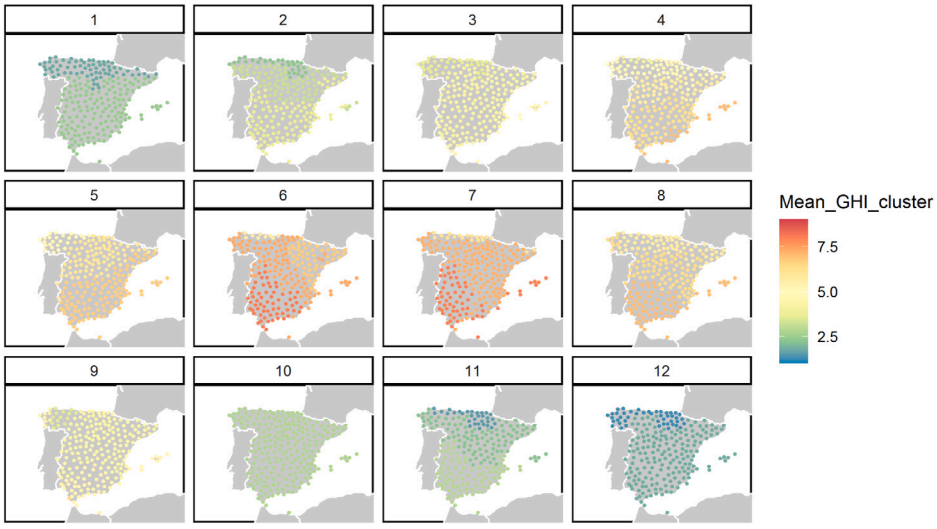

Fig. 15. Dynamic GHI map. Year 1999.

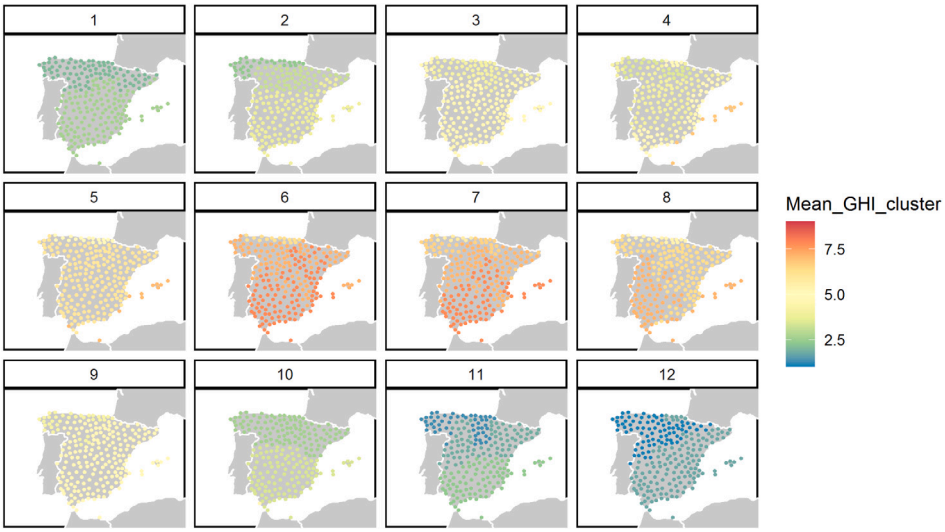

Fig. 16. Dynamic GHI map. Year 2000.

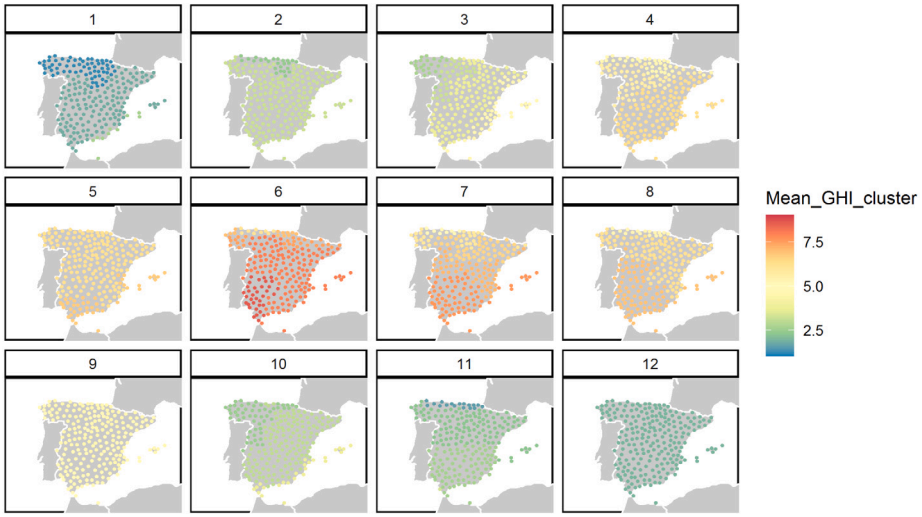

Fig. 17. Dynamic GHI map. Year 2001.

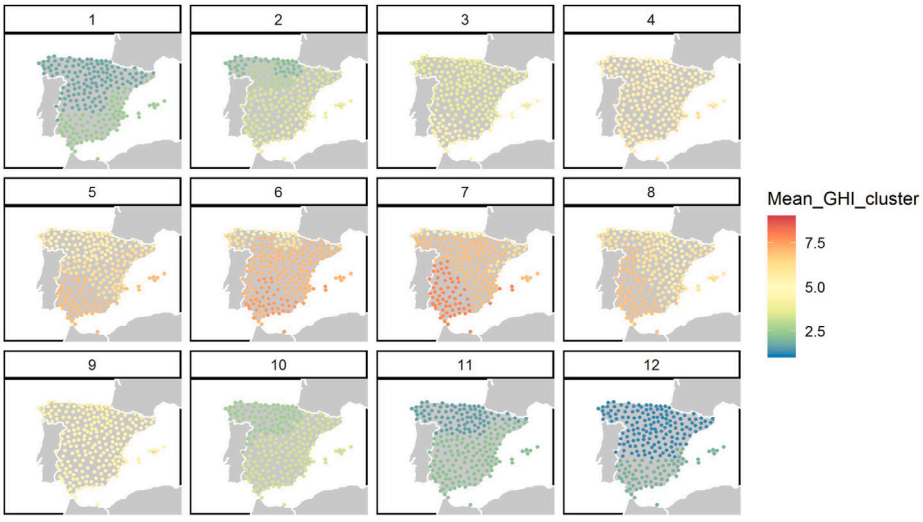

Fig. 18. Dynamic GHI map. Year 2002.

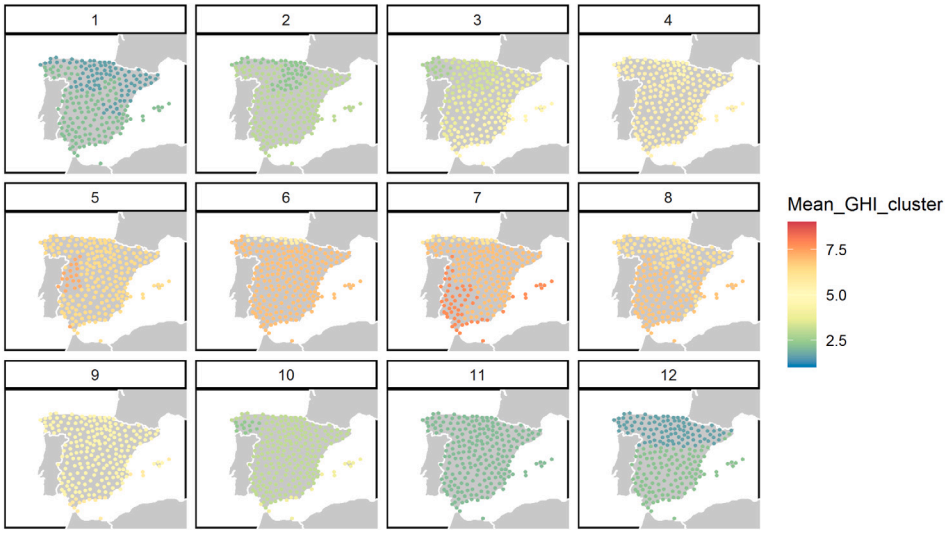

Fig. 19. Dynamic GHI map. Year 2006.

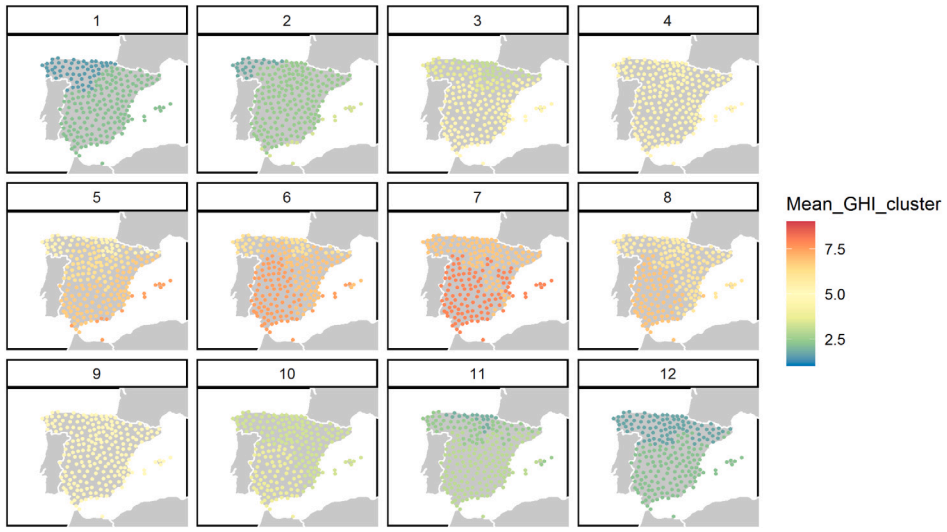

Fig. 20. Dynamic GHI map. Year 2007.

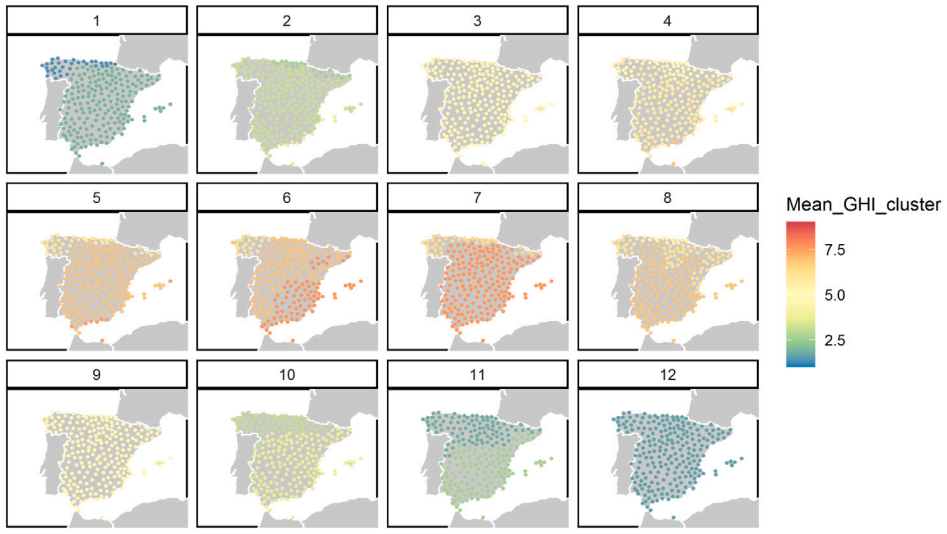

Fig. 21. Dynamic GHI map. Year 2009.

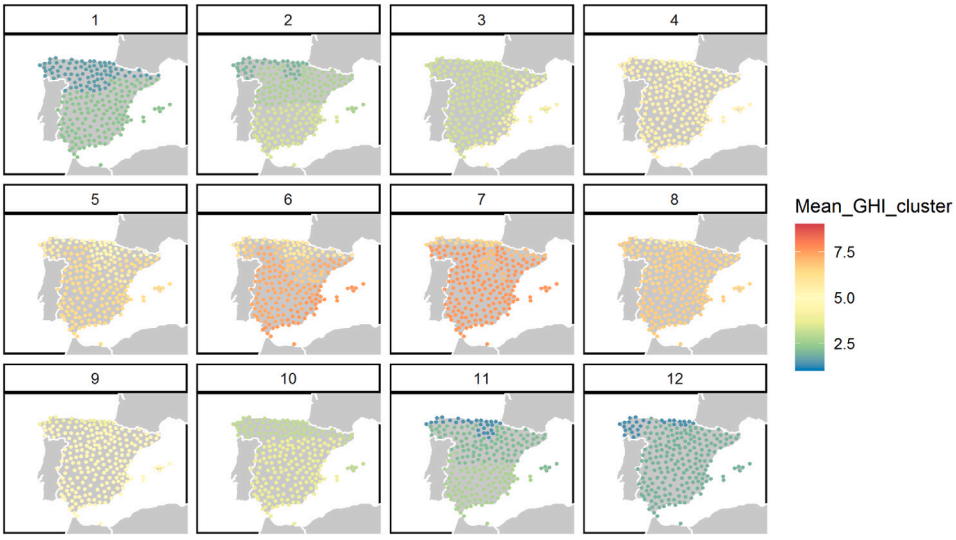

Fig. 22. Dynamic GHI map. Year 2013.

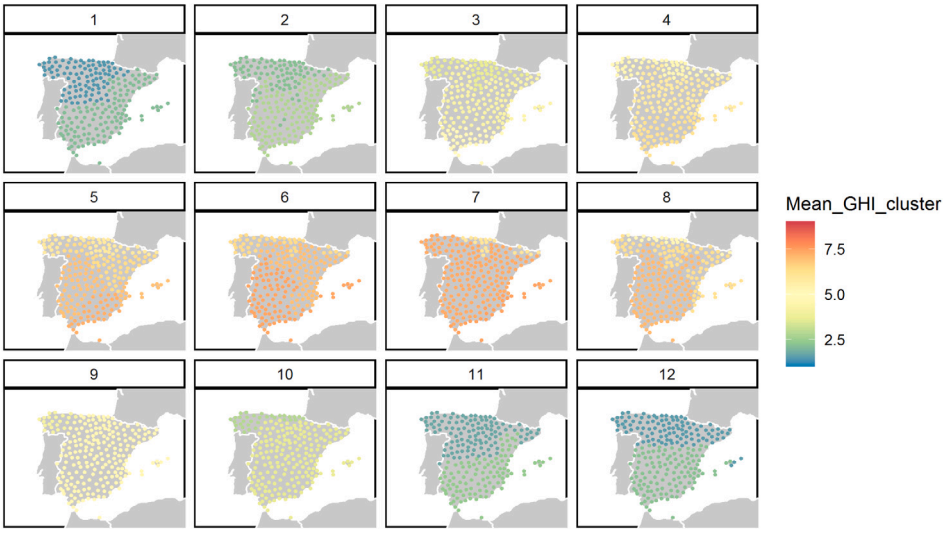

Fig. 23. Dynamic GHI map. Year 2014.

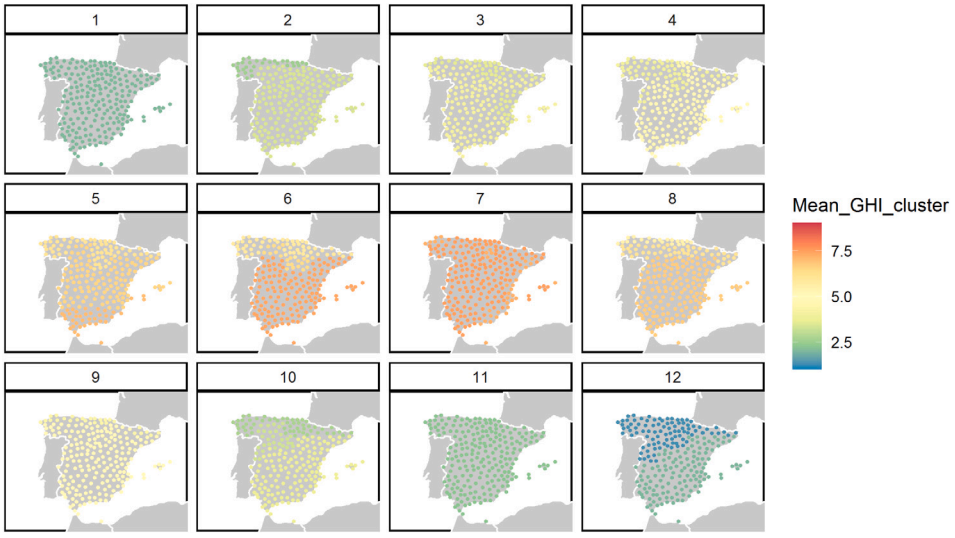

Fig. 24. Dynamic GHI map. Year 2020.

References

[1] S.C. Smith, P. Sen, B. Kroposki, Advancement of energy storage devices and applications in electrical power system, in: 2008 IEEE Power and Energy Society General Meeting-Conversion and Delivery of Electrical Energy in the 21st Century, IEEE, 2008, pp. 1–8.

[2] A. Habte, M. Sengupta, C. Gueymard, A. Golnas, Y. Xie, Long-term spatial and temporal solar resource variability over America using the NSRDB version 3 (1998–2017), Renew. Sustain. Energy Rev. 134 (2020) 110285, http://dx.doi.org/10.1016/j.rser.2020.110285.

[3] R. Perez, M. David, T.E. Hoff, M. Jamaly, S. Kivalov, J. Kleissl, P. Lauret, M. Perez, et al., Spatial and Temporal Variability of Solar Energy, Now Publishers Incorporated, 2016.

[4] A. Woyte, R. Belmans, J. Nijs, Fluctuations in instantaneous clearness index: Analysis and statistics, Sol. Energy 81 (2) (2007) 195–206, http://dx.doi.org/10.1016/j.solener.2006.03.001.

[5] T. Tomson, G. Tamm, Short-term variability of solar radiation, Sol. Energy 80 (5) (2006) 600–606, http://dx.doi.org/10.1016/j.solener.2005.03.009.

[6] R. Perez, S. Kivalov, J. Schlemmer, K. Hemker, T.E. Hoff, Short-term irradiance variability: Preliminary estimation of station pair correlation as a function of distance, Sol. Energy 86 (8) (2012) 2170–2176, http://dx.doi.org/10.1016/j.solener.2012.02.027, Progress in Solar Energy 3.

[7] A. Robock, Volcanic eruptions and climate, Rev. Geophys. 38 (2) (2000) 191–219.

[8] S. Lohmann, C. Schillings, B. Mayer, R. Meyer, Long-term variability of solar direct and global radiation derived from ISCCP data and comparison with reanalysis data, Sol. Energy 80 (11) (2006) 1390–1401, http://dx.doi.org/10.1016/j.solener.2006.03.004, European Solar Conference (EuroSun 2004).

[9] M. Wild, Enlightening global dimming and brightening, Bull. Am. Meteorol. Soc. 93 (1) (2012) 27–37.

[10] P. Juruš, K. Eben, J. Resler, P. Krč, I. Kasanický, E. Pelikán, M. Brabec, J. Hošek, Estimating climatological variability of solar energy production, Sol. Energy 98 (2013) 255–264, http://dx.doi.org/10.1016/j.solener.2013.10.007.

[11] C.A. Gueymard, S.M. Wilcox, Assessment of spatial and temporal variability in the US solar resource from radiometric measurements and predictions from models using ground-based or satellite data, Sol. Energy 85 (5) (2011) 1068–1084, http://dx.doi.org/10.1016/j.solener.2011.02.030.

[12] M. Wild, Global dimming and brightening: A review, J. Geophys. Res.: Atmos. 114 (D10) (2009).

[13] B. Müller, M. Wild, A. Driesse, K. Behrens, Rethinking solar resource assessments in the context of global dimming and brightening, Sol. Energy 99 (2014) 272–282, http://dx.doi.org/10.1016/j.solener.2013.11.013.

[14] S. Solanki, N. Krivova, J. Haigh, Solar irradiance variability and climate, Astron. Nachr. - ASTRON NACHR 323 (2013) http://dx.doi.org/10.1146/annurev-astro-082812-141007.

[15] A. Hempelmann, W. Weber, Correlation between the sunspot number, the total solar irradiance, and the terrestrial insolation, Sol. Phys. 277 (2) (2012) 417–430.

[16] R. Lee III, M. Gibson, N. Shivakumar, R. Wilson, H. Kyle, A. Mecherikunnel, Solar irradiance measurements: minimum through maximum solar activity, Metrologia 28 (3) (1991) 265.

[17] R.B. Lee III, M.A. Gibson, R.S. Wilson, S. Thomas, Long-term total solar irradiance variability during sunspot cycle 22, J. Geophys. Res. Space Phys. 100 (A2) (1995) 1667–1675.

[18] Y. Utomo, Correlation analysis of solar constant, solar activity and cosmic ray, in: Journal of Physics: Conference Series, Vol. 817, IOP Publishing, 2017, 012045.

[19] A. Pérez-Burgos, J. Bilbao, A. De Miguel, R. Román, Analysis of solar direct irradiance in Spain, Energy Procedia 57 (2014) http://dx.doi.org/10.1016/j.egypro.2014.10.070.

[20] CIEMAT, ADRASE. http://www.adrase.com.

[21] J. Sancho, J. Riesco, C. Jiménez, M. Sanchez de Cos, J. Montero, M. López, Atlas de radiación solar en españa utilizando datos del SAF de clima de EUMETSAT, Minist. Agric. 162 (2012).

[22] I.F. Tullot, Atlas de la RadiaciÓN Solar En EspaÑa, Ministerio de Transportes, Turismo y Comunicaciones. Instituto Nacional de . . . , 1984.

[23] N. Vera Mella, Atlas Climático de Irradiación Solar a Partir de Imágenes Del SatéLite NOAA. Aplicación a la Península Ibérica, Universitat Politècnica de Catalunya, 2005.

[24] C. Gutiérrez, M.Á. Gaertner, O. Perpiñán, C. Gallardo, E. Sánchez, A multi-step scheme for spatial analysis of solar and photovoltaic production variability and complementarity, Sol. Energy 158 (2017) 100–116, http://dx.doi.org/10.1016/j.solener.2017.09.037.

[25] F.J. Rodríguez-Benítez, C. Arbizu-Barrena, F.J. Santos-Alamillos, J. Tovar-Pescador, D. Pozo-Vázquez, Analysis of the intra-day solar resource variability in the iberian peninsula, Sol. Energy 171 (2018) 374–387, http://dx.doi.org/10.1016/j.solener.2018.06.060.

[26] J. Polo, Solar global horizontal and direct normal irradiation maps in Spain derived from geostationary satellites, J. Atmos. Sol.-Terr. Phys. 130–131 (2015) 81–88, http://dx.doi.org/10.1016/j.jastp.2015.05.015.

[27] S. Moreno-Tejera, M. Silva-Pérez, I. Lillo-Bravo, L. Ramírez-Santigosa, Solar resource assessment in seville, Spain. Statistical characterisation of solar radiation at different time resolutions, Sol. Energy 132 (2016) 430–441, http://dx.doi.org/10.1016/j.solener.2016.03.032.

[28] R. Urraca, E. Martinez-de Pison, A. Sanz-Garcia, J. Antonanzas, F. Antonanzas-Torres, Estimation methods for global solar radiation: Case study evaluation of five different approaches in central Spain, Renew. Sustain. Energy Rev. 77 (2017) 1098–1113, http://dx.doi.org/10.1016/j.rser.2016.11.222.

[29] M.C. Bueso, J.M. Paredes-Parra, A. Mateo-Aroca, A. Molina-García, A characterization of metrics for comparing satellite-based and ground-measured global horizontal irradiance data: A principal component analysis application, Sustainability 12 (6) (2020) 2454, http://dx.doi.org/10.3390/su12062454.

[30] E. Wang, D. Cook, R.J. Hyndman, A new tidy data structure to support exploration and modeling of temporal data, J. Comput. Graph. Statist. 29 (3) (2020) 466–478, http://dx.doi.org/10.1080/10618600.2019.1695624.

[31] H. Wickham, Tidy data, J. Stat. Softw. 59 (10) (2014) http://dx.doi.org/10.18637/jss.v059.i10.

[32] S. Kampakis, How to keep data tidy, in: The Decision Maker's Handbook to Data Science, A Press, Berkeley, CA, 2020, pp. 45–49, http://dx.doi.org/10.1007/978-1-4842-5494-3{_}4.

[33] N.J. Tierney, D.H. Cook, Expanding tidy data principles to facilitate missing data exploration, visualization and assessment of imputations, 2018.

[34] R. Somasundaram, R. Nedunchezhian, Evaluation of three simple imputation methods for enhancing preprocessing of data with missing values, Int. J. Comput. Appl. 21 (10) (2011) 14–19.

[35] A.A. Prasad, M. Kay, Assessment of simulated solar irradiance on days of high intermittency using WRF-solar, Energies 13 (2) (2020) 385, http://dx.doi.org/10.3390/en13020385.

[36] Y. Zhao, X. He, D. Zhou, M.G. Pecht, Detection and isolation of wheelset intermittent over-creeps for electric multiple units based on a weighted moving average technique, IEEE Trans. Intell. Transp. Syst. (2020) 1–14, http://dx.doi.org/10.1109/TITS.2020.3036102.

[37] B. Ben Atitallah, J.R. Bautista-Quijano, H. Ayari, A.Y. Kallel, D. Bouchaala, N. Derbel, O. Kanoun, Comparative study of digital filters for a smart glove functionalized with nanocomposite strain sensor, in: 2021 18th International Multi-Conference on Systems, Signals & Devices (SSD), IEEE, 2021, pp. 1366–1371, http://dx.doi.org/10.1109/SSD52085.2021.9429298.

[38] A.C. Rencher, A review of "methods of multivariate analysis, second edition", IIE Trans. 37 (11) (2005) 1083–1085, http://dx.doi.org/10.1080/07408170500232784.

[39] F. Murtagh, P. Contreras, Algorithms for hierarchical clustering: an overview, WIREs Data Min. Knowl. Discov. 7 (6) (2017) http://dx.doi.org/10.1002/widm.1219.

[40] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O.P. Patel, A. Tiwari, M.J. Er, W. Ding, C.-T. Lin, A review of clustering techniques and developments, Neurocomputing 267 (2017) 664–681, http://dx.doi.org/10.1016/j.neucom.2017.06.053.

[41] A. Bouguettaya, Q. Yu, X. Liu, X. Zhou, A. Song, Efficient agglomerative hierarchical clustering, Expert Syst. Appl. 42 (5) (2015) 2785–2797, http://dx.doi.org/10.1016/j.eswa.2014.09.054.

[42] R. Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2021, URL https://www.R-project.org/.

[43] C.K. Reddy, B. Vinzamuri, A survey of partitional and hierarchical clustering algorithms, in: Data Clustering, Chapman and Hall/CRC, 2018, pp. 87–110, http://dx.doi.org/10.1201/9781315373515-4.

[44] Ministerio de Agricultura, Pesca y Alimentacion del Gobierno de España, Sistema de informacion agroclimatica para el regadio (red SIAR), 2020, online resource, accessed December 2020. URL https://eportal.mapa.gob.es//websiar/Inicio.aspx.

[45] National Aeronautics and Space Administration (NASA) Langley Research Center (LaRC), POWER data access viewer, single point data access, 2021, online resource, accessed March 2021. URL https://power.larc.nasa.gov.

[46] S.W.D. Center, The international sunspot number, 2021, International Sunspot Number Monthly Bulletin and online catalogue. URL http://www.sidc.be/silso/.