energies

MDPI

Article

Photovoltaic Power Generation Forecasting for Regional Assessment Using Machine Learning

Monica Borunda [1,*], Adrián Ramírez [2], Raul Garduno [3], Gerardo Ruíz [2], Sergio Hernandez [2] and O. A. Jaramillo [4]

[1] CONACYT-Tecnológico Nacional de México -Centro Nacional de Investigación y Desarrollo Tecnológico, Cuernavaca 62490, Mexico
[2] Facultad de Ciencias, Universidad Nacional Autónoma de México, Ciudad de Mexico 04510, Mexico
[3] Instituto Nacional de Electricidad y Energias Limpias, Cuernavaca 62490, Mexico
[4] Instituto de Energías Renovables, Universidad Nacional Autónoma de México, Temixco 62580, Mexico
* Correspondence: monica.bp@cenidet.tecnm.mx; Tel.: +52-55-4939-1608

Abstract: Solar energy currently plays a significant role in supplying clean and renewable electric energy worldwide. Harnessing solar energy through PV plants requires problems such as site selection to be solved, for which long-term solar resource assessment and photovoltaic energy forecasting are fundamental issues. This paper proposes a fast-track methodology to address these two critical requirements when exploring a vast area to locate, in a first approximation, potential sites to build PV plants. This methodology retrieves solar radiation and temperature data from free access databases for the arbitrary division of the region of interest into land cells. Data clustering and probability techniques were then used to obtain the mean daily solar radiation per month per cell, and cells are clustered by radiation level into regions with similar solar resources, mapped monthly. Simultaneously, temperature probabilities are determined per cell and mapped. Then, PV energy is calculated, including heat losses. Finally, PV energy forecasting is accomplished by constructing the *P*50 and *P*95 estimations of the mean yearly PV energy. A case study in Mexico fully demonstrates the methodology using hourly data from 2000 to 2020 from NSRDB. The proposed methodology is validated by comparison with actual PV plant generation throughout the country.

Keywords: clustering; machine learning; solar resource assessment; photovoltaic energy forecasting; regional *P*50 and *P*95 forecasts

Citation: Borunda, M.; Ramírez, A.; Garduno, R.; Ruíz, G.; Hernandez, S.; Jaramillo, O.A. Photovoltaic Power Generation Forecasting for Regional Assessment Using Machine Learning. *Energies* **2022**, *15*, 8895. https://doi.org/10.3390/en15238895

Academic Editor: Emanuele Ogliari

Received: 1 November 2022
Accepted: 22 November 2022
Published: 24 November 2022

1. Introduction

Solar energy is one of the most favorable sources of renewable energy for providing electrical energy in vast quantities without the burden of emitting pollutants to the environment. Because of its global availability, the construction of photovoltaic (PV) power plants (and the installation of PV panels for generation in situ) has become a major trend worldwide [1]. With this purpose, assessing the availability of solar resources and forecasting the photovoltaic energy yield are two major and fundamental issues to be known when making decisions about site selection for PV power plants [2]. Other factors include the physical features of the land, environmental issues, land use regulations, social concerns, and electrical infrastructure availability [3]. Regarding solar irradiation, at least 1100 kWh/m^2 per year is usually required to guarantee technical and economic feasibility [4], but in general, places with higher solar irradiation are preferred.

On the one hand, the assessment of solar resources requires the sufficient collection of reliable radiation data for any specific site of interest, potentially covering very large areas, entire regions, or even a whole country, when exploring a territory of interest to find the best site to place a PV plant. Solar radiation data include global horizontal irradiance (GHI), beam normal irradiance (BNI), diffuse horizontal irradiance (DHI), and global-tilted irradiance (GTI). Currently, radiation data can be obtained from three major sources:

measurement stations equipped with pyrheliometers and pyranometers [5], geostationary weather satellites with remote-sensing capabilities [6], and numerical weather prediction (NWP) models [7]. Most accurate radiation data are obtained with the measurement stations when the sensors are properly calibrated, and the least accurate data are provided by NWP models, although accuracy can be improved using data from measurement stations. A major drawback of ground-based measurements is that assessment over a wide area requires many stations, which also requires frequent condition monitoring and maintenance. This heavily increases the cost of obtaining enough and dependable data. Furthermore, the amount of data publicly available may be scarce from measurement stations, whereas data are abundant when provided by NWP models.

On the other hand, forecasting photovoltaic energy heavily depends on the time scale of the prediction horizon. Prediction horizons vary with the intended use of the forecast, ranging from a few seconds to several years [8]. Intra-hour forecasting, with horizons from 15 min to 2 h ahead, and periods from 30 s to 5 min, is required for ramping events and variability operations. Short-term forecasting, with horizons from 1 to 6 h ahead, and periods of 1 h, is required for loads following the operation. Day-ahead forecasting, with horizons from 1 to 3 days ahead and periods of 1 h, is required for unit commitment, transmission scheduling, and day-ahead markets. Medium-term forecasting, with horizons from 1 week to 2 months ahead, and periods of 1 day, is required for hedging, planning, and asset optimization. Long-term forecasting, with horizons from one to several years, and periods of 1 day, 1 month, or 1 year, is required for resource assessment, site selection, and bankable documentation [9]. The long-term forecasting of PV energy is mainly used to demonstrate the technical and economic viability of PV plant projects. In this regard, since PV energy is directly related to solar radiation, the availability of large solar radiation databases for most of the world makes it very appealing to carry out solar radiation forecasting and, consequently, PV energy forecasting, using data-driven approaches to retrieve data from databases such as the National Solar Radiation Database (NSRDB) [10]. However, the large amount of data that needed to be processed to obtain meaningful information for analysis and decision-making calls for the use of data science methods such as data mining and big data [11], and AI tools such as machine learning [12].

This paper proposes a fast-track methodology to solve the problems of the long-term assessment of solar resources and the forecasting of PV energy to identify viable regions for PV generation, as is required by the initial feasibility analysis to locate potential sites for building PV power plants. This methodology provides the fundamental results to guide the selection of specific locations, where additional criteria should be further investigated before making final decisions on site selection. The proposed methodology allows the retrieval of large amounts of hourly solar radiation data for almost any region of interest in the world, divided into land cells, for as many years as possible, processing and accommodating this information per month per land cell, obtaining statistical measures and the clustering of land cells with similar energy characteristics, and presenting the results of solar resource assessment as solar radiation maps where land cells can be easily picked out as candidate regions. Then, the forecasting of PV energy can be easily carried out in a first approximation, considering PV panel efficiency and operating temperature. The main results are provided as the $P50$ and $P95$ energy forecast atlases for the region of interest. Section 2 introduces the methodology to process radiation data, and methods to calculate the solar energy and the PV energy output. In Section 3, the assessment and prediction of solar resources and PV energy yield are presented for Mexico as a case study, using solar radiation and temperature data from the NSRDB. The results are validated by comparison with the PV energy yield published by actual PV plants throughout the country. In Section 4, a discussion of the major results is included, highlighting the advantages and drawbacks of the methodology. Finally, in Section 5, conclusion statements are provided for the methodology, the case study, and the worldwide applicability.

1.1. The Current State of the Art

Global horizontal irradiation (GHI) information is very important for different applications, such as hydrology, meteorology, and renewable energy for photovoltaic and photothermal systems, as well as for economic and environmental matters. The evaluation and prediction of the GHI can be developed using different methods, which are divided into three categories: physical models, machine learning models, and hybrid models. The accuracy of the models depends on the dataset, time step, forecasting horizon (minutes, hours, days, months, or years), and performance indicators for developing solar energy technology.

Three important characteristics of solar radiation forecasting methods can be found in the literature: (a) the forecast horizon, which is the length of time into the future for which solar energy forecasts are to be prepared; (b) the spatial resolution, which is the measurement of the smallest object in the ground area drawn for the sensor's instantaneous field of view; and (c) the forecast theme, which refers to whether researchers are predicting solar irradiance or PV plant power directly. In the literature, some models for solar irradiance and power forecasting are reported: persistence models, physical models, time series models, machine learning models, deep learning models, artificial intelligence models, and hybrid and ensemble models [12,13]. Table 1 summarizes some reviews of solar irradiance and power forecasting models.

Table 1. Current reviews of solar irradiance and power forecasting models.

References/Year	Title of the Paper	Main Contributions
[12]/2017	Machine learning methods for solar radiation forecasting: A review	This paper reviews more than 100 models for solar irradiance and power estimation.
[13]/2019	Review on forecasting of photovoltaic power generation based on machine learning and metaheuristic techniques	This paper reports better forecasting accuracy for solar power output, in comparison to individual machine learning (ANN, SVM, and ELM) and mathematical techniques.
[14]/2020	Solar irradiance resource and forecasting: a comprehensive review	This paper presents an overview of solar irradiance resources, radiometers, sensor network datasets, and forecast error metrics, and a detailed review of the methods used for forecasting irradiance in different time horizons.
[15]/2021	A Comprehensive Application of Machine Learning Techniques for Short-Term Solar Radiation Prediction	This paper proposes a comprehensive application of machine learning techniques for short-term solar radiation prediction.
[16]/2021	Review on Photovoltaic Power and Solar Resource Forecasting: Current Status and Trends	This paper analyzes articles from specialized magazines with a focus on forecasting techniques for PV power, solar irradiation, and irradiance.
[17]/2021	A Comprehensive Review and Analysis of Solar Forecasting Techniques	This paper reports a comprehensive review of potential models with different techniques having produced significant information, obtained from the study of 170 papers.
[18]/2022	Systematic Review on Impact of Different Irradiance Forecasting Techniques for Solar Energy Prediction	This paper explains more than 100 models by their characteristics and metric performance, with the merits and drawbacks of each.
[19]/2022	Recent advances in intra-hour solar forecasting: A review of ground-based sky image methods	This paper provides a systematic review of GSIIHSF, which is a branch of IHSF and employs GSIs to make predictions.

In recent years, machine learning models have been applied to describe GHI assessment and prediction, since solar radiation data are often difficult to obtain. Machine learning, a subfield of computer science classified as a method of artificial intelligence, is

used in diverse domains. Its advantage lies in its approach to solving problems that are impossible to represent using explicit algorithms. Machine learning models can be used in three different ways in the assessment and forecasting of GHI [12]:

- Structural models are based on other meteorological and geographical parameters.
- Time-series models only consider the historically observed data of solar irradiance as input features.
- Hybrid models consider both solar irradiance and other variables as exogenous variables.

A few different machine learning models have been applied to estimate GHI evaluation and prediction. The main steps of the machine learning model include data preparation, feature selection, data preprocessing, model development, and output set methods. The main machine learning models can be classified as generalized (GM), ensemble-based (EM), decomposition-based (DM), transition-based (TM), postprocessing-based (PM), decomposition-cluster-based (DCM), and cluster-based (CM) [20].

In the case of the clustering model (CM), the input data are classified into different groups by a particular algorithm. The data grouped in sets share similar characteristics between them. In this way, these data have similar patterns or characteristics. A CM is a machine learning model that does not require supervision, that is, it does not need user intervention since the model can find hidden and complex structures in its data inputs without knowing the data outputs. The objective of the CM is to obtain high similarity within groups and low similarity between groups during the grouping of the input data. It can be said that this is the main internal quality criterion of a cluster. However, a high acceptance of these internal criteria does not necessarily mean good efficiency in a cluster application. Many clustering methods of high-dimensional data can be found in the literature; for more information, see reference [21]. In the following paragraphs, the main works published in recent years are presented.

In 2013, Zagouras et al. [22] established a CM for optimizing the location of measuring sites for the newly built Hellenic Network of Solar Energy (www.helionet.gr accessed on 31 October 2022). This CM is a k-means algorithm used for cluster analysis based on the dominance of the cloud effect on solar irradiance and the advantage of the high spatial resolution of a geostationary satellite. Through the validation of the clustering method, their results show that the variability of surface solar irradiance due to cloudiness over Greece could be sufficiently monitored with the establishment of 22 ground-based instruments.

In 2015, Polo et al. [23] presented the spatial variability of long-term solar radiation in Vietnam by clustering solar radiation into different regions using sunshine duration measurements. They proposed a model using the Angstrom equation, which is based on canonical correlation analysis, achieving good performance. They characterized the dispersion of long-term solar radiation and analyzed the spatial distribution by clustering techniques. Additionally, they developed a comparison with the Köppen climatic information, defining 3–4 well-defined zones of different solar radiation variability.

One year later, in 2016, Jiménez-Pérez and Mora-López [24] proposed a model to forecast and estimate hourly global solar radiation. The model consists of a clustering algorithm to identify the type of day based on decision trees, artificial neural networks, and support vector machines. Their model was validated using data recorded in Malaga. Their results show that it is possible to predict next-day hourly values of solar radiation values with an rMAE of 15.2% for one of the input datasets, while the rMAE is 16.7% for the other set of input parameters.

In 2018, P. Govender et al. [25] developed two forecasting methods based on CM. The first model is based on k-means clustering to predict daily cloud cover profiles. The other describes a rule for predicting cloud cover profiles. They reported that the two methods had a comparable success rate of approximately 65%; the cloud cover clustering method was better for sunny and cloudy days, and the 50% rule was better for mixed cloud conditions. In 2018, a paper by Rodríguez-Benítez et al. [26] reported the evaluation of the modes of intra-day variability of solar resources in the Iberian Peninsula. The GHI was

associated with meteorological patterns and the impact on solar production was evaluated. Their analysis was performed for annual and seasonal variability. Considering two years of measured global horizontal irradiance (GHI) and direct normal irradiance (DNI) data gathered at four stations, the modes of variability were identified using hierarchical cluster analysis. They used three-hour statistics describing the mean and used the variability in solar radiation as input data for the cluster analysis. They evaluated the synoptic weather patterns associated with each cluster resulting from the cluster analysis by using cloud cover and sea-level pressure data. Their results indicate the existence of four modes of variability of solar resources for annual analysis. Their seasonal analyses show similar results to the annual analyses, but with marked seasonal differences.

In 2019, Lopes-de-Lima et al. [27] studied the northeastern Brazilian region, the country's most abundant solar energy resource. They investigated the surface solar irradiation variability and trends based on clustering analysis. Their results point out a remarkable variability in seasonal and annual scales. The cluster analysis provided five regional patterns, presenting quite interesting complementary temporal regimes for the incoming solar irradiation.

In 2020, Theocharides et al. [28] proposed a forecasting methodology that considers a data quality evaluation stage, for the development of a data-driven PV power output machine learning model (ANN). The proposed methodology also considers the evaluation of climatic clustering (K-means clustering). Their results show that the optimized model has a mean absolute error of 4.7%. Finally, they validated their model, finding a forecast precision of 4.7% and an absolute error of 6.3%.

In 2021, the following articles were published:

Jayalakshmi et al. [29] proposed a multi-temporal scale model for the prediction of solar irradiance. Their model is based on a multitask learning algorithm and is implemented with a short-term memory (LSTM) neural network model. Its performance for various time windows was investigated. The estimation of the hyperparameters involved in the proposed LSTM model was performed using a hybrid swarm optimizer. The proposed model was validated, comparing the existing methodologies for the forecast of a single time scale. Their results show that the strategy exhibits a highly consistent performance for forecasting across all timescales, with improved metric results.

Behr et al. [30] analyzed daily values from a 25-year dataset (1991–2015) obtained by satellite sensors, representing the long-term, large-scale evolution of incident surface solar radiation, and larger-scale cloud dynamics in a spatial area of $0.05^\circ \times 0.05^\circ$. They reported that the most significant long-term increase in solar radiation was observed in spring. They reported that there is little solar–solar complementarity in different regions since the dynamics between mechanisms are similar and do not show significant differences. The methods they proposed to assess complementarity showed the areas where solar potential is not yet fully exploited.

Pham Thi Thanh Nga et al. [31] applied k-means clustering to solar irradiation based on satellite data (Himawari-8 satellite) in different regions of Vietnam. The satellite data were validated with observations recorded at five stations in the period from October 2017 to September 2018. They defined six cluster groups, demonstrating a better agreement with the conventionally classified seven climatic zones than the four climatic zones of the Köppen classification. They obtained the spatial distribution and seasonal variation in the regionalized solar irradiation. Additionally, they found the highest and the lowest daily average solar radiation in two clusters in the southern region, where the South Asian summer monsoon dominates in the rainy season.

Watanabe et al. [32] described a method based on a self-organizing map and cluster analysis. They analyzed five consecutive days for the regional and seasonal characteristics of GHI and then used one hour of accumulated GHI data from ground observation stations in Japan. Their results show that there are three major regions in Japan. Additionally, they conducted another cluster analysis to investigate the seasonal characteristics of the occurrence of time series patterns. Their findings indicate that consecutive cloudy days

occur frequently in winter and during the rainy season, whereas consecutive clear days occur frequently in spring and summer.

Borunda et al. [33] used k-means and k-medoids algorithms to perform a cluster analysis of solar radiation in several representative locations in Mexico, obtaining a preliminary seasonality atlas for solar resources.

In 2022, Ali-Ou-Salah et al. [34] presented a new hybrid approach based on seasonal CM and an artificial neural network (ANN) for forecasting 1 h ahead of GHI. They used a fuzzy c-means algorithm (FCM) to cluster 3 years of monthly average experimental data from Évora city. The meteorological dataset was divided into training subsets based on the seasonal clustering results. Furthermore, an ANN model for each subset was designed to forecast hourly global solar radiation. In the same year, Maldonado-Salguero et al. [35] proposed a CM to determine the spatio-temporal solar resource variability through GHI analysis. They used a hierarchical clustering technique to classify the spatial data. They proposed different time windows—from short-term to long-term data—to evaluate GHI, considering different information sources. Based in Spain, and considering a 22-year period (1999–2020), they reported 1,936,917 observations from an online satellite database. Their approach provides an alternative method for the comprehensive spatio-temporal clustering and characterization of GHI evolution.

Another paper reported in 2022 was that of Salinas-González et al. [36]. They reported a multivariate analysis considering four variables: cloudy sky index, albedo, Linke turbidity factor (TL2), and altitude in satellite image channels. They considered principal component analysis (PCA) to reduce the database's dimensionality (satellite images). In their model, cluster analysis with unsupervised learning was performed, and two clustering techniques were compared: k-means and Gaussian mixture models (GMMs). By considering k-means, they obtained a minimum number of regions with a similar degree of homogeneity. This case study was developed for Mexico. They considered the optimal number of regions to be 17. These regions were compared in terms of the annual average values of daily irradiation data from ground stations using multiple linear regression, showing the regions that are strongly related to solar irradiance.

Table 2 summarizes the literature linked to the objectives of this publication, reporting the place where the modeling is carried out, the time horizon, and the main conclusions.

In recent years, the penetration of photovoltaic generation has increased, mainly due to strategies and objectives targeting climate change. For the proper implementation of PV power, it is essential to correctly forecast production, allowing the negative effects associated with the need for solar resources to be minimized. This also has a direct impact on the surplus or deficit of electricity generation, which can disturb the electrical network and cause instability. In addition, the demand for supplemental or reserve energy must be considered to avoid such fluctuations. Regional PV forecasting is crucial for transmission and distribution system operators to operate networks under the relevant grid codes. Forecast models are the basis for developing and improving smart grids. Smart networks must allow the exchange of information between suppliers and customers. Thanks to the great innovations made in intelligent communication, monitoring, and management systems, it is possible to develop intelligent photovoltaic networks. Solar resource forecasting models are required when making estimations due to the complexity of the topology of the transmission and distribution systems, and the predictability in the management of the dispatch to the electrical network [37,38].

Table 2. Summary of the state of the art for the clustering model (CM).

Ref/Year	Location	Forecast Horizon	Data	Main Conclusions Reported by the Authors
[22]/2013	Greece	Annual	2009–2010	"Estimated values of the cloud modification factor during local noon time with a spatial resolution of 0.05, derived from the SEVIRI instrument on-board MSG satellite for the 2009–2010 time period, were used for cluster analysis with the k-means algorithm"
[23]/2015	Vietnam	Trimester	2003–2012	"A mode inspired by the Angstrom equation has been developed by using daily clear sky global irradiation computed with REST2 model and sunshine duration by using the canonical correlation analysis and fitting the results to four cubic polynomials, corresponding to each trimester of the year"
[24]/2016	Spain	Hourly	2010–2013	"Two procedures have been proposed and two experiments have been performed using different input data sets depending on the used independent variables. Results show that the best method combination is SVM-C to estimate the cluster of each observation and SVM-R to estimate the daily clearness index"
[25]/2018	South Africa	Hourly	2014–2015	"Clustering of Bnfor Durban, South Africa, produced four classes with diurnal patterns as follows: sunny all day (Class A), cloudy all day (Class B), sunny morning and cloudy afternoon (Class C) and cloudy morning and sunny afternoon (Class D)"
[26]/2018	Iberian Peninsula	Hourly	2015–2017	" . . . a hierarchical cluster analysis, applied over radiation statistics from four stations within the study area, is used. Independent yearly and seasonal analyses are conducted. The number of groups is determined using the mean MSLP anomaly field of each group, ensuring that all the associated synoptic weather patterns are significantly different and meaningful in each cluster analysis"
[27]/2019	Brazil	Annual	2005–2015	"Five clustered regions (HR) have a geographical location consistent with the regional climate characteristics and typical meteorological systems operating in each HR. The HR5, the driest area, has the highest daily average of global solar irradiation. The inter-annual variability is high in the mid-eastern area (HR3) due to the cloudiness associated with typical meteorological phenomena in the region"
[28]/2020	Cyprus/USA	Daily	2018	" . . . the model was validated both, at a hot as well as a cold semi-arid climatic location, and the obtained results demonstrated close agreement by yielding forecasting accuracies of mean absolute percentage error of 4.7% and 6.3%, respectively. The validation analysis provides evidence that the proposed model exhibits high performance in both forecasting accuracy and stability"
[29]/2021	India	Minutes	2020	"The model is evaluated with performance metrics such as MSE (mean square error), MAPE (mean absolute percentage error), and DA (direct accuracy), and to signify the obtained performance is not affected by the algorithm's stochastic parameters, a statistical analysis is undertaken. The proposed model outperformed others with better metric results for single-time scale forecasting and multi-time scale forecasting with better metric results"

Table 2. *Cont.*

Ref/Year	Location	Forecast Horizon	Data	Main Conclusions Reported by the Authors
[30]/2021	Germany	Seasonal	1991–2015	"The rarely studied inter-annual variability of SIS and CFC is much greater than their long-term variation. This has also a substantial impact on the strategic planning of PV electricity production. The coupling with other renewables and extensive, long-term storage must be considered to compensate for the inter-annual fluctuations in exploitable solar energy"
[31]/2021	Vietnam	Monthly	2016–2018	"The results of k-means clustering applied to the 3-yr satellite-based GHI illustrated the best 6-cluster groups with good spatial homogeneity for regionalization in Vietnam. This regionalization demonstrated a better agreement with the conventional classification of the seven climatic zones rather than the four Köppen classified climatic zones"
[32]/2021	Japan	Annual/Seasonal	2013–2019	"In the analysis of seasonal characteristics, another cluster analysis is performed using a two-level approach. In this analysis, time series data are divided into four groups and the number of stations at which the same cluster occurs simultaneously is investigated. It is found that the cluster in which cloudy conditions are maintained for 5 days has peaks of the number of stations that are simultaneously assigned to the cluster in the rainy season and in winter, whereas the cluster with five consecutive clear days has the peaks in spring and summer"
[33]/2021	Mexico	Annual/Seasonal	2000–2020	"This work performs a cluster analysis to determine the seasonality of the solar radiation of different locations. We use k-means and k-medoids algorithms, and even though both are partitioning algorithms, we end up preferring k-medoids to find the seasonality since the centroids of the clusters belong to data from the dataset and therefore a straightforward interpretation is generated"
[34]/2022	Portugal	Monthly	2012–2016	"In this paper, a new hybrid approach based on seasonal clustering technique and ANN model has been presented for forecasting hourly global solar radiation"
[35]/2022	Spain	Monthly	1999–2020	"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"
[36]/2022	Mexico	Annual	2015	"K-Means and GMM are both unsupervised clustering techniques but work differently. K-means groups data points using Euclidean distance for cluster membership. K-means is widely used due to its simplicity and speed. GMM uses a probabilistic assignment of data points to clusters . . . "

2. Materials and Methods

Solar radiation continuously changes throughout both the day and the year. Figure 1 shows a solar radiation chart for a typical summer and winter day for a site in the north of Mexico, Hermosillo, Sonora, as an example. The integral under the curves should be calculated to compute the solar energy in a horizontal plane, corresponding to the PV panel placed on the ground.

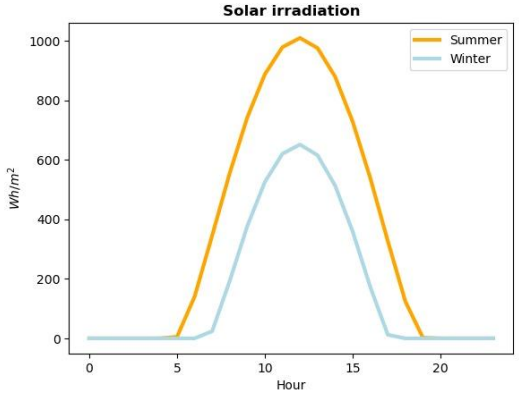

Figure 1. Daily solar radiation chart for a typical summer and winter day in the north of Mexico, Hermosillo, Sonora.

However, these curves change daily, and as the site moves further from the Equator, the difference between the curves throughout the year becomes greater. Additionally, the daily curves differ from year to year. Thus, radiation charts of typical days are commonly used by each of the stations to calculate the available solar energy at a given site.

There exist large datasets of meteorological information for many locations stored in public sites, such as the NSRDB [39], which contain many years of daily, hourly, and 5-min data for any site at a given latitude and longitude. This information is extremely useful, and many machine learning techniques can be used for prediction purposes. The versatility of the methodology proposed in this work is that it uses all information contained in all available data, regardless of meteorological and geographical situations for a small or extended region.

Big data for solar radiation are a deep ocean of information, encoding a great deal of important information. One of the main objectives is to obtain information on the solar energy that can be captured to produce photovoltaic energy. The evolution of the behavior of solar energy by region is of the utmost importance in order to evaluate photovoltaic use. In addition to solar radiation, there are many other important factors to consider for the optimal performance of photovoltaic panels. For example, ambient temperature is a fundamental factor in the performance of solar conversion in a photovoltaic system, since, as temperature increases, the performance of solar cells decreases [40], and therefore, the PV panel efficiency decreases [41,42]. In this section, a methodology is developed for the analysis of the regional variation in solar energy and ambient temperature throughout the year with the intention of (a) evaluating the best sites for PV deployment, and (b) forecasting the produced PV energy. To achieve these goals, a statistical approach is used, together with machine learning. In particular, the average solar energy is statistically calculated, the clustering of solar energy is carried out, the ambient temperature behavior is studied, and, finally, the *P*50–*P*95 estimations for the produced PV energy are calculated. In the following subsection, the proposed methodology is described.

2.1. Methodology

The proposed methodology is shown in Figure 2. The first step is to build a grid in the region of interest. The grid design considers for each uniform cell an evaluation area that is determined by the user, which can be larger or smaller depending on the specific

requirements. Then, the center of each grid cell must be located by its latitude and longitude in a geodetic reference geographic coordinate system. The center of each cell is preferred as a representative site of the place. All the available time series for the GHI and ambient temperature in each of the grid cells are then downloaded.

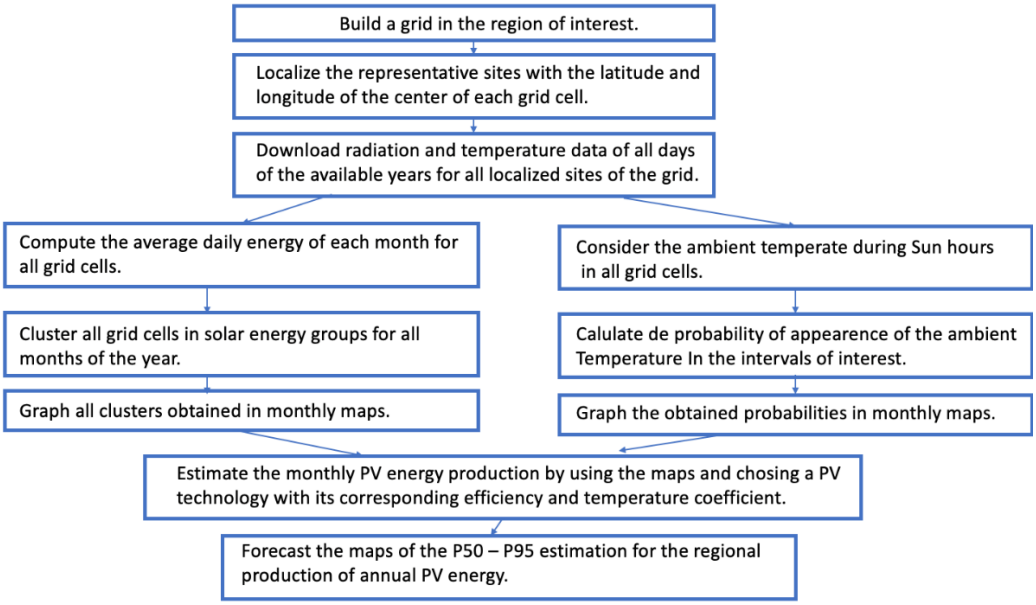

Figure 2. Methodology to obtain a regional map of mean daily solar energy and a map of probability of occurrences of high temperatures per month, and the atlases to forecast the annual PV production in a region using *P*50–*P*95 estimations.

Then, on one hand, the average daily and monthly solar energy are calculated during the hours of sunlight. Subsequently, the sites are grouped into energy intervals for each month. Finally, the maps of the resulting energy clusters for each month of the year are obtained. This provides monthly behavior at a regional level for the incident solar energy.

On the other hand, the hourly probability of the appearance of ambient temperatures, defined by the user, is calculated in each cell, and graphed in maps each month. This provides the monthly behavior at the regional level of the ambient temperature and offers information about sites where heat losses can be higher.

Finally, using radiation and temperature maps, it is possible to inspect the best sites for PV generation. Using an hourly computation, the PV generation is calculated, and assuming a normal distribution over the years, a forecast of the generated PV energy can be conducted. Therefore, the results provide a guide for constructing featured maps given GHI and ambient temperature datasets. The following subsections explain the methodology step by step.

2.2. Calculating Solar Energy

Daily, hourly, or even every 5 min, GHI data are available on the NSRDB site for a given latitude and longitude. The first step consists of downloading the daily radiation data for all cells, for all days of the year, for all available years. Next, the cumulative daily solar energy of day j of month m of the year y, $E_{day_{d,m,y}}$, is calculated as

$$E_{day_{d,m,y}} = \sum_{h=1}^{H} W_{h,d,m,y} \Delta t_h, \quad (1)$$

where $W_{h,d,m,y}$ is the GHI of day d of month m of year y, during the time interval Δt_h, and h runs from 1 to the H available data for that day, such that, if there were 8 h of sun during the day, and considering hourly data, $H = 8$.

The mean daily energy of m month in y year is

$$\overline{E_{month_{m,y}}} = \frac{1}{D}\sum_{d=1}^{D} E_{day_{d,m,y}}, \tag{2}$$

where D is the number of days in each month. Additionally, the mean daily energy of month m over N years is

$$\overline{E_{month_m}} = \frac{1}{N}\sum_{y=1}^{N} \overline{E_{month_{m,y}}}. \tag{3}$$

Given these definitions, the next step is to cluster the mean daily solar energy into energy intervals. The following subsection provides the basics of the grouping process.

2.3. Clustering

Many algorithms are used for clustering. K-means is one of the most used unsupervised algorithms in data mining due to its simplicity, since it groups data in a very intuitive way, through Euclidean distance minimalization. Given a set of n data $(x_1, x_2, \ldots, x_n)$, where each data point is a d-dimensional vector, the algorithm groups it in $k(\leq n)$ clusters $C = \{C_1, C_2, \ldots, C_k\}$ such that the Euclidean distance between the objects and the mean of the points μ_i in C_i, which are the centroids, is minimized [43]

$$argmin \sum_{i=1}^{k} \sum_{x \epsilon C_i} ||x - \mu_i||^2. \tag{4}$$

The Silhouette method is used to provide the best number of groups and test the goodness of the clustering. It is used when in the dataset exists an intrinsic natural number of clusters [44]. The Silhouette value measures the similarity between objects in the same cluster compared with objects in other clusters. Considering i as a data point in the i-th cluster C_i, the distance between i and all other data in the same cluster is defined by

$$a(i) = \frac{1}{|C_i| - 1}\sum_{j \in C_i, i \neq j} d(i,j), \tag{5}$$

where $d(i,j)$ is the distance between the points i and j in C_i. Likewise, the smallest mean distance of i to all points in the other group is

$$b(i) = min \frac{1}{|C_k|}\sum_{j \in C_k} d(i,j), \tag{6}$$

To evaluate if the data point i is properly grouped, one calculates

$$s(i) = \begin{cases} \frac{b(i)-a(i)}{max\{a(i),b(i)\}}, \; if \; |C_i| > 1, \\ 0, \; if \; |C_i| = 1. \end{cases} \tag{7}$$

where $-1 \leq s(i) \leq 1$. If $s(i) = 1$, this means the data are well-grouped, but if $s(i) = -1$, then it should be in the neighboring group. The Silhouette score SS is a measure of the goodness of the clustering and corresponds to the mean $\boldsymbol{s}(i)$ overall data of the group. The Silhouette coefficient SC provides the best number of clusters k and is given by the maximum $\boldsymbol{s}(i)$ overall data. It is given by

$$SC = max\boldsymbol{s}(k). \tag{8}$$

In this way, the average daily solar energy of the region of interest is clustered according to the intervals of energy per unit area to find the regions with the best solar resources. Once the incidence of radiation is known, the next step is to calculate the photovoltaic power. Section 2.4 provides the methodology used for its calculation.

2.4. PV Power Generation

PV systems can be either grid-connected, stand-alone, or hybrid. Stand-alone PV directly satisfies the load requirements by having a system size that supplies the required demand. On the other hand, grid-connected PV systems are coupled to the grid. Hybrid PV systems present both characteristics. A PV power plant, also known as a solar farm, is a large-scale PV system that converts sunlight directly into electric power, which is then fed into the power networks in bulk quantities. Figure 3 shows the typical configuration of a PV power plant. The arrays of PV panels convert the incoming solar radiation into electricity, producing direct current (DC) with a voltage of up to 1500 V. The power electronic inverters convert DC into alternating current (AC) at 60 Hz. AC voltage is raised to 22 kV from the output of the inverter to the PV plant feeders by the mid-voltage transformers. Electric power is added up from all feeders at the substation high-voltage transformer, where the voltage is raised to the required level at the point of interconnection (POI) for transmission to other areas of the power system.

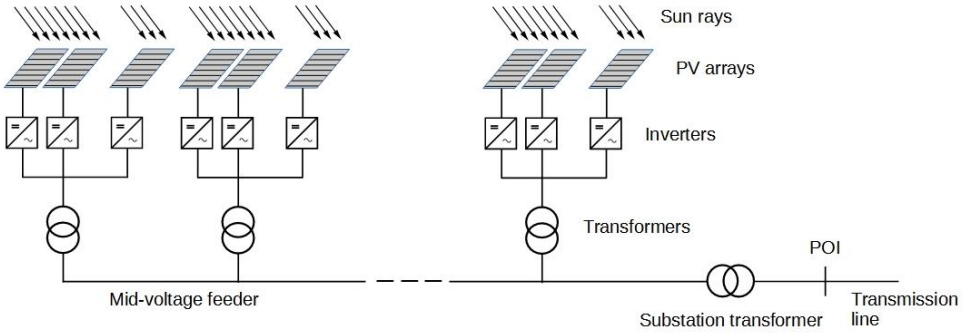

Figure 3. Typical PV power plant topology.

Inside each PV panel, the solar radiation is converted into electricity by cells made of semiconductor materials (i.e., silicon). In the cells, the semiconductor atoms free up electrons after absorbing the photons of the incident sunlight, which is known as the photoelectric effect [45]. When provided with a conducting path, the free electrons can flow, producing an electric current. Therefore, PV panels are modules composed of many PV cells conveniently interconnected to combine their currents and voltages to obtain values favorable for practical use. The power production of a PV panel mainly depends on incident global radiation and ambient temperature, as given by [46]

$$P = P^{STC}\frac{W}{W^{STC}}\left[1+\gamma_C\left(T_C - T_C^{STC}\right)\right]\left[1+C_C ln\frac{W}{W^{STC}}\right], \quad (9)$$

where the superscript STC indicates standard test conditions. P is the produced power due to the incident radiation W, T_C is the cell's temperature, and γ_C and C_C are the temperature coefficient at maximum power and the radiation coefficient of the cell, respectively. Therefore, the PV panel produces P^{STC} at an incident radiation of W^{STC}. STC is indicated in the specification sheet of a PV panel and usually corresponds to an incident radiation of 1000 W/m^2, cell temperature of 25 °C, and AM 1.5 G, which refers to two standard terrestrial solar spectral irradiance spectra, namely a direct normal and a standard total spectral irradiance. The temperature coefficient quantifies the power variation as the cell temperature increases. Likewise, the radiation coefficient measures the efficiency variation as the irradiance decreases at a constant temperature. γ and C_C are negative numbers, corresponding to performance losses, and are provided by the manufacturer. C_C is an order of magnitude smaller than γ_C; thus, the produced power can be calculated by

$$P = P^{STC}\frac{W}{W^{STC}}\left[1+\gamma_C\left(T_C - T_C^{STC}\right)\right], \quad (10)$$

As the cell's temperature increases, heat reduces the cell's efficiency. The cell's temperature depends on the ambient temperature T and on the PV panel material as follows

$$T_C = T + \left[\frac{NOCT - 20\ ^\circ\text{C}}{800\ \text{W}}\right] W, \tag{11}$$

where the nominal operating cell temperature, $NOCT$, is given by the manufacturer.

It is important to forecast the annual PV power production for the installation of power plants and to guarantee the monthly and yearly minimal production. This can be achieved using Equations (10) and (11), considering the specific characteristics of a selected PV module.

PV power production strongly depends on the temperature during the day, as shown in Equation (10). Thus, it is important to consider the average daily temperature of the month, m, calculated as

$$\overline{T_{day_m}} = \frac{1}{N}\frac{1}{D}\frac{1}{H}\sum_{y=1}^{N}\sum_{d=1}^{D}\sum_{h=1}^{H} T_{h,d,m,y}, \tag{12}$$

where $T_{h,d,m,y}$ is the temperature value at time interval Δt_h of day d of month m of year y. The probability of the occurrence of high temperature at a given site can be calculated with all available data for each grid cell. The results can be shown in monthly maps. It is important to infer the efficiency losses for temperatures higher than 25 °C in order to select the best sites for PV deployment, based not only on their incident radiation but on the prevention of large drops in performance due to high ambient temperature.

Once a given site and PV module are selected, the cumulative daily PV energy produced on day d of month m of year y is given by

$$E^{PV}_{day_{d,m,y}} = \frac{P^{STC}}{W^{STC}}\sum_{h=1}^{H} W_{h,d,m,y}\left[1 + \gamma_C\left(T_{h,d,m,y} + \left[\frac{NOCT - 20\ ^\circ\text{C}}{800\ \text{W}}\right] W_{h,d,m,y} - T_C^{STC}\right)\right]\Delta t_h. \tag{13}$$

Then, similarly to Equation (3), the mean daily PV energy of month m is calculated as

$$\overline{E^{PV}_{month_m}} = \frac{1}{N}\frac{1}{D}\sum_{y=1}^{N}\sum_{d=1}^{D} E^{PV}_{day_{d,m,y}}. \tag{14}$$

Finally, the computation of the annual forecasted PV energy is described in the following subsection.

2.5. Forecasting PV Energy

The Gaussian distribution (GD) is one of the most frequently observed data distributions in nature, since this distribution better describes systems for which entropy is maximized, and GD is a very straightforward distribution. The probability density function (PDF) of a GD is characterized by its mean value, μ, and its standard deviation, σ^2, as

$$PDF(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}. \tag{15}$$

Thus, the probabilistic approach supposes that the annual PV energy production distribution obeys a GD over several years of operation, and this distribution is used to obtain estimates of the PV energy yield in the years to come based on statistical levels of confidence. Two common estimates are the $P50$ and $P95$ energy yield estimates, which indicate that the annual PV energy yield will be exceeded by 50% probability and 95% probability, respectively, as shown in Figure 4. Note that $P50 = \mu$, and $P95$ is calculated by solving

$$PDF(x > P95) \overset{yields}{\rightarrow} P(x > P95) = \frac{1}{\sigma\sqrt{2\,\pi}}\int_{P95}^{\infty} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = 0.95. \tag{16}$$

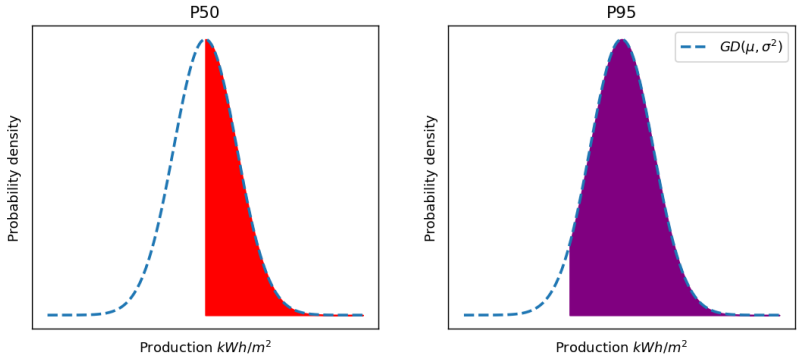

Figure 4. Gaussian distributions with the *P*50 and *P*95 estimations are shown in the left and right figures, respectively.

In this case, the mean value, $\mu = \overline{E_{year}^{PV}}$, is calculated by computing the mean PV energy produced in N years,

$$\overline{E_{year}^{PV}} = \frac{1}{N} \sum_{y=1}^{N} \sum_{m=1}^{M} \sum_{d=1}^{D} E_{day_{d,m,y}}^{PV}, \tag{17}$$

whereas the standard deviation, $\sigma = \mathrm{SD}$, is calculated as

$$\mathrm{SD} = \sqrt{\frac{\sum_{y=1}^{N} \left(\overline{E_{year_y}^{PV}} - \overline{E_{year}^{PV}} \right)^2}{N}}. \tag{18}$$

Thus, the forecast of the annual PV energy is given by the *P*50–*P*95 estimations. This criterion is the most frequently used for PV plant planning and financial analysis in order to calculate the initial investment and its return, as a first step to assess the feasibility of the deployment of PV technology. The proposed methodology allows the best regions for PV power generation to be determined, either at different times of the year or annually, as demonstrated by the case study presented in the next section.

3. Results

A case study in Mexico is considered in this section. According to the methodology described before, a grid was built for Mexico consisting of 731 cells. The locations were chosen by dividing a rectangular region around the Mexican territory into a 50 × 50 mesh grid. Only cells whose center was in the continental Mexican territory were considered. For the center of each grid cell, hourly global horizontal radiation, GHI, and ambient temperature, T, alongside other meteorological data, were downloaded from the NSRB. The data acquisition was conducted with a python script using the NSRBD Application Programming Interface (API) [47]. For Mexico, 20 years of information was available, from 2000 to 2020. Hourly data for radiation, ambient temperature, and other meteorological variables were available. It is worth noting that a larger number of cells could be considered; however, due to time restraints on downloading data from the NSRBD API, we were limited to 731 cells. It is possible to download up to 5000 files per day from the NSRBD API. Each file contains one year of meteorological information. Thus, 20 years of data for the 731 cells, corresponding to 14,620 files, required 3 days for data download. Finer spatial resolution is possible, requiring more time to download data. The obtained data were compressed into a single HDF5 file to reduce the file size from several GB to 350 MB. The compressed data and the python scripts used in this work are available in a link contained in the Supplementary Materials.

3.1. Clustering Maps of Mean Daily Solar Energy

Once the radiation dataset was available, the accumulated daily solar energy and the average daily solar energy for each month were calculated according to Equations (1)–(3)

for each cell. Then, these values were clustered into five energy intervals, as shown in Figure 5. These maps correspond to five clusters of sites grouped with the K-means algorithm such that sites within a group have a similar mean daily incident solar energy. The centroid of each cluster corresponds to the mean daily energy, the value of which is given by the color. The lightest color corresponds to the highest mean daily incident solar energy, 7.69 kWh/m^2, and the darkest color to the lowest, 4.04 kWh/m^2. For all maps, the Silhouette score was calculated, resulting in a value of approximately 0.6, which indicates that the goodness of the clustering is acceptable. This information allows regions with similar solar potential to be grouped, and identification of the regions with the best solar potential throughout the year. The highest mean daily solar radiation was present in the northwestern part of the country during April, May, and June. However, this region corresponds to the lowest mean daily solar radiation during the cold months. Therefore, these maps provide useful information for regional PV deployment throughout the year. Thus, as the first step in site selection, energy needs throughout the year must match resource availability.

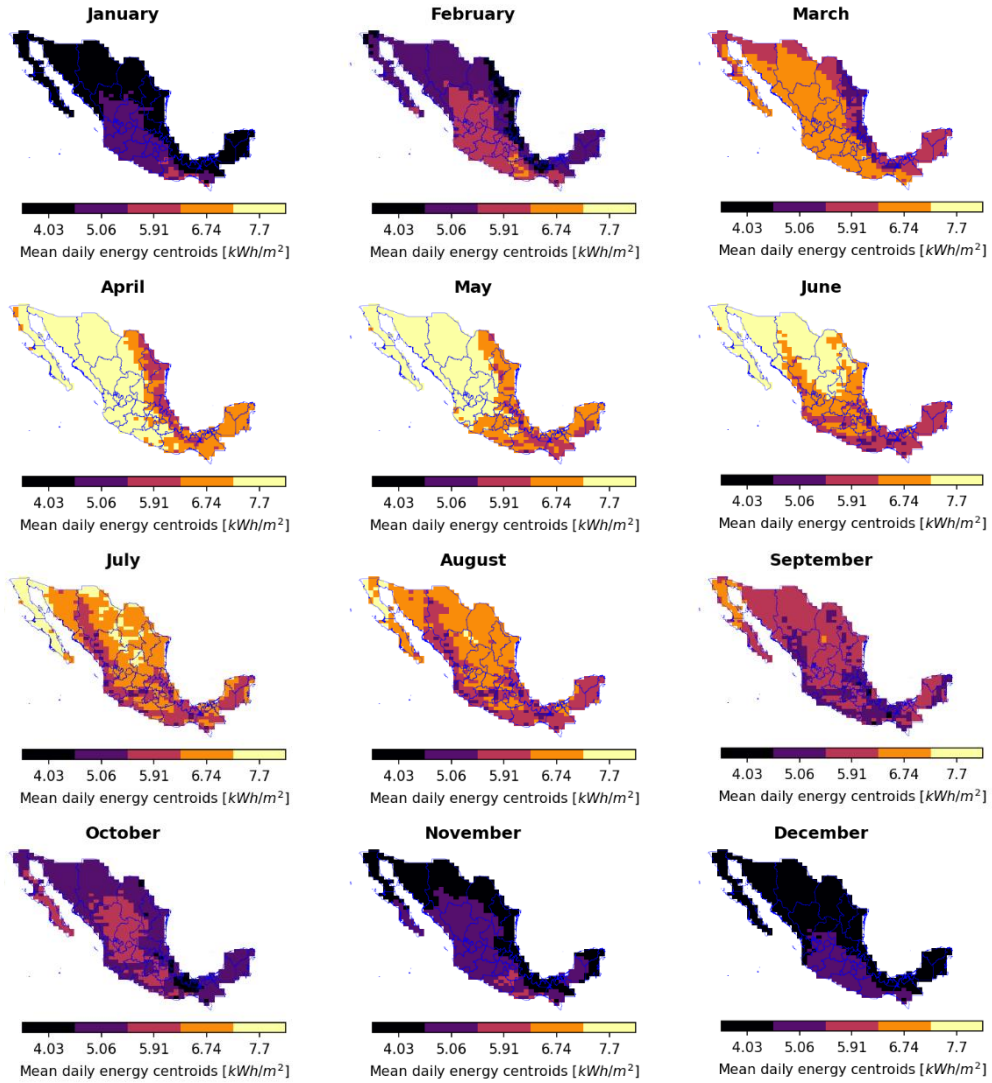

Figure 5. Cluster maps of mean daily solar energy for each month in Mexico.

However, PV generation depends on the ambient temperature, as shown in Equation (9); therefore, the probability of the occurrence of high temperatures in each region is analyzed in the following subsection.

3.2. Probabilities of Temperature Occurrences

The increase in ambient temperature is fundamental for the operation of a photovoltaic module. As the ambient temperature exceeds 25 °C, the photovoltaic module suffers higher heat losses, resulting in a drop in efficiency, according to Equation (10). Since the ambient temperature changes throughout the day, the photovoltaic plant can work with maximum efficiency during some hours, while during other hours it operates with less efficiency. In this subsection, the hourly ambient temperature at each site during sunny hours is considered and the probability of the occurrence of high temperatures is calculated.

This analysis can be performed for each degree Celsius above 25 °C, but for simplicity, the probability of occurrence is first calculated for temperatures between 25 °C and 30 °C. Figure 6 shows the results for the twelve months of the year. Below each map, a gradient color line indicates the probability that the ambient temperature is between 25 and 30 degrees. Figures 3 and 4 show that there are regions with good solar potential but with a high probability that their ambient temperature exceeds 25 °C, implying that the efficiency of photovoltaic technology will decrease in those places.

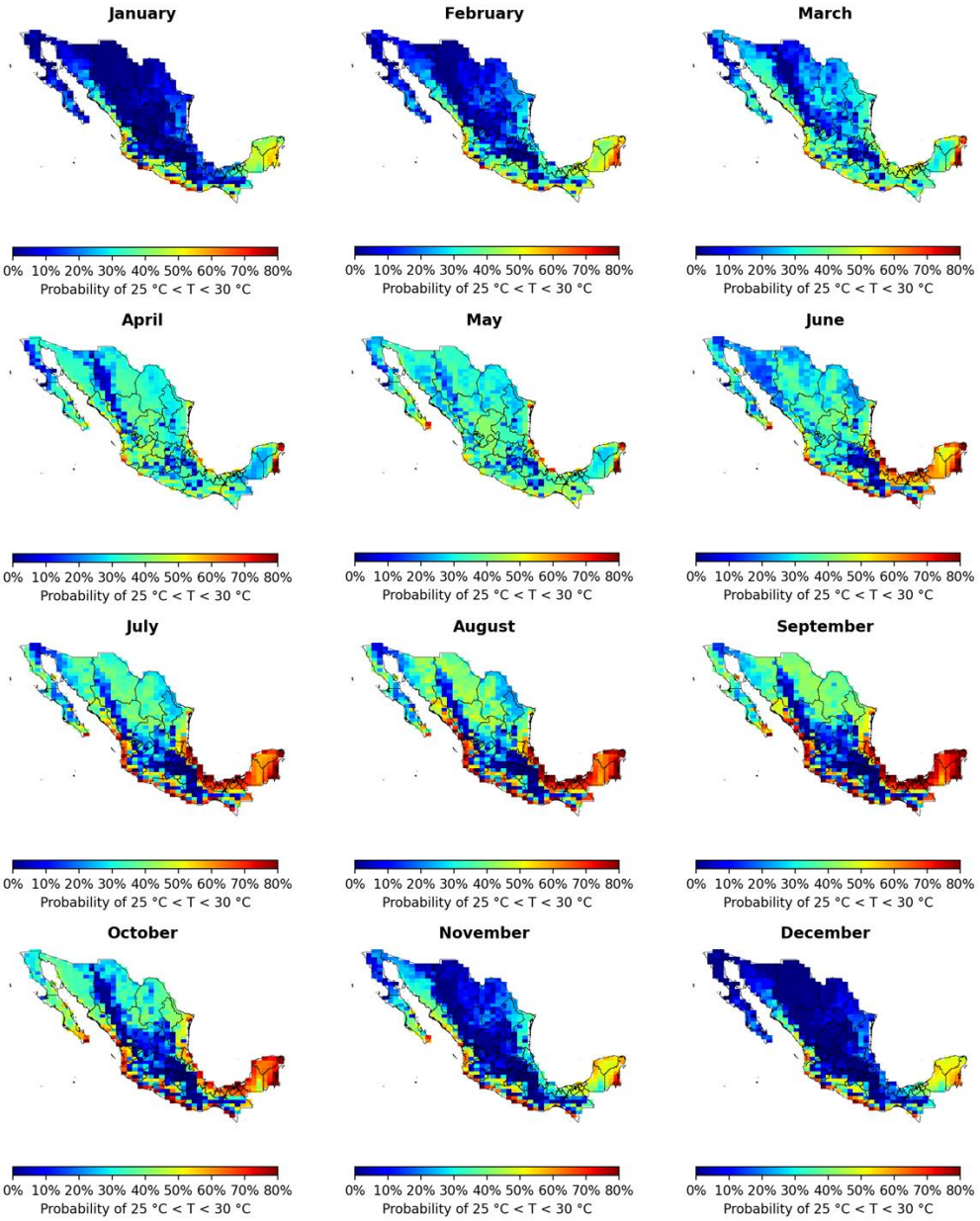

Figure 6. Probability map of ambient temperatures between 25 °C and 30 °C per month in Mexico.

In addition, Figure 7 shows the probability that the ambient temperature is between 30 and 35 degrees. Thus, those regions that have good solar resources but, with a high probability of exceeding 30 °C, will present an even greater decrease in the performance of photovoltaic technology. These results suggest that the regions with the greatest solar resources are not the most convenient for the installation of photovoltaic plants. If the ambient temperature increases, the photovoltaic module performance may be less than in regions with lower solar radiation. In other words, the selection of a site for the installation of a photovoltaic power plant should not be based on the identification of the greatest solar resources; rather, the regions that present the greatest solar resources combined with a lower temperature should be considered. In particular, the northwestern part of the country exhibits a higher probability of an ambient temperature greater than 30 °C during April, May, June, July, and August. Thus, even though the northwestern part of Mexico could be the best candidate for PV deployment given its high irradiance, this region also presents high temperatures, leading to bigger heat losses and thus obtaining lower PV energy production.

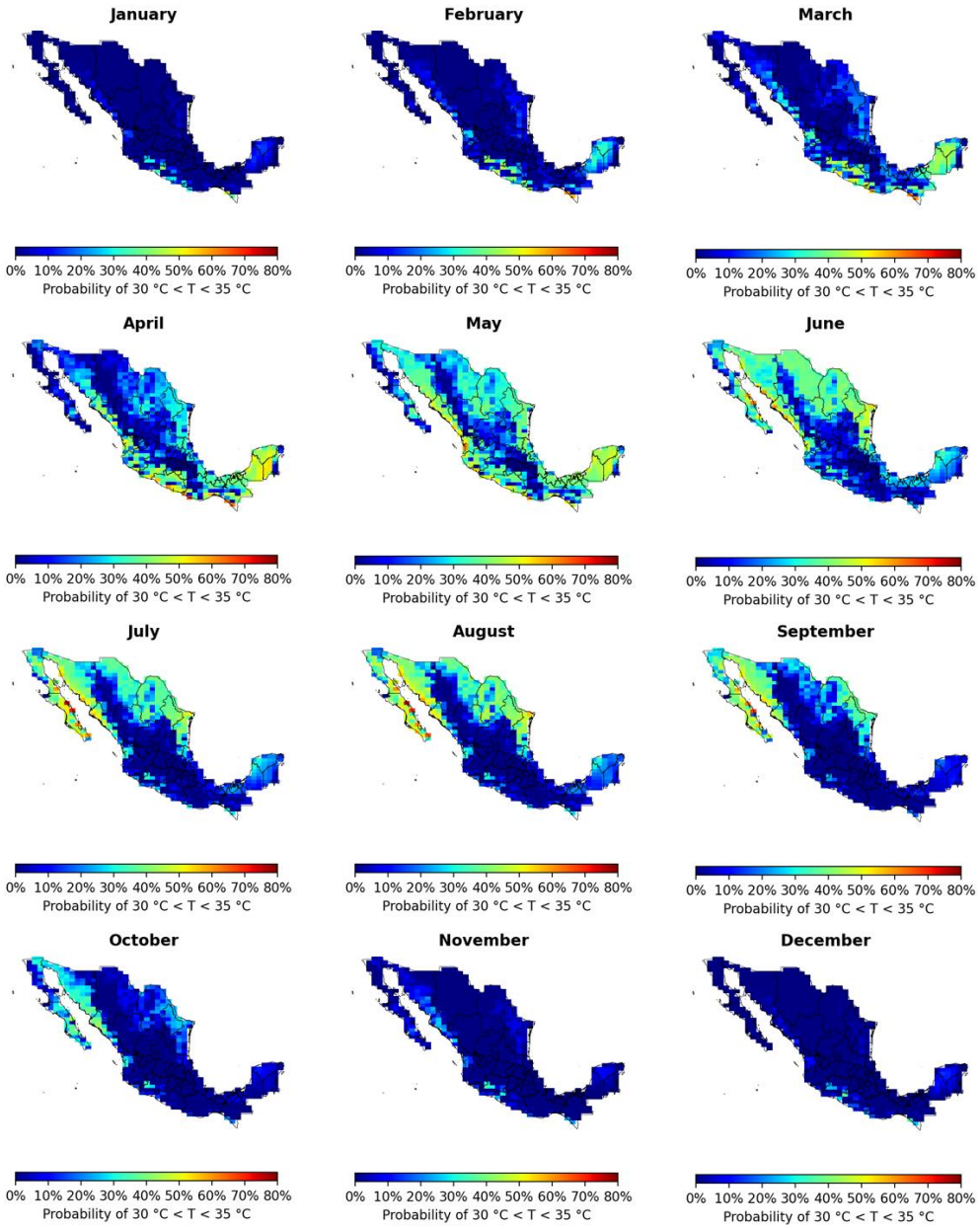

Figure 7. Probability map of ambient temperatures between 30 °C and 35 °C per month in Mexico.

This information is enough to compute the mean daily PV energy for a given PV module, which is performed in the following subsection.

3.3. PV Energy Calculation

In this subsection, the JAM78S30 580/MR PV module from JA Solar was chosen to compute the generated PV energy per square meter. The electrical specifications needed for the computation can be obtained from the free specifications sheet [48] and are shown in Table 3. The operating conditions correspond to a NOCT of 45 °C at an irradiance of 800 W/m^2, an ambient temperature of 20 °C, wind speed of 1 m/s, and *AM*1.5 G.

Table 3. Some of the main electrical parameters of the JAM78S30 580/MR PV module from JA Solar, with a surface area of 2.64 m^2, at an STC corresponding to an irradiance of 1000 W/m^2, cell temperature of 25 °C, and *AM*1.5 G.

Electrical Parameters at STC	Value
Rated Maximum Power (W)	580
Open Circuit Voltage (V)	53.11
Maximum Power Voltage (V)	44.35
Short Circuit Current (A)	13.84
Module Efficiency (%)	20.7
Temperature Coefficient	−0.350%/°C

First, the daily photovoltaic energy per square meter was computed using Equation (13) for the 365 days of 20 years for all grid cells. Then, the mean daily PV energy in each month of the year was calculated using Equation (14), and the results are represented in the map of the country by color levels ranging from 0.6 kWh/m^2 to 1.6 kWh/m^2, corresponding to purple and yellow, respectively. The 12 maps are shown in Figure 8. Lighter-colored regions indicate higher energy, and darker ones indicate lower energy. It is important to remark that the results shown in Figure 5 are not enough to draw conclusions about site deployment, since, as shown in Figure 8, the regions with the highest radiation do not present the highest photovoltaic energy generation.

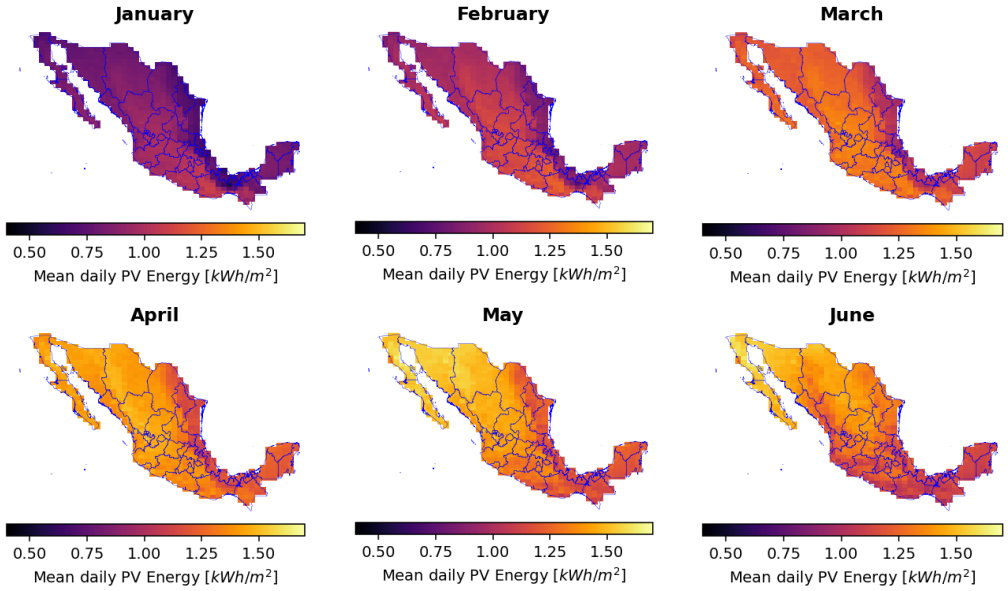

Figure 8. *Cont.*

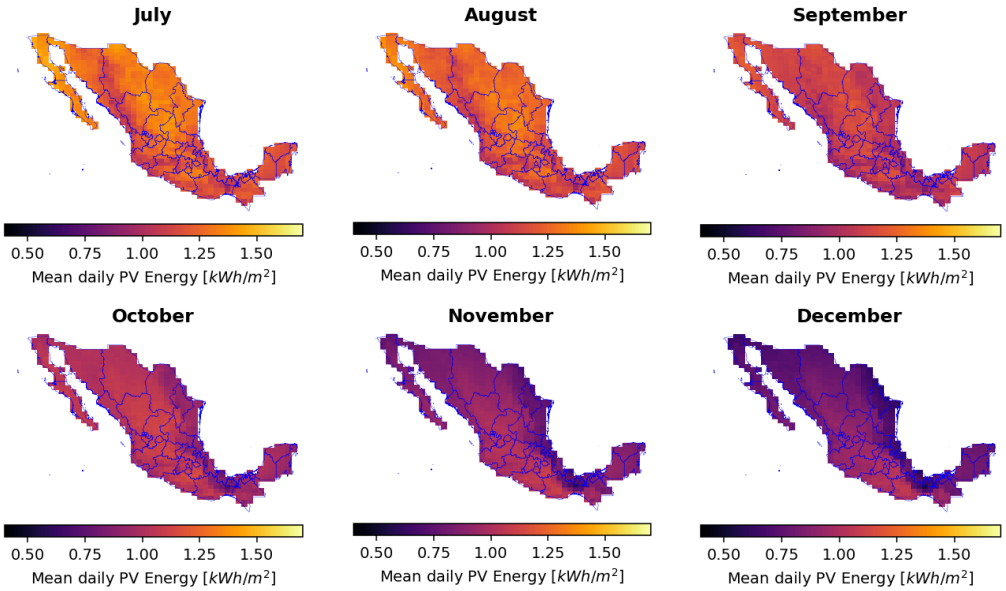

Figure 8. Maps of Mexico for the mean daily PV energy per square meter for each month of the year.

The last step to assess PV deployment is the regional forecasting of PV energy production.

3.4. Forecasting PV Energy

As mentioned before, due to the intermittency of solar resources, a probabilistic approach is more adequate for forecasting PV energy production. Following the methodology shown in Section 2.5, it is assumed that the annual production of the PV module obeys a GD. The mean and the standard deviation of the GD for each grid cell are computed using Equations (17) and (18), as well as hourly radiation and temperature information from all available data. This information allows obtaining the $P50$–$P95$ estimations of PV energy production for each cell grid, as defined in Equation (16). The information is graphically presented in atlases of the $P50$–$P95$ estimations for the PV energy production of the country.

The forecast of the annual PV energy produced in Mexico is shown in the atlases for the $P50$–$P95$ estimation in Figure 9. The dark colors correspond to regions where the least PV energy can be generated, with a minimum of 300 kWh/m^2, and the light colors correspond to the highest energies, up to 450 kWh/m^2. The atlas for the $P50$ estimation corresponds to the left map, and the atlas for the $P95$ estimation corresponds to the right map. The blue dots in the $P95$ atlas correspond to the location of real PV power plants used for the validation of these results in the following subsection. As expected, both maps follow the same pattern. However, the $P50$ atlas shows regions with a 50% probability of PV energy production of at least the energy corresponding to the color. Moreover, the $P95$ atlas shows regions with a 95% probability of obtaining a PV energy production corresponding to at least the energy represented by the colors. Thus, in Mexico, it is 50% likely that at least 350 kWh/m^2 and at most 450 kWh/m^2 will be produced in a year. Moreover, it is 95% probable that at least 300 kWh/m^2 will be produced. Additionally, it is remarkable to note that the best site for PV energy production corresponds to the region close to Puebla, in the center of Mexico. This information is very important for financial investment risk analysis.

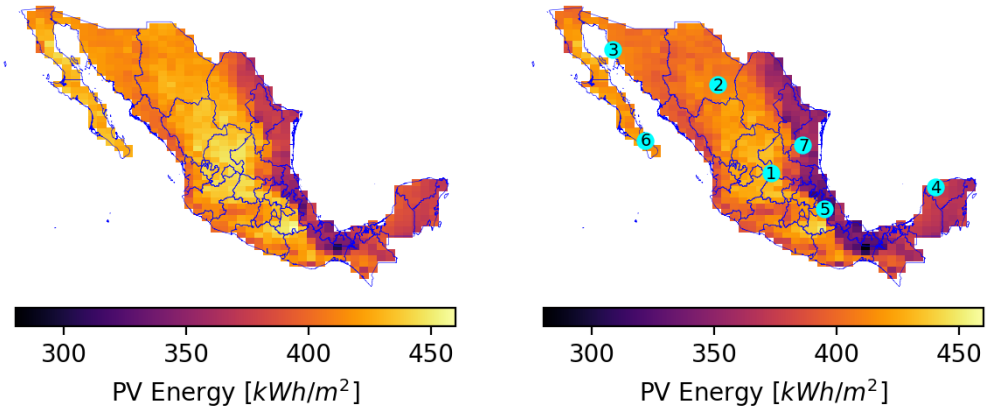

Figure 9. Atlases for the *P*50–*P*95 estimations of the annual PV energy per square meter in Mexico. The blue dots correspond to the location of the PV plants used for validation in Section 3.5.

3.5. Validation

*P*50 estimations of the annual PV energy are more adequate when the PV system operation conditions are normal throughout the year. However, weather conditions can bring unexpected events that may decrease energy production. Thus, the validation of the results is conducted for *P*95 forecasting results as follows. Seven PV solar plants throughout Mexico, the location of which is depicted by the blue points in Figure 9, were selected based on the public information on power generation available online. Table 4 shows the findings. The first column corresponds to the name of the PV power plant and its location. The second column corresponds to the *P*95 forecasts for the produced yearly energy, whereas the third column corresponds to the yearly PV energy reported on the websites of the PV power plants. Finally, the fourth column corresponds to the discrepancy between the reported yearly and the forecasted *P*95 energies.

Table 4. Validation of the results for *P*95 forecasting by comparing PV forecasted energy with the real generation of some PV power plants throughout Mexico.

PV Plant	P95 Estimations for Forecasted Yearly Energy (kWh/m^2)	Reported Yearly Energy (kWh/m^2)	Discrepancy (%)
1. Villa de Arriaga, San Luis Potosí	427.46	348	22
2. Ciudad Camargo, Chihuahua	409.15	367	10
3. Puerto Libertad, Sonora	415.03	393	5
4. San Ignacio, Yucatán	377.71	352	6
5. Cuyoaco, Puebla	422.26	382	10
6. Aura Solar III, Baja California	415.86	353	17
7. Parque Bicentenario, Tamaulipas	367.71	318	13

As shown in Table 4, the discrepancy between the forecasted *P*95 values and the reported PV energy produced for the selected PV power plants ranges between 6% and 22%. There are several reasons for this discrepancy: The atlases for the *P*50–*P*95 estimations for PV energy production were created using a specific PV module, described at the beginning of Section 3.3, with an efficiency of 20.7% at STC. However, the PV technology used in the solar power plants of this section is unknown; thus, the efficiency of the panels is also unknown. Typical values of commercial PV panel efficiency are around 17%. Moreover, it is also important to know the surface of the PV panels, but this information is missing. Furthermore, the analysis considers heat losses due to temperature, but further losses have not been considered so far. The main losses that a PV system faces are due to inaccuracy and variability in the meteorological data, which can reach up to 3%. Other losses are mainly due to shadings, incidence angle modifier (IAM), dirt on the PV modules, module

and string mismatches, wiring ohmic loss, inverter efficiency, and the degradation of the modules. These losses can reach more than 10%.

Additionally, this analysis considers a PV system operating full-time. The plant capacity factor, CF, measures the real production of a plant considering the operating time, and is calculated as follows:

$$CF = \frac{\text{Actual Energy Generated}}{\text{Theoretical Energy Generated operating at full time}}. \quad (19)$$

Thus, the $P50$–$P95$ estimations are for $CF = 1$; however, the CF for the PV plants reported in Table 4 is unknown.

4. Discussion

The proposed methodology can be used free of charge to explore vast regions almost anywhere in the world, for which there is coverage by free-access databases of solar radiation and temperature, providing fundamental information when making decisions about where to build photovoltaic installations, especially PV plants. A remarkable benefit is that the resulting information is provided on a regional basis, not local, which allows for the rapid exploration of large areas for viability in the long term.

This methodology can be scaled up or down to accommodate the area to be investigated for PV viability. In general, a large area is divided into a grid of smaller areas or land cells. The smallest size of the land cells depends on the spatial resolution of the databases; squared areas as small as 2 km per side are achievable. Nevertheless, with small land cells, the amount of radiation and temperature data required can grow enormously, increasing the retrieval time and the processing time, as well as the number of land cells that need to be graphically depicted on the maps. For each land cell, hourly data are required to calculate meaningful statistics per day, per month, and per year, for as many years as possible. Data clustering from machine learning has proved to be a great tool, allowing sense to be made of problems that may easily have millions of raw data points to analyze.

Even though several works have studied solar resources in regions, as shown in Section 1.1, most of them neglect the effect of ambient temperature on PV energy production. In this regard, the proposed methodology, in the first step, uses GHI data as the main factor to determine the incoming energy to the PV panels, and in the second step, it accounts for the effect of temperature on the efficiency of the PV panels to compute the PV energy outcome. The results show that high ambient temperature may significantly decrease PV energy production, contrary to expectations. Hence, the maps that statistically show the combined effect of solar radiation and the probability of reaching high temperatures throughout the year allow the sites and times of the year for which PV energy production is better to be identified. In general, it is found that regions with high radiation and low temperatures provide the highest energy yield and can be considered viable.

The simplest approach for forecasting PV energy production is deterministic and provides a number based on the mean values of the relevant variables. However, a probabilistic approach is far more convenient, based on statistical levels of confidence which state that the PV energy production will exceed a specific value with a given probability, assuming that the PV energy yield can be described with a Gaussian probability distribution. The $P50$ and $P95$ estimations are PV energy values such that PV energy generation will exceed them with 50% and 95% probability, respectively. The construction of the $P50$ and $P95$ annual PV energy forecast atlases provides a powerful fast-track procedure to find regions of interest that are viable for the installation of PV plants. Of course, local radiation and temperature measurements must be obtained, as well as the specific electrical parameters PV panel operations, if necessary, to fine-tune the results and achieve a better approximation of the PV energy yield and the attainable profits.

This methodology provides the $P50$ and $P95$ forecasts of annual PV energy. To obtain a better appreciation of the validity of these results, a comparison is drawn against the yearly production of several PV plants scattered throughout Mexico. Discrepancies between the

*P*95 forecasts and the yearly production vary from 5% to 22%. These discrepancies can easily be explained by the conditions not considered by the proposed fast-track methodology, as mentioned in Section 3.5, that is, a discrepancy of 3.7% due to lower PV panel efficiencies, 3% because of technical losses due to the physical installation, and a very conservative 5% decrease due to the actual PV plant capacity factor. These percentages may easily explain a discrepancy of approximately 11.7%, which is close to the mean of the observed discrepancies and demonstrates the validity of the proposed methodology.

As a direction for future work, the development of a similar study considering smaller land areas will be undertaken, focused on a community isolated from the power grid and incorporating GIS tools to locate a PV plant with greater precision. Another future direction is to complement this work with long-term PV energy forecasting, with the results of previous research about cloud kinetics forecasting using ANN [49] to compensate for solar intermittency in the short term. Finally, it will also be interesting to continue the calculations per cluster once the clusters are formed based on the solar radiation levels, and not per land cell, as in this work. It is expected that the PV energy distribution will better approximate a Gaussian distribution.

5. Conclusions

This paper introduces a fast-track methodology to carry out the long-term assessment of solar resources and forecasting of PV energy to uncover viable regions for PV energy generation, as is required by the initial feasibility analysis to locate potential sites for building PV plants. After applying this methodology, the results in the form of PV energy yield forecasts will be very helpful for risk analysis to ensure the success of the potential PV plants.

The proposed methodology is tested by applying it to the whole country of Mexico as a case study. The area of interest is approximately 2 million km^2, with a large diversity of topographical and meteorological conditions, besides the good solar radiation levels throughout the country. These characteristics provide many combinations, the viability of which for PV energy generation needs to be evaluated from a set of more than 100 million hourly radiation and temperature data points, spanning a period of 21 years for 731 land cells of approximately 50 km per side. It should be noted that the amount and the size of the land cells can be scaled down or up depending on the availability and resolution of the databases. Additionally, the graphical approach allows a very intuitive appreciation of the best regions to place PV plants, of course, relying on the numerical data as a backup.

It is demonstrated that the proposed methodology is a fast, cost-effective, and reliable way to address issues in the assessment of solar resources and the forecasting of PV energy yield, which is required for the placement of PV plants with a guaranteed probability of success. This methodology can be applied to any area of interest in the world, having arbitrary size and topography, since free-access solar radiation databases currently include information for any point for which the longitude and latitude lie on land. Additionally, any PV panel technology can be considered to forecast the PV energy yield, assuming that the efficiency and temperature variation are known.

Finally, it is noteworthy that the places with the highest radiation levels do not necessarily have the best PV energy yields, as could be expected. Hence, methodologies such as the one presented in this paper must be utilized to obtain useful figures of merit for the assessment of solar resources and the forecasting of PV energy yield.

Supplementary Materials: The following supporting information can be downloaded at: https://github.com/FelosRG/Photovoltaic-Energy-Mexico accessed on 15 November 2022.

Author Contributions: Conceptualization, M.B., A.R., R.G. and O.A.J.; methodology, M.B. and R.G.; software, A.R., G.R. and S.H.; validation, A.R.; formal analysis, M.B. and R.G.; investigation, M.B., R.G. and O.A.J.; resources, G.R. and S.H.; data curation, A.R.; writing—original draft preparation, M.B., R.G. and O.A.J.; writing—review and editing, M.B.; visualization, M.B., A.R., R.G., G.R., S.H.

and O.A.J.; supervision, M.B.; project administration, M.B.; funding acquisition, M.B., G.R. and S.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Programa Espacial Universitario (PEU) from Universidad Nacional Autónoma de México (UNAM).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Meteorological datasets are freely available at the NSRDB.

Acknowledgments: This work is part of the project "Predicción del Recurso Solar usando imágenes satelitales para impulsar el desarrollo sostenible en comunidades aisladas con energía asequible y no contaminante" approved in the Programa Espacial Universitario (PEU) from Universidad Nacional Autónoma de México (UNAM). The authors thank the PEU program for their support in the publication of this work. Monica Borunda also thanks CONACYT for her Catedra Research Position with ID 71557, and CENIDET for its hospitality.

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

References

1. International Energy Agency. *Renewables 2021 Analysis and Forecast to 2026*; International Energy Agency: Paris, France, 2021; pp. 26–29.
2. Al Garni, H.Z.; Awasthi, A. Solar PV Power Plants Site Selection: A Review. In *Advances in Renewable Energies and Power Technologies*; Yahyaoui, I., Ed.; Elsevier: Amsterdam, The Netherlands, 2018; Volume 1, pp. 57–75.
3. Brewer, J.; Ames, D.P.; Solan, D.; Lee, R.; Carlisle, J. Using GIS analytics and social preference data to evaluate utility-scale solar power site suitability. *Renew. Energy* **2015**, *81*, 825–836. [CrossRef]
4. Kereush, D.; Perovych, I. Determining criteria for optimal site selection for solar power plants. *Geomat. Land Manag. Landsc.* **2017**, *4*, 39–54. [CrossRef]
5. World Meteorological Organization. Measurement of Radiation. In *Guide to Instruments and Methods of Observation*, 2020 ed.; WMO: Geneva, Switzerland, 2020; Volume I—Measurement of Meteorological Variables, pp. 255–307.
6. Zhang, X.; Liang, S. Solar Radiation. In *Advanced Remote Sensing Terrestrial Information Extraction and Applications*, 2nd ed.; Liang, S., Wang, J., Eds.; Academic Press: Oxford, UK, 2020; pp. 157–191.
7. Schulz, B.; El-Ayari, M.; Lerch, S.; Baran, S. Post-processing numerical weather prediction ensembles for probabilistic solar irradiance forecasting. *Sol. Energy* **2021**, *220*, 1016–1031. [CrossRef]
8. Kostilev, V.; Pavlosky, A. Solar power forecasting performance—Towards industry standards. In *Environmental Science Engineering*; Energynautics GmbH Mühlstraße: Langen, Germany, 2011; pp. 1–8.
9. Charabi, Y.; Gastli, A.; Al-Yahyai, S. Production of solar radiation bankable datasets from high-resolution solar irradiance derived with dynamical downscaling Numerical Weather Prediction model. *Energy Rep.* **2016**, *2*, 67–73. [CrossRef]
10. Sengupta, M.; Xie, Y.; Habte, A.; Buster, G.; Maclaurin, G.; Edwards, P.; Sky, H.; Bannister, M.; Rosenlieb, E. *The National Solar Radiation Database (NSRDB) Final Report: Fiscal Years 2019–2021*; National Renewable Energy Laboratory: Golden, CA, USA, 2022.
11. Khare, V.; Bunglowala, A. Solar-Wind Energy Assessment by Big Data Analysis. In *Innovation in Energy Systems—New Technologies for Changing Paradigms*; Ustun, T.S., Ed.; IntechOpen: London, UK, 2019; pp. 1–23.
12. Voyant, C.; Notton, G.; Kalogirou, S.; Nivet, M.L. Machine learning methods for solar radiation forecasting: A review. *Renew. Energy* **2017**, *105*, 569–582. [CrossRef]
13. Akhter, M.N.; Mekhilef, S.; Mokhlis, H.; Shah, N.M. Review on forecasting of photovoltaic power generation based on machine learning and metaheuristic techniques. *IET Renew. Power Gener.* **2019**, *13*, 1009–1023. [CrossRef]
14. Kumar, D.S.; Yagli1, G.M.; Kashyap, M.; Srinivasan, D. Solar irradiance resource and forecasting: A comprehensive review. *IET Renew. Power Gener.* **2020**, *14*, 1641–1656. [CrossRef]
15. Wang, L.; Shi, J.A. Comprehensive Application of Machine Learning Techniques for Short-Term Solar Radiation Prediction. *Appl. Sci.* **2021**, *11*, 5808. [CrossRef]
16. Carneiro, T.C.; de Carvalho, P.C.M.; Santos, H.A.d.; Lima, M.A.F.B.; Barga, A.P.d. Review on Photovoltaic Power and Solar Resource Forecasting: Current Status and Trends. *J. Sol. Energy Eng.* **2022**, *144*, 010801. [CrossRef]
17. Singla, P.; Duhan, M.; Saroha, S. A comprehensive review and analysis of solar forecasting techniques. *Front. Energy* **2022**, *16*, 187–223. [CrossRef]
18. Sudharshan, K.; Naveen, C.; Vishnuram, P.; Krishna Rao Kasagani, D.V.S.; Nastasi, B. Systematic Review on Impact of Different Irradiance Forecasting Techniques for Solar. Energy Prediction. *Energies* **2022**, *15*, 6267. [CrossRef]
19. Lin, F.; Zhang, Y.; Wang, J. Recent advances in intra-hour solar forecasting: A review of ground-based sky image methods. *Int. J. Forecast.* 2020, in press. [CrossRef]

20. Zhou, Y.; Liu, Y.; Wang, D.; Liu, X.; Wang, Y. A review of global solar radiation prediction with machine learning models in a comprehensive perspective. *Energy Convers. Manag.* **2021**, *235*, 113960. [CrossRef]
21. Bouveyrona, C.; Brunet-Saumardb, C. Model-based clustering of high-dimensional data: A review. *Comput. Stat. Data Anal.* **2014**, *71*, 52–78. [CrossRef]
22. Zagouras, A.; Kazantzidis, A.; Nikitidou, E.; Argiriou, A.A. Determination of measuring sites for solar irradiance, based on cluster analysis of satellite-derived cloud estimations. *Sol. Energy* **2013**, *97*, 1–11. [CrossRef]
23. Polo, J.; Gastón, M.; Vinde, J.M.; Pagola, I. Spatial variability and clustering of global solar irradiation in Vietnam from sunshine duration measurements. *Renew. Sustain. Energy Rev.* **2015**, *42*, 1326–1334. [CrossRef]
24. Jiménez-Pérez, P.F.; Mora-López, L. Modeling and forecasting hourly global solar radiation using clustering and classification techniques. *Sol. Energy* **2016**, *135*, 682–691. [CrossRef]
25. Govender, P.; Brooks, M.J.; Matthews, A.P. Cluster analysis for classification and forecasting of solar irradiance in Durban, South Africa. *J. Energy South. Afr.* **2018**, *29*, 51–62. [CrossRef]
26. Rodríguez-Benítez, F.J.; Arbizu-Barrena, C.; Santos-Alamillos, F.J.; Tovar-Pescador, J.; Pozo-Vázquez, D. Analysis of the intra-day solar resource variability in the Iberian Peninsula. *Solar Energy* **2018**, *171*, 374–387. [CrossRef]
27. Lopes-de-Lima, F.J.; Martinsa, F.R.; Santos-Costab, R.; Rodrigues-Gonçalvesb, A.; Paes-dos-Santos, A.P.; Bueno-Pereira, E. The seasonal variability and trends for the surface solar irradiation in the northeastern region of Brazil. *Sustain. Energy Technol. Assessments* **2019**, *35*, 335–346. [CrossRef]
28. Theocharides, S.; Makrides, G.; Livera, A.; Theristis, M.; Kaimakis, P.; Georghiou, G.E. Day-ahead photovoltaic power production forecasting methodology based on machine learning and statistical post-processing. *Appl. Energy* **2020**, *268*, 115023. [CrossRef]
29. Jayalakshmi, N.Y.; Shankar, R.; Subramaniam, U.; Baranilingesan, I.; Karthick, A.; Stalin, B.; Rahim, R.; Ghosh, A. Novel Multi-Time Scale Deep Learning Algorithm for Solar Irradiance Forecasting. *Energies* **2021**, *14*, 2404. [CrossRef]
30. Behr, H.D.; Jung, C.; Trentmann, J.; Schindler, D. Using satellite data for assessing spatiotemporal variability and complementarity of solar resources—A case study from Germany. *Meteorol. Z.* **2021**, *30*, 515–532. [CrossRef]
31. Nga, P.T.T.; Ha, P.T.; Hang, V.T. Satellite-Based Regionalization of Solar Irradiation in Vietnam by k-Means Clustering. *J. Appl. Meteorol. Climatol.* **2021**, *60*, 391–402.
32. Watanabe, T.; Oka, K.; Hijioka, Y. Assessment of characteristics of surface solar irradiance on consecutive days using a self-organizing map and clustering methods. *Meteorol. Appl.* **2021**, *28*, 1. [CrossRef]
33. Borunda, M.; Ramirez, A.; Liprandi, N.; Rodríguez, M.; Sánchez, A. Seasonality Atlas of Solar Radiation in Mexico. In *Advances in Computational Intelligence, MICAI 2021. Lecture Notes in Computer Science*; Batyrshin, I., Gelbukh, A., Sidorov, G., Eds.; Springer: Cham, Switzerland, 2021; p. 13067.
34. Ali-Ou-Salah, H.; Oukarfi, B.; Mouhaydine, T. Short-term solar radiation forecasting using a new seasonal clustering technique and artificial neural network. *Int. J. Green Energy* **2022**, *19*, 424–434. [CrossRef]
35. Maldonado-Salguero, P.; Bueso-Sánchez, M.C.; Molina-García, Á.; Sánchez-Lozano, J.M. Spatio-temporal dynamic clustering modeling for solar irradiance resource assessment. *Renew. Energy* **2022**, *200*, 344–359.
36. Salinas-González, J.D.; García-Hernández, A.; Riveros-Rosas, D.; Moreno-Chávez, G.; Zarzalejo, L.F.; Alonso-Montesinos, J.; Galván-Tejada, C.E.; Mauricio-González, A.; González-Cabrera, A.E. Multivariate Analysis for Solar Resource Assessment Using Unsupervised Learning on Images from the GOES-13 Satellite. *Remote Sens.* **2022**, *14*, 2203. [CrossRef]
37. Fotis, G.; Dikeakos, C.; Zafeiropoulos, E.; Pappas, S.; Vita, V. Scalability and Replicability for Smart Grid Innovation Projects and the Improvement of Renewable Energy Sources Exploitation: The FLEXITRANSTORE Case. *Energies* **2022**, *15*, 4519. [CrossRef]
38. Sijakovic, N.; Terzic, A.; Fotis, G.; Mentis, I.; Zafeiropoulou, M.; Maris, T.I.; Zoulias, E.; Elias, C.; Ristic, V.; Vita, V. Active System Management Approach for Flexibility Services to the Greek Transmission and Distribution System. *Energies* **2022**, *15*, 6134. [CrossRef]
39. National Solar Radiation Database, NREL. Available online: https://nsrdb.nrel.gov/ (accessed on 22 July 2022).
40. Dubey, S.; Sarvaiga, J.N.; Seshadri, B. Temperature Dependent Photovoltaic (PV) Efficiency and Its Effect on PV Production in the World—A Review. *Energy Procedia* **2013**, *33*, 311–321. [CrossRef]
41. Green, M.A.; Hishikawa, Y.; Dunlop, E.D.; Levi, H.D.; Hohl-Ebinger, J.; Ho-Baillie; Anita, W.Y. Solar cell efficiency tables (version 51). *Prog. Photovolt. Res. Appl.* **2017**, *26*, 3–12. [CrossRef]
42. Clean Energy Reviews. Available online: https://www.cleanenergyreviews.info/blog/most-efficient-solar-panels (accessed on 22 July 2022).
43. Pelleg, D.; Moore, A. Accelerating exact k-means algorithms with geometric reasoning. In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD'99, California, CA, USA, 15–18 August 1999.
44. Rousseeuw, P.J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
45. Einstein, A. Concerning an Heuristic Point of View toward the Emission and Transformation of Light. *Annalen der Physik* **1905**, *17*, 132–148. [CrossRef]
46. Lorenzo, E. *Electricidad Solar Fotovoltaica. Vol. 3, Ingeniería fotovoltaica. Mairena de Aljarafe (Sevilla)*; PROGENSA: Sevilla, España, 2014.
47. Available online: https://developer.nrel.gov/docs/solar/nsrdb/psm3-download/ (accessed on 22 October 2022).

48. Available online: https://www.jasolar.com/uploadfile/2022/0513/20220513051007792.pdf (accessed on 2 November 2022).
49. Borunda, M.; Ramírez, A.; Garduno, R.; Ruiz, G.; Hernandez, S.; Jaramillo, O.A. Convolutional and Dense ANN for Cloud Kinetics Forecasting Using Satellite Images. In *Advances in Computational Intelligence*; MICAI 2022 Lecture Notes in Computer Science; Pichardo Lagunas, O., Martínez-Miranda, J., Martínez Seis, B., Eds.; Springer: Cham, Switzerland, 2022; Volume 13612, pp. 212–224.