A machine learning approach to PV-climate classification

de las Heras, Francisco Javier Triana; Isabella, Olindo; Vogt, Malte Ruben

DOI
10.1016/j.renene.2025.123685

Publication date
2026

Document Version
Final published version

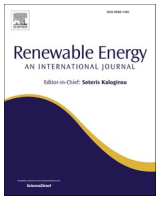
Published in
Renewable Energy

Renewable Energy 256 (2026) 123685

Contents lists available at ScienceDirect

Renewable Energy

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



A machine learning approach to PV-climate classification

Francisco Javier Triana de las Heras [1], Olindo Isabella, Malte Ruben Vogt [*]

Photovoltaic Materials and Devices, Delft University of Technology, Mekelweg 4, 2628 CD Delft, the Netherlands

ARTICLE INFO

Keywords:
Photovoltaic
Solar energy
Machine learning (ML)
Performance
Climate zones
Climate classification

ABSTRACT

Photovoltaic (PV) system performance is linked to climatic conditions in which the system operates. This leads to the Köppen-Geiger-Photovoltaic (KGPV) climate classification. KGPV is created by overlaying four irradiation levels with the commonly used Köppen-Geiger climate zones. Potential drawbacks of this approach are that the climate features are not considered in a combined manner in the sorting process and that the KGPV zones inherent a dependence on precipitation. We propose a machine-learning approach to address this deficiencies and improve PV climate classification. First, supervised learning is used to evaluate the correlation between climate features and a PV system's specific energy yield. We find that the inclusion of the darkest and brightest irradiation months as well as UV irradiation improves accuracy, while wind speed, relative humidity, precipitation and annual mean daily temperature difference have little impact on accuracy. Subsequently, k-means clustering combined with comprehensive qualitative analysis, identifies a PV classification based on seven climate features and 21 clusters. A mountainous climate characterized by moderate to low temperature and high irradiation is uncovered compared to KGPV. Moreover, this new PV climate classification reduces the sum of squared errors by 58 % compared to KGPV clearly signifying a more accurate PV climate classification approach.

1. Introduction

Global solar photovoltaic (PV) energy surpassed 1 TW cumulative capacity in 2022 [1]. While it took 68 years to install the first terawatt, the latest numbers of 0.346 TW added in 2023 [2] indicate that the second terawatt of PV capacity will be achieved in 2025 taking no more than three years. PV growth during the last decade has been impressive with current capacity being more than 21 times greater than in 2010. This has been accompanied by an 88 percent reduction in the global weighted average Levelized Cost of Energy (LCOE) of utility-scale photovoltaics [3,4]. Solar PV has shown the highest learning rates of all renewable energy technologies, becoming the lowest-cost option for new electricity generation in most of the world. Even though it already contributed to more than 5.5 percent of global electricity generation in 2023 [5,6], worldwide deployment of PV is still low compared to future projections. Technological advances, cost reduction, depletion of fossil fuels, environmental concerns, and growing energy demand are expanding PV in all climate zones.

Machine Learning (ML) is a unique technique, which has been used increasingly in many fields in the past years [7]. A rising number of PV related studies also utilize machine learning [8,9]. These studies commonly aim to forecast PV power production [10,11], but other goals such as for example material discovery for PV [12,13] or PV reliability and fault detection [14,15] are also increasing. Clustering methods have been utilized in PV power forecast to group weather conditions [16,17] or to group China in terms of PV tilt angle [18] or reliability assessment [19,20].

The Köppen-Geiger (KG) climate classification is the most well-known way to sort the earth's climates into different climate zones [21,22]. Consequently, the original KG classification has been modified differently by several authors, resulting in a variety of classifications without standardisation and difficulties to compare results [23]. In recent years machine learning techniques have also been used to create new climate classification [24,25], which in some cases had specific applications such as for example vegetation [26,27], streamflow [28] or buildings [29].

Climate classification has also been applied to the PV technology. However, machine learning has not yet been utilized to find a global PV climate classification. Among the classical approaches to PV climate classification Dash et al. [30] divided India into 6 climatic zones and studied their most efficient PV technologies. The International Energy Agency (IEA) PV Power Systems (PVPS) Programme Task 13 [31]

* Corresponding author.
E-mail address: m.r.vogt@tudelft.nl (M.R. Vogt).
[1] Now with: Chalmers University of Technology, 412 96 Gothenburg, Sweden.

https://doi.org/10.1016/j.renene.2025.123685
Received 27 July 2024; Received in revised form 15 May 2025; Accepted 6 June 2025
Available online 13 June 2025

presented guidance for customized operations and maintenance (O&M) service in seven different climate zones. Micheli et al. [21] studied the impact of performance losses due to soiling in climate zones. The KG classification is solely based on temperature and precipitation [22], resulting in an unsatisfactory scheme for conducting a comprehensive analysis of PV performance as irradiation is not considered. Therefore, several authors aimed to supplement KG by including other relevant parameters. Skandalos et al. [32] analysed the effect of local climatic conditions on photovoltaic building integration for some global locations, concluding that the optimised design depends on the climate zone. Karin et al. [33] developed a climate classification to identify which types of degradation may be expected in different geographic areas in the USA.

In 2019, Ascencio-Vásquez et al. [23,34] went a step further, developing a new worldwide classification: the Köppen-Geiger-Photovoltaic (KGPV) climate classification. This new classification expands KG by considering solar irradiation. It follows two criteria. First, based on the KG scheme, zones are classified in terms of temperature and precipitation, differentiating among Tropical (A), Desert (B), Steppe (C), Temperate (D), Cold (E), and Polar (F). Secondly, solar irradiation is considered to distinguish among Very High irradiation zones (K), High irradiation zones (H), Medium irradiation zones (M), and Low irradiation zones (L). By including the Global Horizontal Irradiation, KGPV is a significant improvement for PV applications. Among the 24 possible combinations, KGPV selected 12 to classify the world, neglecting the rest of them due to reasoning not based on climate parameters such as a poor land-surface ratio or population density. One aspect which needs further research is that the climate parameters are not considered in a combined manner in the sorting process as temperature and precipitation are derived from KG and irradiation is layered over them later in an independent process. Moreover, Ascencio-Vásquez et al. concluded, additional parameters such as wind speed, relative humidity or UV irradiation should be examined [23]. Like its PV climate classification predecessors, KGPV is also completely based on classical statistical analysis.

In this work, this research gap is closed by introducing a machine-learning-based (ML-based) approach to develop a worldwide climate classification directly applicable to PV. First, supervised learning serves to identify and assess the climate variables for correlation with a PV System's specific energy yield to identify possible additional climate parameters. Subsequently, the selected variables are used to create the classification using the k-means clustering algorithm. In contrast to KGPV the climate variables are considered in a combined manner during the classification process. As secondary objective additional climate variables such as wind speed, relative humidity, UV irradiation and also the irradiation in the brightest and darkest months are also considered in the feature selection process to determine their relevance for PV climate classification.

The paper is structured as follows. First, data collection and preprocessing are illustrated in section 2. In section 3, the first step, the use of machine learning for selection and weighing of climate variables, is described. In section 4, the second step is presented. It comprises the methodology for creating the climate zones, and the analysis of the classification obtained. The results are discussed and compared to KGPV in section 5. Finally, the main conclusions are summarised in section 6.

2. PV performance and climate data collection methodology

For developing the classification, data on climate and specific energy yield is required. This is the key to relating climate to PV performance. Since the objective is to develop a worldwide classification, an extensive and accurate dataset is essential. In this project, a worldwide grid with resolution 0.5° latitude by 0.5° longitude is utilized. The climate data is extracted from renowned climate research centres and institutions for the period 1991 to 2021 [35–38]. Specific energy yield values are used as provided by Ascencio-Vásquez et al. in Ref. [23].

2.1. Specific energy yield

PV performance comprises several factors. In particular, the International Electrotechnical Commission (IEC) 61724 standard defines the following principal PV system performance indices: energy generated by PV systems (E_{ac}), reference yield (Y_r), final or specific energy yield (Y_f), performance ratio (PR), capacity utilization factor (CUF), and PV system efficiency (η_{sys}) [39]. Among these parameters, Y_f is selected as the basis for the classification due to its relevance and comprehensibility. It enables direct comparison between systems with different capacities, as opposed to the E_{ac}. The final yield, or specific energy yield, is defined as the net daily, monthly, or annual electrical energy output of the PV plant divided by its rated power. It is given by the following expression [39]:

$$Y_f = \frac{Energy\ generated\ (E_{ac})}{Rated\ power\ of\ PV\ plant\ (P_{STC})} \quad (1)$$

It is measured in hours or, equivalently, kWh/kW$_p$. In this work, Y_f always refers to an annual basis.

To enable direct comparison with KGPV without introducing a potential bias of using different specific energy yield data the theoretical worldwide specific energy yield values calculated by Ascencio-Vásquez et al. to assess the KGPV climate classification [23] are also used in this study. It is based on crystalline silicon (c-Si) modules, which account for 95 % of the market. They simulated a typical day for each month, multiplied by the number of days in each month, and summed up to the annual value. The impact of temperature, balance-of-system efficiency, and spectral and angular-reflection losses were considered. On the other hand, shading, soiling, and snow losses were neglected. No quantification, of the energy yield uncertainty was given in Ref. [23].

2.2. Climate data collection and preprocessing

The climate dataset built here comprises all climate variables to be tested for their significance in developing a PV climate classification. This dataset serves as the baseline for the subsequent feature selection procedure. The final data set used for feature selection is available in the supplementary files.

It consists of a matrix whereby each row corresponds to a particular location (sample) and each column contains the value of a climate variable (feature) for that location, except for the first two columns which contain the latitude and longitude, respectively. The climate features are selected based on technical expertise. Moreover, KG and KGPV classifications are used as references. As indicated by Ascencio-Vasquez et al. in KGPV, the evolution of temperature has been remarkably different since 1990 [23]. To enhance comparability and reproducibility, for every climate feature, monthly average data is extracted from 1991 to 2021, and averaged, in turn for each month of the year, to obtain a typical average year.

Table 1 summarises the 12 final features included in the dataset. These are derived from the raw climate variables shown in Table 2. This

Table 1
Climate features conforming the climate dataset. The final data set used for feature selection is available in the supplementary files.

Feature	Description
T_{ann}	Annual mean near-surface (2 m) temperature (°C)
T_{max}	Monthly mean temperature of the warmest month (°C)
T_{min}	Monthly mean temperature of the coldest month (°C)
DTR_{ann}	Annual mean daily temperature difference (°C)
P_{ann}	Accumulated annual precipitation (mm)
P_{min}	Accumulated precipitation of the driest month (mm)
RH_{ann}	Annual mean relative humidity (percent)
GHI_{ann}	Accumulated annual Global Horizontal Irradiation (J/m^2)
GHI_{max}	Maximum accumulated monthly Global Horizontal Irradiation (J/m^2)
GHI_{min}	Minimum accumulated monthly Global Horizontal Irradiation (J/m^2)
UV_{ann}	Accumulated annual UV irradiation (J/m^2)
WS_{ann}	Annual mean near-surface (2 m) wind speed (m/s)

Table 2
Raw climate variables used to determine the features.

Raw variable	Description	Source
TMP	Mean near-surface (2 m) temperature (°C)	CRU [35]
DTR	Mean daily temperature difference (°C)	CRU [35]
P	Accumulated precipitation (mm)	GPCC [36]
RH	Mean relative humidity (percent)	CDS [37]
GHI	Accumulated Global Horizontal Irradiation (J/m^2)	CDS [38]
UV	Accumulated UV irradiation (J/m^2)	CDS [38]
WS	Mean near-surface (10 m) wind speed (m/s)	CDS [38]

data is extracted from three different sources: the Climate Research Unit (CRU) of the University of East Anglia (CRU TS V. 4.06) [35]; the Global Precipitation Climatology Centre (GPCC), and, more specifically, the GPCC Full Data Reanalysis Version 5 [36]; and the Copernicus Climate Change Service Data Store (CDS), a service implemented by the European Centre for Medium-Range Weather Forecasts (ECMWF). From the latter, the datasets "Essential climate variables for assessment of climate variability from 1979 to present" [37] and "ERA5 monthly averaged data on single levels from 1940 to present" [38] were used. The climate data is quality controlled by checking for abnormalities such as missing values or outliers. The temperature and precipitation data used by Ascencio et al. [23] for the PV performance data is from the same sources as used in this work and in their follow-up [34] study, they also used the ERA5 data, which is used in this work.

UV irradiation and wind speed data are pre-processed as described in appendix A. Similarly, the resolution and range for the different data sources are harmonized during pre-processing as described in appendix B.

3. Use of machine learning for selection and weighing of climate variables

As a first step to generate an objective classification via ML, not only identifying the climate variables is essential, but also their levels of importance, or, mathematically speaking, their weights. These two issues are solved with the help of supervised learning. A linear regression model is built to predict worldwide specific energy yields from the knowledge of several climate parameters. The predicted values can be compared with known data to obtain a measure of the error of the model. This provides a method to analyse the relevance of the climate variables on PV performance: the more relevant a parameter is, the lower the error of the prediction. Hence, the model can be optimised by selecting the most significant variables. However, feature selection proves to be a challenging issue. Pearson coefficients, automated feature selection, and technical expertise are combined to choose 79 possible sets of features. Then, these are evaluated individually, and the results are carefully analysed to decide the features to use in the second step.

In this project, Python 3.10 was used. One of the strengths of Python is its third-party packages. Among the numerous packages used in this work, the *scikit-learn* deserves to be highlighted [40]. Scikit-learn is the most prominent Python library for ML, containing several state-of-the-art ML algorithms, as well as a thorough documentation. It is an open-source project, which has been widely used in industry and academia [41]. A more detailed description of our procedure can be found in Ref. [42].

3.1. Methodology for climate variable selection and weighing

The approach consists in implementing linear regression to analyse the suitability of each feature for predicting accurately the specific energy yield. In short, linear models make a prediction using a linear function of the input features [41]. In mathematical notation:

$$y_p = w_1 \cdot x_1 + w_2 \cdot x_2 + w_3 \cdot x_3 + \ldots + w_n \cdot x_n + b, \quad (2)$$

where y_p denotes the prediction, and x_i denotes the feature i. The parameters learnt by the model are the weights associated with each feature, w_i, and the interception, b.

A scheme of the methodology is illustrated in Fig. 1. Following the best practice in supervised learning, the dataset is partitioned into two subsets: the training data and the test data. Here, the training data corresponds to 75 percent of the original dataset, while the test data consists of the remaining 25 percent. This is just a general rule of thumb; similar partitions are also applicable [41]. After this step, it is necessary to rescale the data. Data scaling is a common preprocessing method applied before implementing the actual ML algorithm [41]. Data scaling consists in reconstructing the dataset to reduce the impact of different orders of magnitude [43]. This is typically caused by the different units employed among the features. For instance, irradiation has an order of magnitude of 10^8, clearly much higher than other variables like humidity or temperature. These discrepancies in scale and range directly affect the weights' calculation. Therefore, to obtain more reliable and comprehensive results, all data points must be transformed to the same scale [43].

In this work, *StandardScaler,* within the scikit-learn package, is applied. It transforms the data so that every feature has a mean equal to 0 and a variance of 1 [41]. Besides being easy to understand, this technique has proved successful in the optimisation of machine learning algorithms [43]. Thus, StandardScaler is used to scale the data. It should be stressed that, to avoid introducing bias, the transformation used to standardize the test data must be the same as the one used for the training data. [41]. This is indicated in Fig. 1 by the dashed arrow.

Finally, the linear regression model is built using the training data. Following the nomenclature introduced above, here the features, x_i, are the climate variables, while the target, y, is the specific energy yield. After fitting the weights, the model can be evaluated using the test data. In other words, the specific energy yield for each sample of the test data is predicted using the corresponding climate features, and the prediction is compared to the known value. Thus, the performance of the model can be measured. The lower the error, the higher the correlation between the climate features and the specific energy yield.

In principle, all features could be fed into the model and the algorithm would calculate their optimum weights. Then, these weights could be used for developing the classification. However, this would result in a complex classification, difficult to analyse and understand. Moreover, many climate variables are related to each other or have minor importance, so these can be discarded for the classification criteria. On the other hand, an insufficient number of features would result in poor accuracy. Therefore, it is essential to make a wise selection.

3.2. Climate variable correlation results

The climate features are not independent from each other, and there exist particular combinations which remarkably improve the performance of the model. For that reason, the importance of an individual feature depends on the other features with which it is combined. Consequently, the optimum combination can vary significantly when changing the number of features selected. Therefore, it is necessary to try several combinations to find the optimum set. The high number of possible combinations forces to simplify the procedure.

Certain statistical parameters and tools can facilitate the analysis. A first insight into the relevance of a feature is provided by the Pearson correlation coefficient or Pearson's r. It evaluates the linear correlation between the feature and the target [44]. Pearson's r measures the dependence of the individual variable, so no information is gained regarding the interactions between the climate features. Hence, feature selection cannot be based solely on these numbers. Nevertheless, it provides a sense of the importance of each feature and can be utilized to make some decisions. For instance, Ahmed et al. [43] considered significant only those variables with a Pearson coefficient higher than 0.4. Table 3 illustrates the calculated Pearson coefficients.

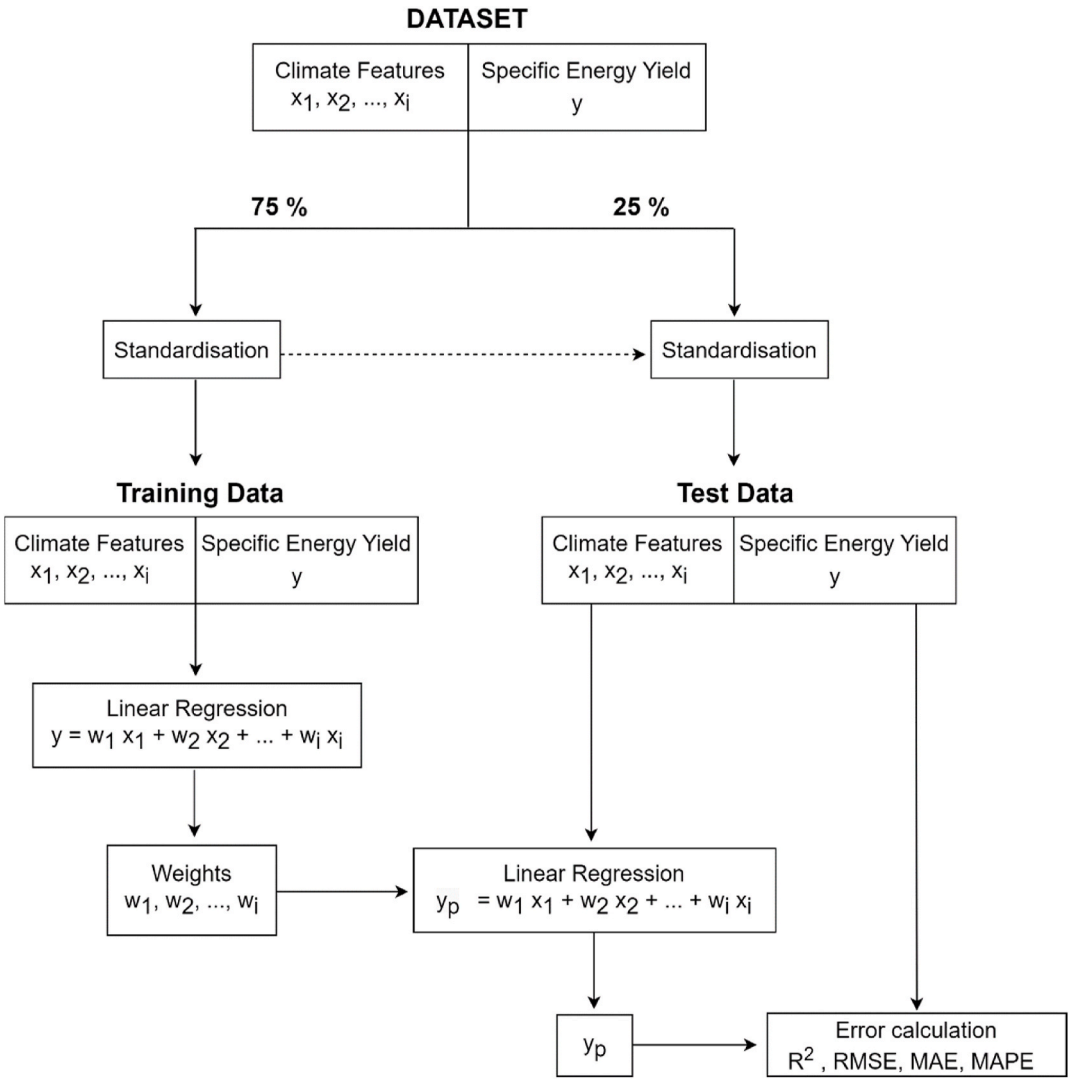


Fig. 1. Scheme of the methodology applied to select and weight the most relevant climate variables to the specific energy yield. Solid lines represent the flow of data. The dashed arrow indicates that the transformation used to standardize the test data is the same as the one used for the training data to avoid introducing bias.

Table 3
Pearson correlation coefficient for every climate feature. Values close to +1 or −1 indicate a high linear dependence between the feature and the specific energy yield.

Feature	Pearson correlation coefficient
T_{ann}	+0.66
T_{max}	+0.63
T_{min}	+0.60
DTR_{ann}	+0.76
P_{ann}	−0.13
P_{min}	−0.27
RH_{ann}	−0.72
GHI_{ann}	+0.92
GHI_{max}	+0.78
GHI_{min}	+0.77
UV_{ann}	+0.91
WS_{ann}	+0.01

It is clear from Table 3 that GHI_{ann} and UV_{ann} have a strong linear correlation with the specific energy yield. Therefore, these variables are expected to play a fundamental role in the classification. On the other hand, the low Pearson coefficients for precipitation and wind speed, suggest that these variables will not be so relevant, reinforcing the need to go beyond the original KG classification method. Particularly, WS_{ann} has a coefficient of almost zero, so it will be disregarded for the rest of the analysis.

3.3. Climate variables scaling and weighing results

Overall, for every possible number of features (ranging from 1 to 11), the model has been evaluated using different combinations. The possible combinations have been chosen based on the Pearson coefficients, and technical expertise. Furthermore, Recursive Feature Elimination (RFE) was implemented to guide the procedure. In total, 79 options have been evaluated.

Table 4 summarises the optimum combination found for each number of features, with their associated errors. Error measures include the coefficient of determination (R^2), the root mean square error (RMSE), the mean absolute error (MAE), and the mean absolute percentage error (MAPE). Furthermore, a case consisting of a random variable is included to show the validity of the model and establish a reference. The errors shown in Table 4 are calculated using the test data. Furthermore, since the data is randomly divided into training and test data, running the algorithm again results in slightly different values. To reduce this effect, the algorithm was executed three times, and the error and weights were averaged.

Table 4 provides, for a fixed number of features, the optimum set, the

Table 4
Optimum combination for each possible number of features. The weights given to the climate variables and the error of the model are shown.

Features	T_{ann}	T_{max}	T_{min}	DTR_{ann}	P_{ann}	P_{min}	RH_{ann}	GHI_{ann}	GHI_{max}	GHI_{min}	UV_{ann}	Random	R^2	RMSE	MAE	MAPE (%)
0												1.61	0	360.4	301.28	28.3
1								332.3					0.852	139.61	109.14	9.9
2								533.28		−216.81			0.904	112.37	84.05	7.9
3								894.04	−165.25	−471.95			0.923	99.71	77.88	6.8
4	−65.77							964.93	−200.28	−463.66			0.931	95.27	73.84	6.7
5			−104.31					169.91	−133.00	−435.09	748.66		0.937	90.77	70.31	6.5
6	−98.57					−18.39		266.6	−169.9	−497.71	732.93		0.938	89.45	68.86	6.3
7	500.46	−183.59	−423.15					269.94	−121.51	−426.79	610.23		0.941	87.31	67.64	6.2
8	501.44	−191.16	−409.65			−17.56		252.12	−134.75	−452.43	652.1		0.943	86.03	66.77	6.1
9	495.06	−190.9	−404.93			−17.2	−4.19	210.69	−132.36	−450.36	688.52		0.943	86.06	66.86	6.1
10	500.14	−192.01	−408.37		−1.40	−16.28	−4.01	209.25	−132.57	−450.56	690.07		0.943	85.7	66.49	6.0
11	491.87	−189.94	−397.83	7.48	−2.30	−16.24	−1.13	217.01	−132.73	−447.14	671.82		0.944	85.33	66.75	6.1

weights, and the error associated with that linear regression model. Based on these results, the criteria for developing the classification in the next section can be defined. The first observation is the importance of the irradiation features, when fixing the number of features to three, the optimum combination consists of the three measures of GHI. Only using these variables, an R^2 of 0.923 and a MAPE of 6.8 percent are achieved. For more than three climate variables, the temperature starts to play a fundamental role too, appearing T_{ann} when adding the fourth feature. It is interesting to note that T_{ann} and T_{min} provide more information than T_{max} or DTR_{ann}. From the five features, UV_{ann} becomes the variable with the highest weight, a sign of its importance to the model. The next variable to appear is P_{min}, which, after showing for the first time with six features and being disregarded with seven, is always selected. Finally, with very low weights, RH_{ann}, P_{ann}, and DTR_{ann} are added, in that order.

Besides the features, the errors must be carefully analysed. In particular, the regression score function, R^2. From eight features, the R^2 takes a virtually constant value of 0.943. This suggests that selecting more than eight features is unnecessary, since the model's complexity would increase without tangible improvement. Even for less than eight features there is a clear trade-off between reducing complexity of data collection and accuracy. In a longer format, classification based on four, five, seven, and eight features were considered [42]. From Tables 4 and it is known that best model with seven features has a MAE of 67.6, while the MAPE is 6.2 percent. Based on this, it was found that seven features (T_{ann}, T_{max}, T_{min}, GHI_{ann}, GHI_{max} GHI_{min}, UV_{ann}) provide the best compromise between accuracy and unnecessary complexity. Thus, we will use this combination of seven features for the rest of this work.

For this combination of climate variables, Fig. 2 shows a visual assessment of the predictions. In this figure, the specific energy yield predictions (y-axis) are directly compared to the known values or targets (x-axis). Ideally, a straight line (red line in the figure) should be obtained. In general, the predictions follow a similar behaviour to the actual values, except for regions characterised by a very low specific energy yield, where the discrepancy is higher. Therefore, for these regions, the classification could be less accurate.

Fig. 2. Comparison between the test data specific energy yield targets and the predictions made by the linear regression model using seven features (T_{ann}, T_{max}, T_{min}, GHI_{ann}, GHI_{max} GHI_{min}, UV_{ann}).

4. Use of machine learning to classify PV climate zones

In step 1, as described in the previous section, the most relevant features were identified and the optimum combination for seven features was proposed. In this section, step 2, a method to develop the final classification is illustrated. It consists in creating several classifications using the concept of *k-means*, followed by a careful qualitative analysis based on various parameters and tools. In this way, an optimum classification comprising 21 climate groups is achieved. A more detailed description of our procedure can be found in Ref. [42].

4.1. Classification methodology

The procedure consists of the following steps, shown in Fig. 3. First, the features selected in section 3 are standardised (using StandardScaler) and multiplied by their corresponding weights. Then, k-means is applied to create the clusters from these features. In principle, the desired number of clusters can be fixed beforehand, and a classification is obtained via the k-means algorithm. However, how many clusters should be formed?

The first step to building an optimum classification is to define an evaluation method. This essential requirement is usually the most challenging part of clustering studies [41]. In unsupervised learning, in contrast to supervised, there are no known values to compare and assess the results of the algorithm with. To guide this decision, several mathematical strategies, such as the elbow method, have been proposed. However, unfortunately, these pure quantitative analyses rarely work for the datasets found in practice [41]. Indeed, these methods were applied in this work with poor performance. The reality is that there is not a clear and unique solution to this challenge. Instead, a careful and tedious qualitative exploration procedure is required. This is a common approach in clustering algorithms [41]. Fortunately, numerous tools can facilitate the study and help to make objective decisions. Data visualization is especially relevant for this project. Other insightful parameters are the clusters' centres and sizes. In the following, the procedure will be demonstrated using 19, 20, 21, and 22 clusters.

4.2. Analysing clustering results

The exploration method is illustrated hereafter. For making an objective decision, it is essential to know how the clusters are formed and what characterize them. There exists a visual tool that proves very insightful: the pair plot. A pair plot consists of a matrix of scatter plots, each representing the points for every possible pair of features. The diagonal of this matrix is filled with a histogram of each feature [41]. The data points are coloured according to the clusters they belong to. This figure enables an understanding of how the clusters are formed, and their main properties. Furthermore, it can be used to predict the formation of new groups. The pair plot associated with seven features and 19 clusters is illustrated in Fig. 4.

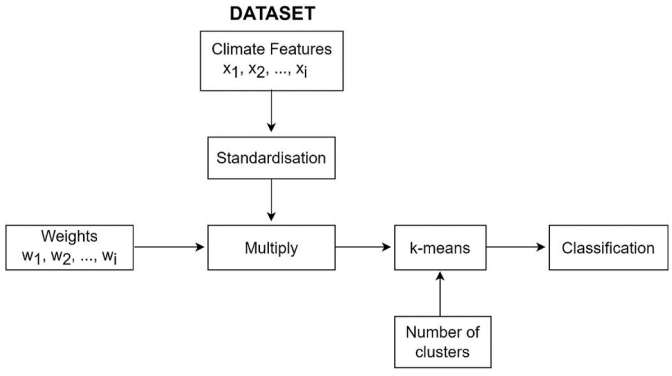


Fig. 3. Procedure to create the classification from the features and weights determined in section 3. The desired number of clusters is given as an input to the algorithm.

The pair GHI_{ann} – T_{ann} is very informative. Fig. 5A illustrates this for 19 clusters. This figure enables understanding the main properties of the clusters. First, the high weight given to the GHI_{ann} is here evident. Initially for lower feature or cluster counts, the clusters are mainly formed based on this criterion. However, there is a point at which temperature starts playing a role too and, consequently, horizontal divisions appear. In particular, the reddish clusters are characterized by a high T_{ann}. Their cluster centres indicate an average T_{ann} of 24 °C. It is seen that these reddish clusters have a similar level of irradiation to the greenish ones but a higher temperature, suggesting a tropical climate. On the other hand, the purple clusters are characterized by a very low T_{ann} (polar), while the orange regions present a high GHI_{ann} but moderate to low T_{ann}, presumably being mountain regions. This reasoning can be expanded, with the help of other pair of features, to obtain a description of every cluster.

Now, it is interesting to analyse what happens when the number of clusters is increased. Fig. 5 shows the pair GHI_{ann} – T_{ann} for 19 (A), 20 (B), 21 (C), and 22 (D) clusters. When increasing the number of clusters to 20, the bluish clusters are split, and a new blue group appear. Similarly, a new green group is formed when the number of clusters increases to 21. Both clusters constitute independent clouds of points and have clearly defined properties. These two new clusters identify relevant climate regions, significantly improving the classification's accuracy. However, when going from 21 to 22 clusters, a grey cluster is formed whose significance in terms of the meaningful features, GHI_{ann} and T_{ann}, is very unclear. It overlaps other clusters and does not show clearly defined properties. This suggests that forming 22 clusters or more is not appropriate and that 21 clusters is the ceiling for seven features.

4.3. PV climate classification results

The final classification is illustrated in Fig. 6. Overall, it comprises 21 climate regions based on seven features: T_{ann}, T_{max}, T_{min}, GHI_{ann}, GHI_{max}, GHI_{min}, and UV_{ann}.

Names and colours associated with the clusters have been proposed. These are indicated by the bar at the right of the figure. They are inspired by the approach followed in KGPV. First, clusters are divided into six climate types: Tropical (Tro), Desert (Des), Mountainous (Mou), Temperate (Tem), Cold (Col), and Polar (Pol). Then, the clusters inside each of these climate types are ordered from minor to greater irradiation. Therefore, both Tro1 and Tro4 are tropical climates, but Tro4 has a higher level of irradiation than Tro1. It is important to note that these numbers only apply inside the same climate type. Hence, even though Pol1 and Tro1 have the same number, they do not have the same level of irradiation. Table 5 summarises the cluster's names, centres, and sizes.

Even though the names are inspired by the KGPV scheme, they are ultimately justified by the properties of the clusters. These are understood with the help of the cluster's centres and pair plot (Fig. 4.). Overall, Tropical clusters are characterised by a high T_{min} and a low seasonal dependence. Desert shows a high T_{max}, GHI_{ann}, and UV_{ann}. On the other hand, Mountainous regions, despite presenting a high GHI_{ann} and UV_{ann}, have moderate to low temperatures. The most relevant features of the Temperate clusters are moderate temperatures and high seasonal dependence. Lastly, Cold is characterized by a low T_{min} and high seasonal dependence, while Polar shows a low T_{ann} and extreme seasonal dependence. An extensive description of the clusters can be found in Ref. [42].

5. Discussion and comparison to KGPV

The main objective is to create a ML driven PV climate classification based on the most relevant climate variables to the specific energy yield. Has this objective been achieved? For this purpose, Fig. 7 illustrates the scatter plot for the pair GHI_{ann} – Y_f. One can see a clear relationship

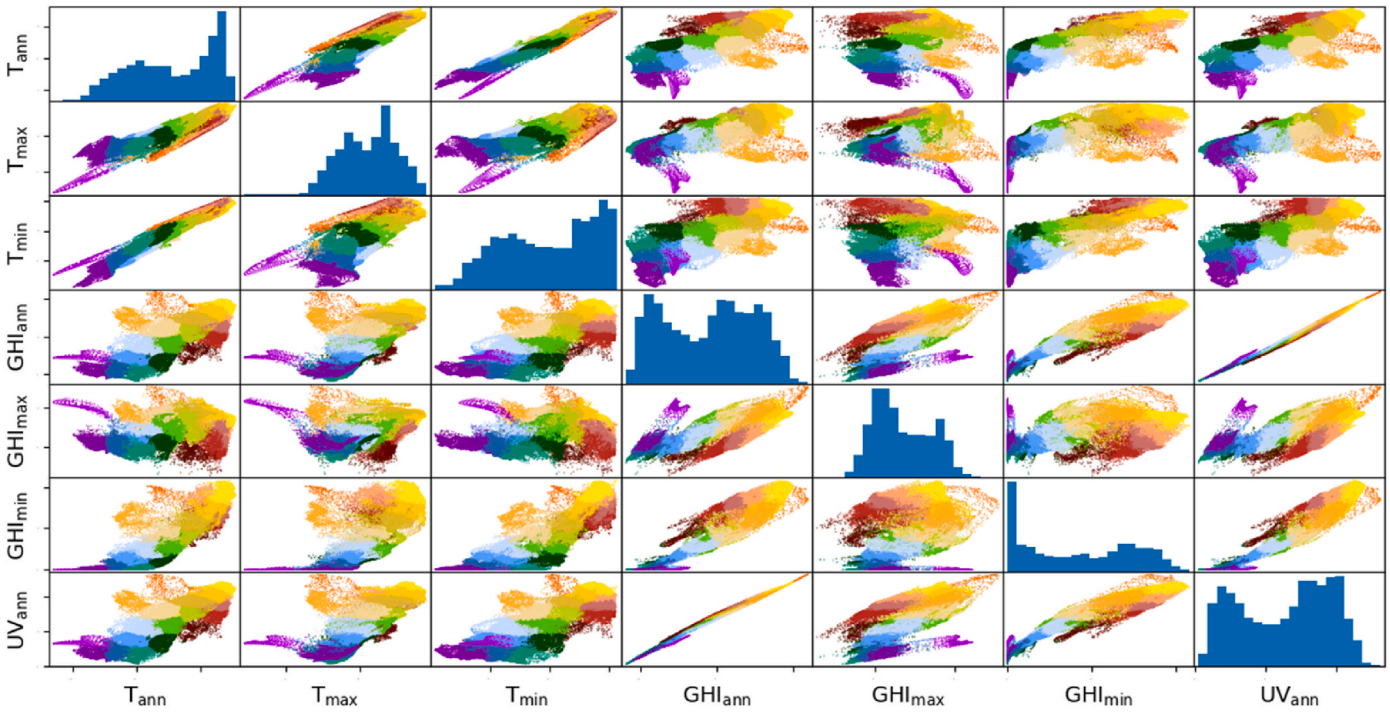


Fig. 4. Pair plot associated with the classification based on seven features and 19 clusters. The units of temperature and irradiation, not shown in the figure for the sake of clarity, are °C and J/m^2, respectively. The principal diagonal of this matrix represents a histogram of each feature, while the scatter plots show how the points are distributed and classified from the point of view of a pair of features.

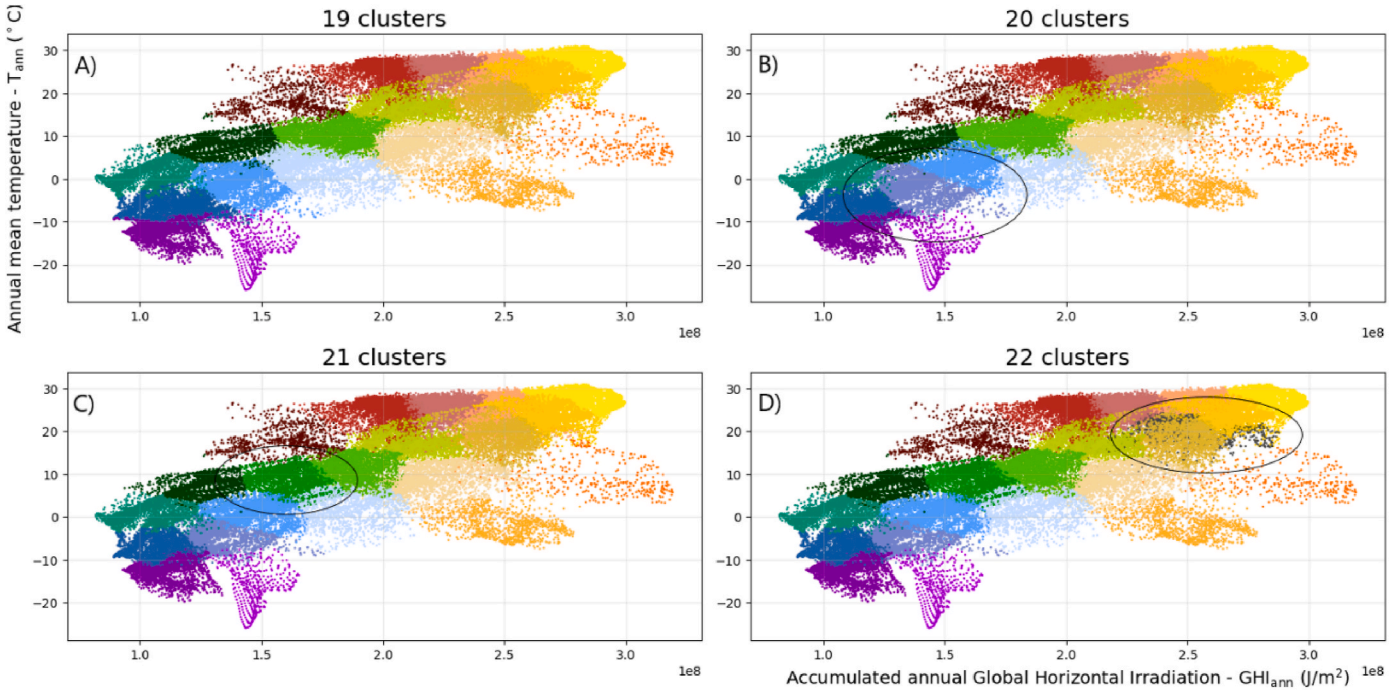


Fig. 5. Scatter plot GHI_{ann} – T_{ann} for 19 (A), 20 (B), 21 (C), and 22 (D) clusters.

between the created clusters and the specific energy highlighting the positive result of the found classification. The impact of temperature is here evident. Of course, considering regions with an equivalent range of temperatures, the higher the irradiation, the higher the specific energy yield. One point for improvement in future works is that the relationship between the clusters and the specific energy yield is less clear for regions with specific energy yield low values. This is in accordance with the conclusions drawn in section 3, where it was seen that the predictions for these regions are less accurate.

The secondary objective is to evaluate additional climate variables namely wind speed, relative humidity, UV irradiation and also the irradiation in the brightest and darkest months. According to the feature selection results listed in Table 4, GHI_{min} is the second most impactful climate variable, when it comes to predicting the energy yield. This is

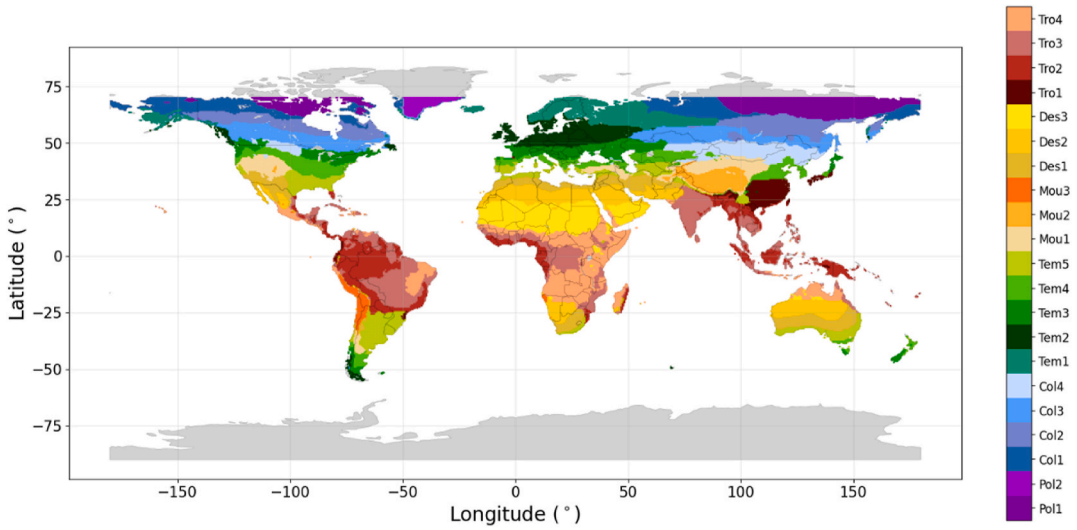


Fig. 6. ML driven PV climate classification. First index: Tro-Tropical, Des-Desert, Mou-Mountainous, Tem-Temperate, Col-Cold, Pol-Polar. The second index orders from minor to greater irradiation the clusters inside a particular climate type. The data to build this figure can be found in the supplementary file.

Table 5
Cluster's names, centres, and sizes. Temperatures are given in °C and irradiations in J/m^2. The size is the number of data points in 0.5° by 0.5° resolution constituting the cluster.

Name	T_{ann}	T_{max}	T_{min}	GHI_{ann} ($\cdot 10^8$)	GHI_{max} ($\cdot 10^7$)	GHI_{min} ($\cdot 10^7$)	UV_{ann} ($\cdot 10^6$)	Size
Tro4	24.54	27.12	21.58	2.49	2.43	1.75	14.83	4098
Tro3	25.64	28.08	22.88	2.23	2.23	1.51	13.59	4788
Tro2	25.05	26.69	22.83	2.00	2.00	1.33	12.46	3618
Tro1	17.52	26.26	7.51	1.61	1.84	0.87	10.26	997
Des3	27.96	33.87	20.30	2.81	2.73	1.86	16.11	3202
Des2	23.56	31.51	14.03	2.67	2.84	1.46	15.3	3613
Des1	19.01	27.9	9.66	2.47	2.83	1.17	14.29	2929
Mou3	9.43	12.15	5.84	2.86	2.98	1.76	16.40	384
Mou2	−1.48	9.64	−13.62	2.48	2.81	1.28	14.26	1108
Mou1	8.89	21.88	−4.94	2.22	2.72	0.88	12.89	2004
Tem5	17.46	25.77	8.88	2.11	2.58	0.89	12.60	2951
Tem4	11.08	23.33	−1.68	1.88	2.44	0.65	11.32	2726
Tem3	8.53	20.71	−3.90	1.60	2.25	0.39	9.83	2284
Tem2	7.14	18.00	−3.44	1.26	1.99	0.17	7.91	2031
Tem1	0.69	15.33	−13.33	1.05	1.90	0.04	6.59	2847
Col4	2.52	19.68	−17.20	1.82	2.35	0.57	10.85	2086
Col3	0.94	17.52	−17.45	1.46	2.15	0.28	8.92	3493
Col2	−4.84	14.82	−25.67	1.31	2.12	0.15	8.02	3295
Col1	−6.84	12.60	−24.79	1.07	2.01	0.02	6.64	3146
Pol2	−16.20	−3.75	−26.27	1.47	3.00	0.01	8.39	450
Pol1	−12.19	12.03	−35.49	1.10	2.09	0.01	6.75	3368

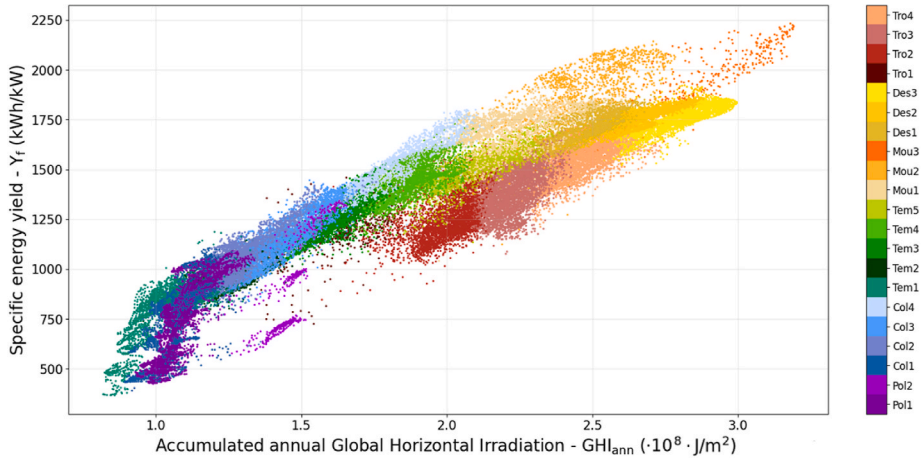


Fig. 7. Scatter plot for the pair GHI_{ann} – Y_f.

followed by GHI_{max} in third and UV_{ann} in fifth position. In contrast, wind speed has low Pearson correlation coefficient, while relative humidity had nearly no measurable impact on the accuracy of the linear regression. Similarly, it is found that the original KG parameters based on precipitation and annual mean daily temperature difference do little to improve the accuracy of the linear regression and are thus also not selected for the ML driven PV climate classification.

To compare the ML driven climate classification obtained in this work and KGPV [23] the scatter plot $GHI_{ann} - T_{ann}$ is shown for both classifications in Fig. 8. Similar types of climates are present in both classifications: Tropical (red), Desert (yellow), Temperate (green), Cold (blue), and Polar (purple). However, one climate zone is very different, defined in KGPV as Steppe (C), which is not clearly identified in this work. Instead, a Mountainous climate type has been proposed. KGPV classifies these regions as Polar, because of the low temperatures. However, their high irradiation strongly suggests considering an independent group. These Mountainous regions are characterised by a high GHI_{ann} and moderate to low temperatures, which make them the regions with the highest PV performance and should not be mixed with the Polar regions of the northern hemisphere.

Regarding the number of PV climate zones, KGPV initially considers 24, but half of them is neglected based on a land-surface ratio and population density criterion, so in practice a total of 12 KGPV climate zones remains. By contrast, in this work 21 PV climate zones have been identified, which is lower than the initial KGPV model, but non-climate variables, such as population or land area, were not used to reduce their number. Consequently, more subdivisions are considered inside each climate type, and a higher level of detail is achieved. For instance, KGPV distinguishes solely between two Tropical climates (AH and AK), in contrast to the four regions found in this work. Furthermore, since the methodologies are different, even equivalently named zones present different climates.

In general, KGPV's climate zones tend to overlap more with each other such as desert high (BH) and temperate high (DH). The clusters found by this work show a much clearer classification regarding the meaningful features GHI_{ann} and T_{ann}.

To quantitively measure for the effectiveness of the two PV climate classifications, in addition to the visual comparison from above, the sum of squared errors is calculated for both [45]. When considering the seven climate variables used to obtain the climate classification the sum of squared errors is 58 % reduced for the machine learning based PV climate classification compared to KGPV. Moreover, calculating the sum of squared errors of the specific energy yield alone for both classifications it is reduced by 63 % for the machine learning based PV climate classification. This showcases that the machine learning approach to PV-climate classification can provide increase accuracy compared to classical approaches such as KGPV.

Overall, the methodology developed in this work proves to be a promising alternative to the previous classifications proposed. The results of the model may be further improved by means of the following recommendations. The feature selection procedure discussed in section 3. Here, the analysis was simplified by selecting 79 combinations. Even though this approach produced a satisfactory result, not all combinations have been checked. Implementing an optimisation algorithm such as Particle Swarm Optimisation (PSO) or the Genetic Algorithm (GA) is recommended to consider more possibilities. Secondly, the accuracy can be improved by considering non-linear dependencies. Multivariate Adaptive Regression Spline (MARS) is a promising approach to integrate

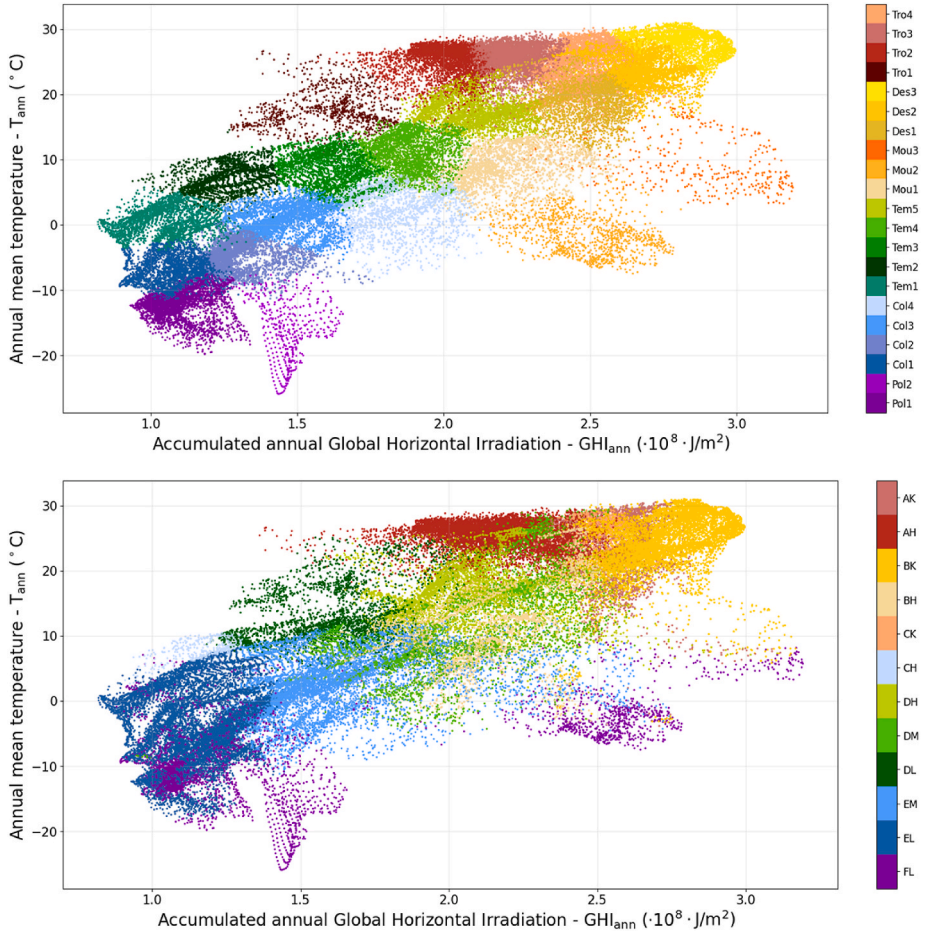


Fig. 8. Scatter plots for the pair $GHI_{ann} - T_{ann}$ for ML driven PV climate classification (top, this work) and KGPV (bottom, [23]).

non-linear dependencies and, at the same time, keep the logical meaning of the weights. Finally, as more high-quality PV performance data from comparable PV systems becomes available from around the globe efforts should be made to validate the PV performance data used in this study.

6. Conclusions

A machine learning-approach is proposed to develop a PV-climate classification. Supervised learning is used to identify and weigh the climate variables more correlated to the specific energy yield, while unsupervised learning is used to create the classification.

Overall, the combination corresponding to the optimum for seven climate features ($T_{ann,}$ $T_{max,}$ $T_{min,}$ $GHI_{ann,}$ GHI_{max} $GHI_{min,}$ UV_{ann}) is selected. This model performs predictions of the specific energy yield with a MAPE of 6.2 percent. Accuracy appears to be higher for medium and high specific energy yields than for low ones.

The inclusion of irradiation in the darkest and brightest months as well as UV irradiation improves accuracy, while wind speed, relative humidity, precipitation and annual mean daily temperature difference has little to no impact on accuracy.

Pair plots are used to determine that 21 clusters is the correct number for seven features. The clusters found by this work show a much clearer classification than KGPV. This is quantified by 58 % reduced sum of squared errors for the machine learning based PV climate classification compared to KGPV. A mountainous climate characterized by moderate to low temperature and high irradiation is uncovered compared to KGPV. This showcases that the machine learning approach to PV-climate classification can provide increase accuracy compared to classical approaches such as KGPV.

As more machine learning based PV climate classifications follow this initial work, the model may be further improved by implementing an optimisation algorithm such as PSO or GA to select the features, and by integrating non-linear dependencies via MARS.

CRediT authorship contribution statement

Francisco Javier Triana de las Heras: Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation. **Olindo Isabella:** Writing – review & editing. **Malte Ruben Vogt:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

Data availability

Data on the climate zone associated with each location is in the supplementary material. Further data will be provided upon seasonable request.

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

The results contain modified Copernicus Climate Change Service information 2023. Neither the European Commission nor ECMWF is responsible for any use that may be made of the Copernicus information or data it contains. TU Delft's Climate Action Program Flagship project "Materials for circular renewable energy technologies" supported this work. The authors want to thank A. Alcañiz for fruitful discussion and J. Ascencio-Vásquez for answering questions about his work on KGPV.

Appendix A. UV and wind speed pre-processing

The ERA5 dataset contains a variable called "Downward UV radiation at the surface" which is used to calculate UV_{ann}. However, as discussed in [34], UV is typically referred to for wavelengths below 400 nm, while the variable given in ERA5 covers the range from 200 to 440 nm. As a consequence, the latter significantly overestimates the UV irradiation. Following the alternative proposed in the same paper, UV is calculated using the approach given in [46]:

$$UV_A = \left(7.210 - 2.365 \cdot k_t^*\right) \cdot 10^{-2} \cdot GHI \tag{A1.1}$$

$$UV_B = \left(1.897 - 0.860 \cdot k_t^*\right) \cdot 10^{-3} \cdot GHI \tag{A1.2}$$

$$UV = UV_A + UV_B \tag{A1.3}$$

$$k_t^* = \max(0.1, min(0.7, k_t)) \tag{A1.4}$$

With k_t representing the clearness index, i.e., the GHI divided by the solar radiation at the top of the atmosphere, both variables available in the ERA5 dataset.

On the other hand, WS is available as the horizontal speed of the wind at a height of 10 m above the surface of the Earth (10 m wind speed). However, temperature data is known at a height of 2 m, so the following correction is applied to obtain the wind speed at the same height [34]:

$$WS_{2m} = \left(\frac{2}{10}\right)^{0.2} \cdot WS_{10m} \tag{A1.5}$$

Appendix B. – Pre-processing of data resolution and range

When building the dataset, as a final step, the specific energy yield (Y_f) is added as another column to the climate dataset. The result is a worldwide grid with resolution 0.5° latitude by 0.5° longitude. Thus, for every location (rows), the dataset contains its climate features and specific energy yield (columns). This enables finding a relation between climate and Y_f.

Nevertheless, some data processing is required to match all climate variables and the specific energy yield, since the raw data present different resolutions. For climate, the resolution of the data given by the CRU TS V. 4.06 and GPCC is 0.5° × 0.5°, while the data from the CDS has a resolution of

$0.25^\circ \times 0.25^\circ$. Moreover, the range of latitude and longitude values is different for each source. Therefore, the data must be pre-processed, and in some cases, interpolated to obtain a final grid of $0.5^\circ \times 0.5^\circ$ with every variable having the same range.

Regarding the specific energy yield, this data is already available with a $0.5^\circ \times 0.5^\circ$ resolution. However, the actual range of values is different. For instance, latitude in the climate dataset starts counting at -55.25°, while in the Y_f, it starts at -55°. Therefore, it is necessary to interpolate to have both datasets referred to exactly the same locations. The same issue is found regarding longitude. As a final remark, latitude values range from -55° to 70°, so Antarctica and most of Greenland are excluded.

Appendix C. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.renene.2025.123685.

References

[1] IRENA, Renewable capacity statistics 2023. https://www.irena.org/Publications/2023/Mar/Renewable-capacity-statistics-2023, 2023.

[2] IRENA, Renewable capacity statistics 2024, International Renewable Energy Agency (2024) 25–28. https://www.irena.org/-/media/Files/IRENA/Agency/Publication/2024/Mar/IRENA_RE_Capacity_Statistics_2024.pdf.

[3] IRENA, Renewable power generation costs in 2021. https://www.irena.org/publications/2022/Jul/Renewable-Power-Generation-Costs-in-2021, 2022.

[4] IEA, Solar PV. https://www.iea.org/reports/solar-pv, 2022.

[5] Ember, 73rd statistical review of world energy, Energy institue (2024). https://www.energyinst.org/_data/assets/pdf_file/0006/1542714/684_EI_Stat_Review_V16_DIGITAL.pdf.

[6] Our world in data, Share of electricity generated by solar power. https://ourworldindata.org/grapher/share-electricity-solar.

[7] P.P. Shinde, S. Shah, A review of machine learning and deep learning applications, in: 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), IEEE, Pune, India, Aug. 2018, pp. 1–6, https://doi.org/10.1109/ICCUBEA.2018.8697857.

[8] A. Alcañiz, D. Grzebyk, H. Ziar, O. Isabella, Trends and gaps in photovoltaic power forecasting with machine learning, Energy Rep. 9 (Dec. 2023) 447–471, https://doi.org/10.1016/j.egyr.2022.11.208.

[9] G.M. Tina, C. Ventura, S. Ferlito, S. De Vito, A state-of-art-review on machine-learning based methods for PV, Appl. Sci. 11 (16) (Aug. 2021) 7550, https://doi.org/10.3390/app11167550.

[10] U.K. Das, et al., Forecasting of photovoltaic power generation and model optimization: a review, Renew. Sustain. Energy Rev. 81 (Jan. 2018) 912–928, https://doi.org/10.1016/j.rser.2017.08.017.

[11] R.A. Rajagukguk, R.A.A. Ramadhan, H.-J. Lee, A review on deep learning models for forecasting time series data of solar irradiance and photovoltaic power, Energies 13 (24) (Dec. 2020) 6623, https://doi.org/10.3390/en13246623.

[12] J. Cai, X. Chu, K. Xu, H. Li, J. Wei, Machine learning-driven new material discovery, Nanoscale Adv. 2 (8) (2020) 3115–3130, https://doi.org/10.1039/D0NA00388C.

[13] S. Sun, et al., Accelerated development of perovskite-inspired materials via high-throughput synthesis and machine-learning diagnosis, Joule 3 (6) (Jun. 2019) 1437–1451, https://doi.org/10.1016/j.joule.2019.05.014.

[14] A. Mellit, G.M. Tina, S.A. Kalogirou, Fault detection and diagnosis methods for photovoltaic systems: a review, Renew. Sustain. Energy Rev. 91 (Aug. 2018) 1–17, https://doi.org/10.1016/j.rser.2018.03.062.

[15] D.S. Pillai, N. Rajasekar, A comprehensive review on protection challenges and fault diagnosis in PV systems, Renew. Sustain. Energy Rev. 91 (Aug. 2018) 18–40, https://doi.org/10.1016/j.rser.2018.03.082.

[16] F. Wang, Z. Xuan, Z. Zhen, K. Li, T. Wang, M. Shi, A day-ahead PV power forecasting method based on LSTM-RNN model and time correlation modification under partial daily pattern prediction framework, Energy Convers. Manag. 212 (May 2020) 112766, https://doi.org/10.1016/j.enconman.2020.112766.

[17] K.A. Baharin, H. Abdul Rahman, M.Y. Hassan, C.K. Gan, Short-term forecasting of solar photovoltaic output power for tropical climate using ground-based measurement data, J. Renew. Sustain. Energy 8 (5) (Sep. 2016) 053701, https://doi.org/10.1063/1.4962412.

[18] W. Liu, J. Li, S. Li, J. Luo, X. Jiang, Research on optimum tilt angle of photovoltaic module based on regional clustering of influencing factors of power generation, Int. J. Energy Res. 45 (7) (Jun. 2021) 11002–11017, https://doi.org/10.1002/er.6584.

[19] W. Liu, X. Jiang, S. Li, J. Luo, G. Wen, Photovoltaic module regional clustering in mainland China and application based on factors influencing field reliability, Renew. Sustain. Energy Rev. 133 (Nov. 2020) 110339, https://doi.org/10.1016/j.rser.2020.110339.

[20] S. Li, W. Liu, S. Hu, B. Xu, A method for determining the applicable geographical regions of PV modules field reliability assessment results based on regional clustering of environmental factors and their weights, Sustain. Energy Technol. Assessments 53 (Oct. 2022) 102620, https://doi.org/10.1016/j.seta.2022.102620.

[21] L. Micheli, M. Muller, S. Kurtz, Determining the effects of environment and atmospheric parameters on PV field performance, in: 2016 IEEE 43rd Photovoltaic Specialists Conference (PVSC), IEEE, Portland, OR, USA, Jun. 2016, pp. 1724–1729, https://doi.org/10.1109/PVSC.2016.7749919.

[22] D.C. Jordan, J.H. Wohlgemuth, S.R. Kurtz, Technology and climate trends in PV module degradation, 27th Eur. Photovolt. Sol. Energy Conf. Exhib. (2012), https://doi.org/10.4229/27THEUPVSEC2012-4DO.5.1, 3118-3124, p. 7 pages, 6571 kb.

[23] J. Ascencio-Vásquez, K. Brecl, M. Topič, Methodology of köppen-geiger-photovoltaic climate classification and implications to worldwide mapping of PV system performance, Sol. Energy 191 (Oct. 2019) 672–685, https://doi.org/10.1016/j.solener.2019.08.072.

[24] J. Zscheischler, M.D. Mahecha, S. Harmeling, Climate classifications: the value of unsupervised clustering, Procedia Comput. Sci. 9 (2012) 897–906, https://doi.org/10.1016/j.procs.2012.04.096.

[25] B. Bochenek, Z. Ustrnul, Machine learning in weather prediction and climate analyses—applications and perspectives, Atmosphere 13 (2) (Jan. 2022) 180, https://doi.org/10.3390/atmos13020180.

[26] Z. Bao, et al., The sensitivity of vegetation cover to climate change in multiple climatic zones using machine learning algorithms, Ecol. Indic. 124 (May 2021) 107443, https://doi.org/10.1016/j.ecolind.2021.107443.

[27] R. Beigaitė, et al., Identifying climate thresholds for dominant natural vegetation types at the global scale using machine learning: average climate versus extremes, Glob. Change Biol. 28 (11) (Jun. 2022) 3557–3579, https://doi.org/10.1111/gcb.16110.

[28] P. Parisouj, H. Mohebzadeh, T. Lee, Employing machine learning algorithms for streamflow prediction: a case study of four river basins with different climatic zones in the United States, Water Resour. Manag. 34 (13) (Oct. 2020) 4113–4131, https://doi.org/10.1007/s11269-020-02659-5.

[29] L. Yang, K. Lyu, H. Li, Y. Liu, Building climate zoning in China using supervised classification-based machine learning, Build. Environ. 171 (Mar. 2020) 106663, https://doi.org/10.1016/j.buildenv.2020.106663.

[30] P.K. Dash, N.C. Gupta, R. Rawat, P.C. Pant, A novel climate classification criterion based on the performance of solar photovoltaic technologies, Sol. Energy 144 (Mar. 2017) 392–398, https://doi.org/10.1016/j.solener.2017.01.046.

[31] IEA-PVPS, Guidelines for operation and maintenance of photovoltaic power plants in different climates. https://iea-pvps.org/key-topics/guidelines-for-operation-and-maintenance-of-photovoltaic-power-plants-in-different-climates/, 2022.

[32] N. Skandalos, et al., Building PV integration according to regional climate conditions: BIPV regional adaptability extending Köppen-Geiger climate classification against urban and climate-related temperature increases, Renew. Sustain. Energy Rev. 169 (Nov. 2022) 112950, https://doi.org/10.1016/j.rser.2022.112950.

[33] T. Karin, C.B. Jones, A. Jain, Photovoltaic degradation climate zones, in: 2019 IEEE 46th Photovoltaic Specialists Conference (PVSC), IEEE, Chicago, IL, USA, Jun. 2019, pp. 687–694, https://doi.org/10.1109/PVSC40753.2019.8980831.

[34] J. Ascencio-Vásquez, I. Kaaya, K. Brecl, K.-A. Weiss, M. Topič, Global climate data processing and mapping of degradation mechanisms and degradation rates of PV modules, Energies 12 (24) (Dec. 2019) 4749, https://doi.org/10.3390/en12244749.

[35] I. Harris, T.J. Osborn, P. Jones, D. Lister, Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset, Sci. Data 7 (1) (Apr. 2020) 109, https://doi.org/10.1038/s41597-020-0453-3.

[36] B. Rudolf, A. Becker, U. Schneider, A. Meyer-Christoffer, M. Ziese, New GPCC full data reanalysis version 5 provides high-quality gridded monthly precipitation data, GEWEX News 21 (2011).

[37] Copernicus Climate Change Service, ERA5 monthly averaged data on single levels from 1979 to present, ECMWF (2019), https://doi.org/10.24381/CDS.F17050D7.

[38] Copernicus Climate Change Service, ERA5 monthly averaged data on single levels from 1940 to present, Copernicus Climate Change Service (C3S) Climate Data Store (CDS) (2019), https://doi.org/10.24381/CDS.F17050D7.

[39] N. Bansal, S.P. Jaiswal, G. Singh, Comparative investigation of performance evaluation, degradation causes, impact and corrective measures for ground mount and rooftop solar PV plants – a review, Sustain. Energy Technol. Assessments 47 (Oct. 2021) 101526, https://doi.org/10.1016/j.seta.2021.101526.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[41] A.C. Müller, S. Guido, Introduction to Machine Learning with Python: a Guide for Data Scientists, first ed., O'Reilly, Beijing Boston Farnham Sebastopol Tokyo, 2016.

[42] F.J. Triana de las Heras, Machine Learning Driven PV-Climate Classification, Delft University of Technology, 2023.

[43] R. Ahmed, V. Sreeram, R. Togneri, A. Datta, M.D. Arif, Computationally expedient photovoltaic power forecasting: a LSTM ensemble method augmented with adaptive weighting and data segmentation technique, Energy Convers. Manag. 258 (Apr. 2022) 115563, https://doi.org/10.1016/j.enconman.2022.115563.

[44] Scikit-learn, Feature selection. https://scikitlearn.org/stable/modules/feature_selection.html.
[45] N. Sankalana, K-means clustering: choosing optimal K, process, and evaluation methods, in: https://medium.com/@nirmalsankalana/k-means-clustering-choosing-optimal-k-process-and-evaluation-methods-2c69377a7ee4. (Accessed 28 February 2025).
[46] L. Wald, A Simple Algorithm for the Computation of the Spectral Distribution of the Solar Irradiance at Surface, 2018, https://doi.org/10.13140/RG.2.2.17025.76648.