

Article

Global Dataset of Solar Power Plants: Multidimensional Integration and Analysis

Anibal Mantilla-Guerra ¹, Christian Mejia-Escobar ², Jorge Azorin-Lopez ³ and Jose Garcia-Rodriguez ^{3,*}

¹ Department of Mining, Central University of Ecuador, Quito 170129, Ecuador; armantilla@uce.edu.ec

² Department of Geology, Central University of Ecuador, Quito 170129, Ecuador; cimejia@uce.edu.ec

³ Department of Computer Technology, University of Alicante, 03690 Alicante, Spain; jazorin@ua.es

* Correspondence: jgarcia@dtic.ua.es

Abstract

The global expansion of renewable energy has increased the strategic importance of photovoltaic (PV) power plants and the demand for comprehensive, high-quality solar datasets to support energy planning, optimization, and data-driven applications. However, existing datasets are often constrained by limited attribute integration, incomplete information, and insufficient global coverage. This study presents a standardized, reliable, and multi-dimensional global dataset for photovoltaic power plant analysis, together with a largely reproducible and documented methodology that automates the collection, generation, and integration of heterogeneous solar-related data from multiple sources. Using this methodology, 27 geographic, topographic, logistical, climatic, and power-related attributes were integrated into a unified dataset comprising 58,978 photovoltaic plant records worldwide. Descriptive statistical analyses were performed to characterize the dataset and assess its informational richness and consistency. The results demonstrate that both the proposed methodology and the resulting dataset provide a robust foundation for photovoltaic energy research and decision-making across diverse application domains. By making this resource publicly available, this work facilitates reproducible research and supports the development of advanced analytical, predictive, and optimization models for academic, industrial, and policy-oriented applications.

Keywords: solar power plants; photovoltaic energy; dataset; descriptive statistics

1. Introduction

Global energy demand continues to increase, while the environmental impacts associated with fossil fuel consumption have intensified the need for cleaner and more sustainable energy sources. Among the available renewable technologies, solar energy has emerged as a cornerstone of the global energy transition due to its abundant availability, technological maturity, and continuous cost reductions. Consequently, photovoltaic (PV) systems have experienced rapid worldwide deployment and now play a fundamental role in advancing sustainable energy generation [1,2].

Among solar energy technologies, photovoltaic (PV) power plants are the most widely deployed. They convert solar radiation, an abundant and renewable energy resource, into clean electricity [3]. Owing to rapid technological advances and the continuous decline in manufacturing costs, PV technology has become a mature and economically viable solution [4]. As photovoltaic infrastructure continues to expand worldwide, access to comprehensive and reliable data on solar power plants has become increasingly important for energy planning, infrastructure management, and the diversification of the energy matrix.



Academic Editors: Guangtan Huang,
Youqiang Liao, Haizeng Pan and
Zheng Liu

Received: 3 June 2026

Revised: 4 July 2026

Accepted: 9 July 2026

Published: 13 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Despite the growing availability of data on solar energy installations, existing datasets remain limited in terms of completeness, consistency, geographic coverage, and attribute integration [5]. Many are fragmented across independent public and private sources, making their combined use difficult. Moreover, most do not consolidate the diverse geographic, technical, environmental, and operational attributes required for comprehensive photovoltaic plant characterization and advanced solar energy analyses within a unified framework. This lack of integrated, standardized, and globally representative information constitutes a significant scientific and technical gap that motivates the present study.

The objective of this study is to develop a high-quality global dataset of photovoltaic power plants by integrating geographic, topographic, climatic, logistical, and power-related attributes within a unified framework. These attributes provide a comprehensive characterization of photovoltaic infrastructure and support a wide range of applications, including solar resource assessment, strategic energy planning, investment risk reduction, site selection and optimization [6–9], public and private decision-making, scientific research, and the development of artificial intelligence models [10,11].

Developing a comprehensive global photovoltaic dataset involves significant technical and methodological challenges. The rapid expansion of photovoltaic infrastructure has increased the complexity of collecting, harmonizing, cleaning, and integrating heterogeneous data into a unified repository. To address these challenges, the selection of dataset attributes was guided by a systematic evaluation of their scientific relevance and technical feasibility, enabling the integration of complementary information from multiple data sources. The resulting data construction workflow is fully reproducible and supported by computational tools that automate key processing tasks, substantially reducing the time and effort required for dataset development.

Geographic Information Systems (GISs) provide the geospatial capabilities required to generate and integrate topographic and land suitability information through the processing of Digital Elevation Models (DEMs) [12,13]. Complementarily, the Python programming language supports the automation of data extraction and retrieval through application programming interfaces (APIs), as well as the generation of derived attributes whose availability is limited or nonexistent in existing sources. These include variables related to solar radiation, climate, plant area, and proximity to road infrastructure [14].

The proposed methodology automates a substantial portion of the data collection, processing, and integration workflow. Using this approach, a global dataset comprising 58,978 photovoltaic power plant records was constructed, with each record integrating 27 geographic, topographic, logistical, climatic, and power-related attributes. The resulting dataset provides a comprehensive, consistent, and reproducible resource with broad global coverage, offering a valuable foundation for photovoltaic energy research and data-driven applications.

The remainder of this manuscript is organized as follows. Section 1 introduces the study and its motivation. Section 2 details the proposed dataset and the automated methodology developed for its construction, and Section 3 presents a comprehensive statistical analysis of its attributes. Section 4 discusses the principal findings, their implications, and limitations. Finally, Section 5 summarizes the main conclusions and identifies directions for future research.

2. Materials and Methods

The project focuses on creating a high-quality dataset of solar photovoltaic plants worldwide. The methodology used to create the dataset is illustrated in Figure 1. This methodology consists of five main stages: a literature review to identify information and data sources; the definition of attributes; the structuring of the dataset; and the collection

and generation of data covering various aspects of solar power plants. Finally, the integration and cleaning of this data into a single dataset are performed.

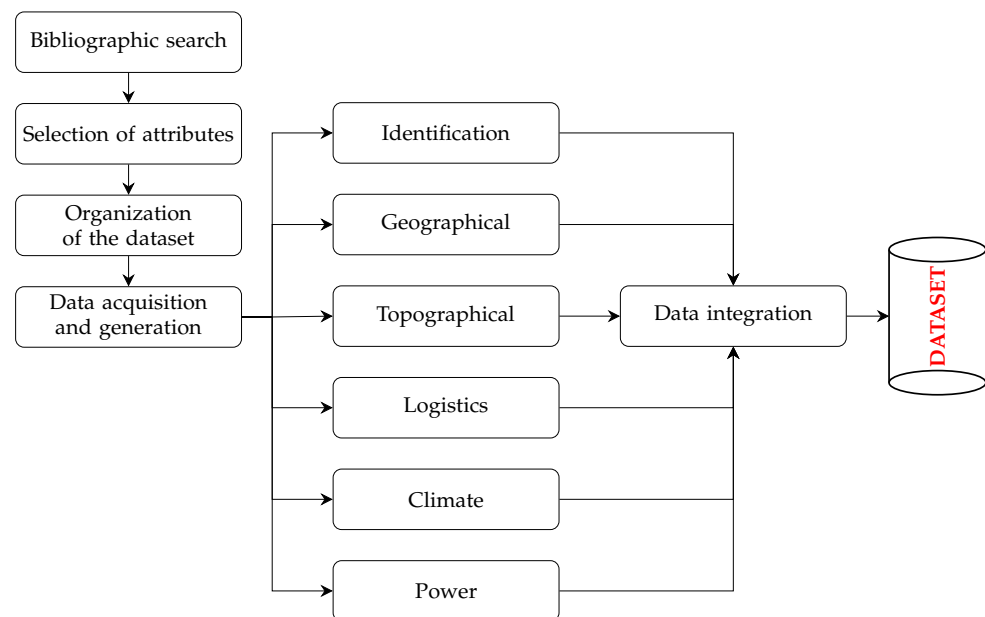


Figure 1. Methodology used to create the multidimensional dataset of solar plants.

The final product successfully combines various types of variables into a single, coherent dataset that integrates both quantitative and qualitative attributes, with its quality rooted in the methodological rigor of the data collection, cleaning, labeling, and validation phases. The importance of creating accurate and up-to-date registries of solar installations in various regions is highlighted, particularly in emerging markets where the potential for solar energy is enormous. The following describes each of the steps required to obtain the final product.

2.1. Bibliographic Search

The goal is to create a dataset that is theoretically robust and technically feasible. By considering this dual approach—what science tells us is necessary and what technical reality tells us is feasible to extract—we ensure that the resulting dataset not only contains the scientifically necessary variables but is also technically feasible to implement and replicate using existing global data infrastructures. Consequently, the initial phase of the project involved a thorough search for high-impact research with scientific credibility, as well as the main existing data sources on solar power plants.

First, AI-assisted research tools such as STORM [15] and Consensus [16] were used, which enabled the filtering of the existing literature based on keyword searches such as “dataset, solar panels, photovoltaic, power, and AI”. The results obtained excluded informal websites, opinion pieces, and studies on other types of renewable energy. The purpose of this broad-ranging search was to identify common patterns in reputable research and prioritize those studies that the scientific community considers critical. Only articles published in the last 10 years in journals indexed in databases such as Elsevier and WoS, as well as academic repositories such as Google Scholar, were included.

To ensure the transparency of the literature review, the search process was documented. Conventional keyword searches yielded 49 publications, while the STORM and Consensus tools returned 49 and 57 results, respectively. After a thorough manual review by the authors to assess the relevance of the texts to the study’s objectives, 8 documents from Consensus were preselected, and the final set of references analyzed was finalized.

Second, the search for available data sources focused on the Web. Since its emergence in the 1990s, the volume of available information has covered every subject. On this global platform, information about solar power plants, energy resources, and geospatial conditions has grown exponentially. Our review of the literature on the datasets most commonly used in climate studies and solar projects is summarized in Table 1.

Table 1. Summary of solar data sources identified during the literature review.

No.	Organization	Access	Main Variables	Period	Applications
1	NASA POWER [17]	Free	GHI, DNI, DHI, temp., wind, clouds	1983–current	Preliminaries, AI
2	PVGIS [18]	Free	GHI, DNI, diffuse, albedo, temp.	2005–current	PV design, simulations
3	Global Solar Atlas [19]	Free/Paid	GHI, DNI, diffuse, temp., production	>20 years	Solar plants
4	SolarGIS [20]	Paid	GHI, DNI, weather, dirt, deterioration	>20 years	Engineering, digital twins
5	Meteonorm [21]	Paid (demo)	GHI, DNI, weather	Sintetic	Filling gaps, simulation
6	ERA5 (Copernicus) [22]	Free	GHI, clouds, wind, temp.	1979–current	Predictive models, AI
7	MERRA-2 (NASA) [23]	Free	GHI, clouds, aerosols	1980–current	AI, weather
8	CERES (NASA) [24]	Free	Incoming/outgoing radiation, clouds	2000–current	Energy balance
9	HelioClim (HC3/HC4) [25]	Free/Paid	GHI, DNI	2004–current	PV homogeneous series
10	BSRN [26]	Free	Surface radiation	1992–current	High precision
11	SURFRAD (NOAA) [27]	Free	Radiation, weather	1995–current	Accurate measurement

The identified solar data sources include well-known organizations in the fields of solar energy, meteorology, and climate modeling. Depending on their applications, three main categories can be distinguished: (1) satellite data sources (NASA, MERRA-2, CERES, HelioClim) and climate reconstructions or reanalyses (ERA5); (2) ground-based measurement networks (BSRN, SURFRAD); and (3) simulation and estimation tools for solar engineering (PVGIS, Global Solar Atlas, SolarGIS, and Meteonorm).

Although there is a close relationship between these organizations, they differ in the aspects they measure, the variables they consider, and the level of accuracy of the information they provide. Given the diverse nature of energy data sources, none can be considered comprehensive on its own. Each has a specific purpose: some offer excellent radiation resolution but lack environmental data, while others provide global climate data but have poor solar accuracy.

Combining these different sources tends to yield better results than using just one. Although it is possible to extract a variety of information, combining it involves dealing with issues such as inconsistencies between global satellite data and point measurements on the ground; in spatial resolutions; and in temporal standardization, as well as differences across geographic regions, among others. Therefore, not only is it justified to create a dataset that brings together the most representative variables in order to eliminate individual biases from each source and provide high-fidelity data for decision-making, but the uniqueness of this work lies in the simultaneous consolidation of geographic, topographic, logistical, climatic, and power variables into a single global repository, an integration that is rarely found in the current literature.

2.2. Selection of Attributes

Selecting the most appropriate variables to build a robust dataset is a task that researchers undertake based on their specific needs. In this case, the selection was made through a two-phase systematic review. On the one hand, a comparative analysis of the documents obtained from the literature review revealed that various variables have been proposed in the literature over time. For this reason, priority was given to those with the greatest academic consensus, relevance, and frequency of occurrence in the sources consulted. On the other hand, the most widely recognized data sources were analyzed to determine which variables are actually available. By cross-referencing both analysis, critical attributes both numerical and categorical were identified and grouped according to techni-

cal criteria into the following key areas: identification, geographic location, topography, logistics, climate, and power.

The inclusion of identifying attributes is justified because they serve as the basis for uniquely and clearly distinguishing each solar plant within the dataset. Geographic attributes make it possible to understand the spatial variability of solar resources, visualize and verify them, and integrate data from other sources into each plant. Topographic attributes describe the characteristics of the terrain that determine exposure to solar radiation and the suitability of a site for the installation of solar power plants, directly influencing energy generation potential, facility design, and project feasibility. Logistics attributes describe the infrastructure and accessibility conditions of the surrounding area that influence the feasibility and construction of solar power plants. Climate variables are essential for assessing the availability of energy resources and the conditions that affect the efficiency of solar modules. Finally, power-related attributes make it possible to correlate a site's potential with its actual generation capacity.

2.3. Organization of the Dataset

Table 2 lists the 27 attributes that compose the proposed dataset. For each attribute, the table specifies the assigned name, category, variable type, unit of measurement, and a brief description of its meaning.

Table 2. Structure of the dataset of solar plants and main characteristics of its attributes.

Attribute	Category	Type	Units	Description
Code	Identification	Qualitative	-	Unique alphanumeric identifier for the plant
Name		Qualitative	-	Denomination of the solar power plant
Country	Geographical	Qualitative	-	National territorial entity where the plant is located
Latitude		Quantitative	Grades	Geographic coordinates associated with parallels
Longitude		Quantitative	Grades	Geographic coordinate associated with meridians
Elevation	Topographical	Quantitative	m	Vertical height of the solar plant
Slope		Quantitative	Grades	Angle of inclination of the terrain
Aspect		Quantitative	Grades	Direction of the slope
Curvature		Quantitative	-	The shape of the terrain
T_slope		Qualitative	-	Classification of slope
T_aspect		Qualitative	-	Classification of aspect
T_curvature		Qualitative	-	Classification of curvature
Distance to road	Logistics	Quantitative	m	Distance from the solar plant to the nearest road
Area		Quantitative	m ²	Total extension of the solar plant
Size		Qualitative	-	Classification of area
Temperature	Climate	Quantitative	°C	Average ambient air temperature at the site
Humidity		Quantitative	%	Water vapor content in the atmosphere
Wind speed		Quantitative	m/s	Airflow velocity at the site
Wind direction		Quantitative	Grades	The cardinal point from which the wind blows
DT_wind	Power	Qualitative	-	Classification of wind direction
Operational status		Qualitative	-	Current phase of the project lifecycle
Solar aptitude		Quantitative	-	A site suitability score for a solar power plant
T_aptitude		Qualitative	-	Classification of solar aptitude
Optimal tilt		Quantitative	Grades	Ideal solar panel angle
Global irradiation		Quantitative	kWh/m ²	Energy incident on a surface over a period of time
Capacity		Quantitative	kWp	Total installed power
PV potential		Quantitative	kWh/kWp	Amount of energy generated per kWp

An innovative feature of the project is the calculation of the area and solar aptitude index for each PV plant. These attributes are considered extremely valuable and are not typically found in general information sources. In addition, to facilitate the analysis and

interpretation of certain numerical variables (area, slope, aspect, curvature, wind, and solar aptitude), their respective categorical versions were created based on these variables (size, T_slope, T_aspect, T_curvature, DT_wind, and T_apitude).

Once the variables that form the dataset have been specified, the dataset is organized into a tabular format. In this structure, each column corresponds to a predefined attribute, while each row represents a registered solar plant instance. The next step is to obtain the data needed to complete that structure, which involves collecting the actual values for each variable. In some cases, these values can be obtained directly from available data sources, while in others, they must be generated during the dataset creation process.

2.4. Data Acquisition and Generation

This section explains where the data was obtained and how it was collected. Effective data collection depends on the integration of various datasets from different sources, such as government agencies and private organizations. Table 3 presents the main sources for viewing and downloading the data for the variables included in the dataset, as well as the method used to obtain that data.

Table 3. Data sources for the variables included in the dataset of solar power plants.

Source of Information	Data Format	Variable	Method
Global Energy Monitor (GEM)	Data downloaded in CSV format	Name	Direct extraction
		Country	Direct extraction
		Latitude	Direct extraction
		Longitude	Direct extraction
		Operational status	Direct extraction
		Capacity	Direct extraction
Global Solar Atlas (GSA)	Raster file in GeoTiff format	Elevation	Direct extraction
		Optimal tilt	Direct extraction
		Global irradiation	Direct extraction
		PV potential	Direct extraction
NASA POWER & ERA5 (Copernicus)	CSV, raster and NetCDF (matrix)	Temperature	Direct extraction
		Humidity	Direct extraction
		Wind speed	Direct extraction
		Wind direction	Direct extraction
Our own generation	Digital Elevation Models (DEMs), raster layers, vector layers, road layers from GEOFABRIK y OpenStreetMap (OSM)	Code	Manual assignment
		Slope	Geospatial calculation
		Aspect	Geospatial calculation
		Curvature	Geospatial calculation
		T_slope	Classification
		T_aspect	Classification
		T_curvature	Classification
		Distance to road	Geometric calculation
		Area	Geometric calculation
		Size	Classification
		DT_wind	Classification
		Solar aptitude	Geospatial calculation
		T_apitude	Classification

All of the sources mentioned are open access and useful for research, planning, and decision-making. The following section describes the data collection process by category: identification, geographic location, terrain characteristics, logistics, climate, and operational characteristics of solar power plants.

2.4.1. Identification Data

The identification data includes the *code* and the *name* of the solar power plant. Each plant is assigned a unique alphanumeric code generated as part of the dataset creation process. It consists of three elements: a unique integer value (ObjectID) that the GIS automatically assigns to each record (row) in the table, the first three letters of the plant's country of origin, and the first letter of the plant's size category (P, M, or G). This design

provides a structured and informative identifier that makes it easier to organize, search, and analyze the dataset. For example, the code “16056-ECU-P” identifies a small-scale solar power plant in Ecuador called “Intiyana Solar Farm”.

The name of each solar plant corresponds to the official designation of the PV project on the website of Global Energy Monitor (GEM), a global reference for the world’s energy infrastructure, covering both fossil fuels and renewable energy [28]. Through its platform and interactive tools, users can view the geographic distribution of energy infrastructure and other key data. The website provides an interactive world map (Figure A1) that allows users to view high-resolution solar data for any location on the planet. GIS data layers can be downloaded for analysis in specialized software programs such as QGIS or ArcGIS.

Although the vast majority of solar plant names come from GEM, in specific cases, data from the Global Solar Atlas (GSA) web platform was used (Figure A2). This is because some names in GEM were encoded in such a way that they are unreadable. GSA is a free online tool developed by the World Bank Group and the company Solargis. It is a key resource for energy planning, research, and, in particular, for identifying the optimal locations for solar power plants worldwide [19].

2.4.2. Geographical Data

The data on *country*, *latitude*, and *longitude* for each solar plant were extracted directly from the GEM portal, under the “Global Solar Power Tracker” option, which is a tracker that collects information on energy projects around the world [29]. Here, users are asked to fill out a basic form (name, institutional email address, and purpose of use), which allows them to download the solar plant inventory. Once access was approved, the system provided the dataset files in standard CSV, Shapefile (SHP), and GeoJSON formats. These files were integrated into the GIS by adding vector layers (SHP and GeoJSON) and delimited text files (CSV), defining the latitude and longitude columns, and assigning the WGS 84 (EPSG:4326) spatial reference system to ensure their correct location. Next, the georeferencing and spatial consistency of the points were verified, generating the resulting layer shown in Figure 2.

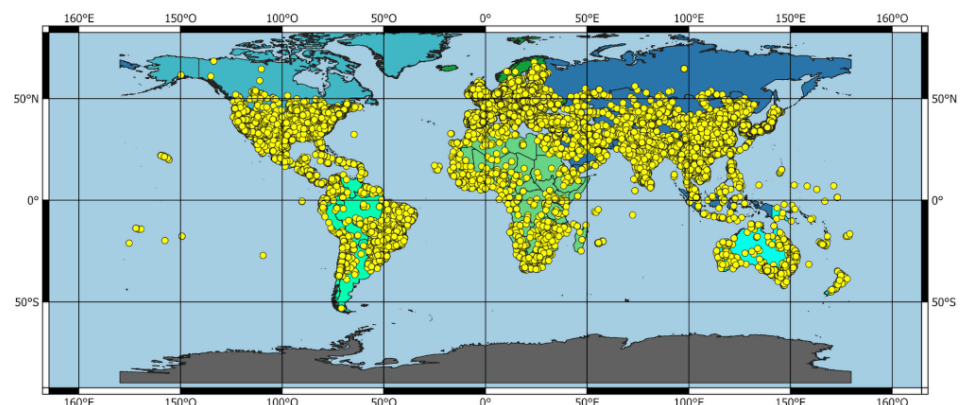


Figure 2. Global location map of the solar plants in the dataset.

2.4.3. Topographical Data

These factors are critical for determining the optimal location for solar power plants. These variables are mathematically derived from a Digital Elevation Model (DEM). A DEM is a three-dimensional representation of the Earth’s surface in the form of a regular pixel-based grid, where each cell or pixel contains an elevation value. As a result, raster maps of *elevation*, *shading*, *slope*, *aspect*, and *curvature* were generated, all of which retain this same matrix format.

The DEM is downloaded directly from the GSA platform in GeoTIFF format, which is compatible with GIS environments and has a spatial resolution of 9 arc-seconds appropriate

for the study's objectives. The coverage area is organized by continent, so we will use South America as an example. Before using the DEM, it is essential to prepare it properly to perform the required spatial analyses by following these steps in ArcMap:

1. Reproject to a coordinate system in meters since the DEMs were originally in WGS84 (degrees). The "South America Albers Equal Area Conic" projection was chosen, which not only allows for the standardization of units of measurement for distance calculations but as an equivalent projection also preserves the proportionality of areas across the entire continent. This ensures that the orientation of the slopes and the surface area of the land identified as suitable for the installation of solar panels are geographically reliable.
2. Verify the integrity of the DEM by identifying the following spatial inconsistencies:
 - Irregular patterns such as artificially flat surfaces, lines or blocks with no data, and abrupt changes in elevation that do not correspond to the actual terrain.
 - Check the value assigned to "NoData" using the Source tab of the raster. If there is no suitable mask, generate an explicit mask using tools such as "Set Null" or the raster editor, thereby ensuring that the valid areas are correctly delineated.
 - Correct artificial depressions or peaks using the Fill tool, as these errors can disrupt the data flow and produce incorrect results (Spatial Analyst → Hydrology → Fill). This is shown in Figure 3.



Figure 3. Fill correction.

- Apply moderate smoothing if necessary: Filter (Low Pass) and Focal Statistics (mean, median).
 - Adjust the Z-factor to ensure consistency between horizontal and vertical units; typically, set it to 1 when working in meters. The Z-factor affects the calculation of slope, curvature, aspect, and shading.
3. Finally, the DEM should be clipped to the study area to improve performance and ensure spatial consistency with the other layers (Data Management → Raster → Raster Processing → Clip). Ensure consistent resolution across multiple layers: Resample if necessary.

Based on the cleaned, continuous, and correctly projected DEM, derived maps are automatically generated with the same pixel resolution of 100×100 m. These maps are generated in ArcGIS v10.3 using the "Terrain Analysis" tool from the "Spatial Analyst" module. The resulting raster maps are shown in Figure 4.

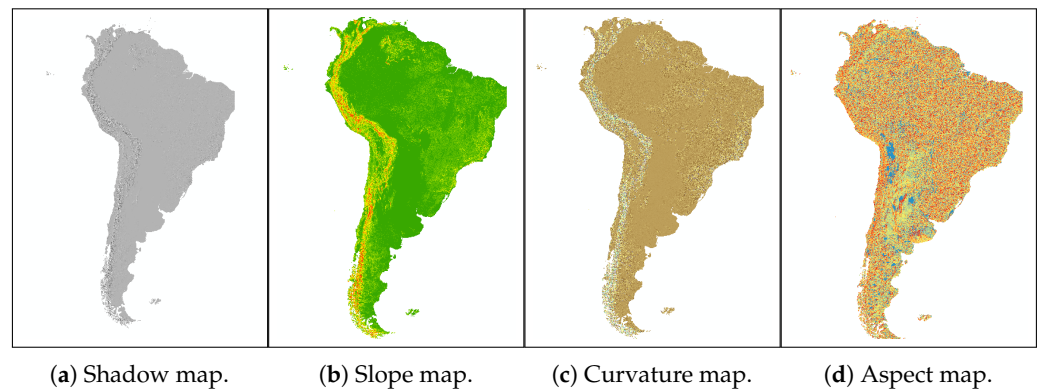


Figure 4. Raster maps of topographic variables derived from the DEM of South America.

The shadow map highlights the contours of the terrain, such as mountains and valleys, using standard lighting parameters (azimuth 315° and elevation 45°). The slope map shows the incline of the terrain expressed in degrees or as a percentage, which is essential for land suitability analysis. The curvature map depicts convex surfaces associated with ridges or watersheds (positive values), concave surfaces associated with areas of accumulation or valleys (negative values), and flat or transitional areas (values close to zero). The aspect map determines the orientation of slopes on a scale from 0° to 360° , assigning a value of -1 to completely flat areas; this is important because exposure to sunlight and wind directly affects the performance of a solar power plant.

The above maps are essential tools for identifying areas with PV potential, as they allow for an accurate assessment of the topographical conditions that directly influence solar energy capture. Each map provides additional information, and when combined, they help identify optimal, safe, and efficient locations for installing solar panels. The maps described represent variables that, in their original form, are numerical. However, they were converted into categorical variables to facilitate their interpretation and statistical analysis. This allows continuous values to be grouped into representative classes, enabling the identification of patterns and comparisons between solar power plants.

The qualitative variable named T_{slope} is generated from the slope classification, which ranges from $0-2^\circ$ (Flat or nearly flat), $2-5^\circ$ (Smooth), $5-10^\circ$ (Moderate), $10-15^\circ$ (Steep), $15-30^\circ$ (Very steep) up to $\geq 30^\circ$ (Steep or abrupt). The slope ranges used are consistent with FAO recommendations for classifying slopes in soil and land-use studies [30,31], allowing the slope of the terrain to be correlated with factors such as erosion, stability, and land-use potential, thereby facilitating their application in solar suitability studies and land-use planning.

The classification of aspect generates the T_{aspect} variable, which has 8 cardinal directions (N, NE, E, SE, S, SW, W, NW), with a special category for flat slopes. This categorization is based on the convention of dividing the 360° of a circle into 45° sectors, which simplifies the interpretation of terrain orientation; it is widely used in geomorphology, hydrology, and terrain analysis in GIS.

Finally, the terrain curvature classification ($T_{curvature}$) consists of three categories: flat or intermediate surfaces (values close to zero, between -0.01 and 0.01), concave surfaces or valleys (negative values), and convex surfaces or ridges (positive values). This interpretation of curvature is part of the methodology described by Wilson and Gallant [32] for consistently distinguishing between flat areas, depressions, and elevations.

2.4.4. Logistics Data

Data on *distance to road*, *area*, and *size* were included to describe the accessibility and extent of the solar plants. The generation of this data through integration and calculation processes constitutes the main contribution of this work, as it provides information that is not available in existing datasets or sources.

Firstly, the distance from the solar plants to the main road network was calculated as follows:

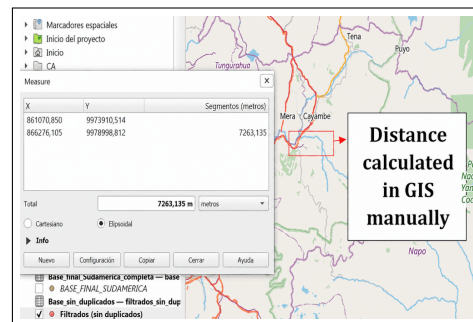
1. Download road maps of South America in a vector file (shapefile) from the GEO-FABRIK repository, a service that facilitates the extraction of OSM road layers at the regional level [33].
2. Filter these layers in QGIS v3.40 using SQL queries, selecting only first- and second-class roads.
3. Merge the resulting vector files into a single layer or shapefile, which is then integrated into the solar plant inventory.
4. A topological correction is performed using the “Correct Geometries” algorithm applied to the road network to correct digitization errors (discontinuities, gaps, etc.).
5. Reproject both layers (roads and plants) to the “South America Albers Equal Area (EPSG:102033)” reference system, ensuring a metric projection at the continental level.
6. Run the “Near” tool in ArcMap v10.3 to identify the nearest neighbor between the two layers, generating the final vector layer that contains the distance in meters.

Thus, the Euclidean distance from the nearest road point to a solar plant was calculated. This result is added directly to the dataset’s attribute table. During this process, the calculation of terrain variables was limited exclusively to the sites located within the coverage area of the generated DEMs. Consequently, the distance to the road network was calculated in a generic manner; however, by using an interactive web platform that integrates these variables, we were able to validate our results and achieve greater accuracy in calculating road distances (Figure 5a,b).

In addition, calculating the *area* required geolocating each PV plant using its latitude and longitude coordinates in order to generate polygons that would allow for an estimate of its surface area. Initially, polygons were manually digitized in a GIS environment by visually delineating the areas corresponding to solar installations. However, this time-consuming and labor-intensive process led to the development of an automated procedure using Python v3.12 code that, based on geographic location (latitude/longitude) and by querying the OSM API, identified polygons tagged as solar plants and automatically calculated their area using a metric reference system. The following steps were implemented:

1. Preparation of the initial data, starting with an Excel file containing the geographic coordinates of the plants, along with information on the country, latitude, and longitude.
2. Geographic classification by mapping countries to continents using a logical dictionary to organize the data by continent.
3. Download geospatial data stored in OpenStreetMap PBF files [34] for each continent from GEOFABRIK.
4. Filter relevant information using the “osmium-tool” utility to extract only those features related to solar power plants from OSM data.
5. Based on the filtered data, geographic data for solar power plants by continent is generated in a GeoJSON file.
6. Geographic filtering (bounding box) to retrieve solar plant data for each country. As a result, a country-specific GeoJSON file of solar power plants is generated.
7. The areas of the polygons are calculated (using geodetic methods) to determine the surface area of each solar panel.

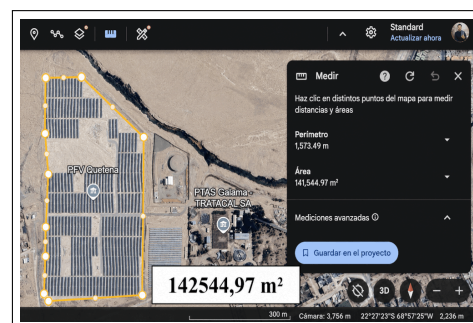
8. A spatial join is performed to link the points in Excel (plant locations) with the solar plant polygons obtained from OSM.
9. Finally, an Excel file is generated containing the consolidated information, including the calculated area and the status of each plant.



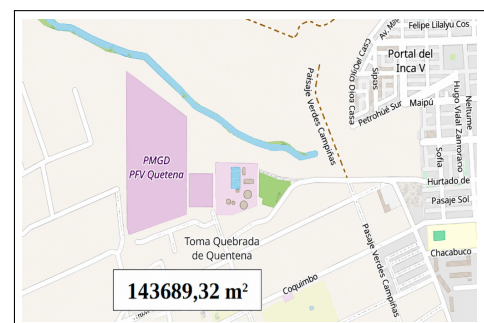
(a) Manual distance to road measurement.

ID	Nombre	Area (m²)	Distancia (m)
1	Plant 1	10000	100
2	Plant 2	20000	200
3	Plant 3	30000	300
4	Plant 4	40000	400
5	Plant 5	50000	500
6	Plant 6	60000	600
7	Plant 7	70000	700
8	Plant 8	80000	800
9	Plant 9	90000	900
10	Plant 10	100000	1000
11	Plant 11	110000	1100
12	Plant 12	120000	1200
13	Plant 13	130000	1300
14	Plant 14	140000	1400
15	Plant 15	150000	1500
16	Plant 16	160000	1600
17	Plant 17	170000	1700
18	Plant 18	180000	1800
19	Plant 19	190000	1900
20	Plant 20	200000	2000
21	Plant 21	210000	2100
22	Plant 22	220000	2200
23	Plant 23	230000	2300
24	Plant 24	240000	2400
25	Plant 25	250000	2500
26	Plant 26	260000	2600
27	Plant 27	270000	2700
28	Plant 28	280000	2800
29	Plant 29	290000	2900
30	Plant 30	300000	3000
31	Plant 31	310000	3100
32	Plant 32	320000	3200
33	Plant 33	330000	3300
34	Plant 34	340000	3400
35	Plant 35	350000	3500
36	Plant 36	360000	3600
37	Plant 37	370000	3700
38	Plant 38	380000	3800
39	Plant 39	390000	3900
40	Plant 40	400000	4000
41	Plant 41	410000	4100
42	Plant 42	420000	4200
43	Plant 43	430000	4300
44	Plant 44	440000	4400
45	Plant 45	450000	4500
46	Plant 46	460000	4600
47	Plant 47	470000	4700
48	Plant 48	480000	4800
49	Plant 49	490000	4900
50	Plant 50	500000	5000

(b) Registered calculated distance.



(c) Area digitized manually.



(d) Area calculated automatically.

Figure 5. Calculation of the distance to road and the area of a solar plant.

Automating this calculation made it possible to systematically determine the areas of the polygons for the rest of the world and, at the same time, served as a means of verifying the values previously obtained using the manual method, since there should be no significant differences between the two procedures (Figure 5c,d).

Finally, the classification of solar power plants by size into small (<20,000 m²), medium (20,000 to 200,000 m²), and large (>200,000 m²) in this study has been adapted from industry standards through an empirical evaluation of the global dataset distribution. Traditional industry frameworks employ fixed thresholds based strictly on economic viability and operational purpose. Small-scale installations designed for residential and commercial self-consumption typically cover areas of less than 2 hectares (<20,000 m²), whereas medium-sized solar farms range from 2 to 5 hectares (20,000 to 50,000 m²) and larger, high-profitability projects cover larger areas, exceeding 5 hectares (>50,000 per m²) [35]. Our approach is intended to more accurately reflect the observed dimensional reality and operational capabilities recorded worldwide, where the scale of medium and large projects exceeds conventional theoretical standards.

2.4.5. Climate Data

This type of data is important because it allows us to assess the environmental conditions that affect the performance of solar installations. The variables considered are *temperature*, *humidity*, *wind speed*, and *wind direction*. These data were obtained from NASA's Prediction of Worldwide Energy Resources (POWER) [17], a NASA initiative of high scientific quality that provides free access to meteorological and solar resource data. The web interface of this platform (Figure A3) features an interactive viewer called "POWER

Data Access Viewer" (DAV), which allows users to select a specific geographic location (latitude/longitude) or a region, specify a time period and variables of interest, view the data on maps or in graphs, and download the data in CSV format.

Initially, data collection was carried out by manually extracting the necessary parameters from the annual averages. However, to optimize the process and ensure continuous spatial coverage without loss of information, the analysis was scaled up to utilize data from the ERA5 reanalysis [22], available in the Climate Data Store (CDS) of the European Union's Copernicus program. Using the "Datasets" option, we downloaded "ERA5 monthly averaged data on single levels," which provides global monthly averages of atmospheric variables with continuous spatial coverage and temporal consistency.

This change in methodology made it possible to transition from individual queries to bulk downloads of global monthly averages in NetCDF (Network Common Data Format), using the subregion extraction option and defining global limits of latitude (-90° to 90°) and longitude (-180° to 180°) to ensure complete coverage of the study area and avoid loss of spatial information. After importing these files into QGIS, the base variables (air temperature at 2 m, dew point temperature at 2 m, and the zonal and meridional wind components at 10 m [u10 and v10]) were processed, and relative humidity, wind speed, and wind direction were calculated using raster operations (map algebra) in QGIS. Previously, monthly averages were calculated to obtain representative annual values. However, it is essential to recognize that the spatial resolution of the ERA5 reanalysis (31 km grid) introduces a structural scale mismatch for site-specific or pixel-level assessments of solar installations. This macroscale resolution introduces uncertainties inherent to positional mismatch, especially in areas characterized by complex topography. Consequently, the climate variables provided here should be interpreted as macro-environmental indicators rather than site-specific physical measurements, a standard margin of uncertainty that is widely accepted in the literature on global-scale energy resource mapping.

Finally, based on wind direction, the categorical variable *DT_wind* was created, which was classified into eight categories corresponding to the cardinal directions and the main intermediate points (N, NE, E, SE, S, SW, W, NW), each representing a 45° sector on the compass.

2.4.6. Power Data

This section examines the *status* of the solar project, i.e., whether it is operational, under construction, proposed, canceled, or on hold. This data is obtained directly from a CSV file of solar plants worldwide downloaded from Global Energy Monitor.

A key contribution of this study is to calculate the *solar aptitude* index (I_{AS}), which is an indicator of great importance for assessing the suitability of an area for the installation of solar energy systems. Raster layers derived from DEMs are used: slope, aspect, shading, and terrain curvature. The degree of importance of each variable is represented by the weights (0.40, 0.25, 0.20, 0.15), since the slope affects both physical feasibility and cost, while aspect influences solar radiation gain, and shading and curvature are secondary morphological factors. These weights were determined based on the opinions of experts in geological engineering and electrical and electronic engineering, using as a conceptual basis the Analytic Hierarchy Process (AHP), which is one of the most widely used multi-criteria decision-making methods in the specialized literature [36,37]. These layers are imported into the GIS and standardized to the same reference system and spatial resolution. They are normalized to a common scale, enabling the comparison and combination of heterogeneous variables. Standardized and weighted layers are integrated using map algebra, by applying a superposition operation of weighted sum (Equation (1)).

$$I_{AS} = (Slope \times 0.40) + (Aspect \times 0.25) + (Hillshade \times 0.20) + (Curvature \times 0.15) \quad (1)$$

The calculated index is classified into suitability categories (low, medium, and high), resulting in a new variable called *T_apptitude*. This makes it easier to interpret spatial data, especially when viewing the solar suitability map, which is a tool designed to identify and classify favorable geographic areas based on their potential for harnessing solar energy. The classification of the solar aptitude index into low (0–0.4), medium (0.4–0.6), and high (0.6–1) categories is based on the standardization of results typical of multi-criteria evaluation methods such as AHP [38]. This approach allows results to be expressed on a scale from 0 to 1 and classified into qualitative categories to facilitate their interpretation. The territory is classified into zones of varying suitability, making it possible to objectively identify the sites with the best conditions for harnessing solar energy, support energy planning, and optimize decision-making for photovoltaic or solar thermal projects.

The *optimal tilt* variable corresponds to the mounting angle that maximizes the output of a fixed PV system oriented toward the equator [39]. In the Solar Atlas, this parameter is estimated using algorithms derived from the Solargis solar model, which evaluate the system's energy output for different tilt angles. Next, the angle that provides the highest energy efficiency over the 2018–2024 period is selected, based on the information available on the official website. This information is represented in raster images, which are imported into the GIS environment. The pixel values corresponding to the optimal tilt data are automatically extracted using a raster-to-point tool based on the exact geographic coordinates of each solar installation, ensuring absolute reproducibility and eliminating human error in processing.

The *global irradiation* is obtained from a GSA raster by selecting the area of interest. This platform provides various metrics, such as direct normal irradiance (DNI), horizontal global irradiance (GHI), horizontal diffuse irradiance (DHI), and global irradiance on an inclined plane (GTI). For this study, we used this last variable, which was overlaid onto the geographic layer of the solar plants using an “Extract Values to Point” operation, allowing us to assign each plant the irradiance value corresponding to its location.

The *capacity* of PV plants is listed in the GEM inventories. The registered value is obtained through a systematic process of collecting and validating public data. Researchers compile information from official sources (government data, permits, and reports from power generation companies), industry reports, and verified media reports, which are then cross-checked against other public and private databases to ensure consistency and quality. Each documented solar plant has a profile page with detailed references [40], and all rated capacities reflect the declared power of the generating units under standard conditions.

Finally, the *PV potential* represents the estimated specific electricity generation for a standard photovoltaic system. The corresponding raster from the Solar Atlas is downloaded, generated using data on solar radiation, panel efficiency, and climate variables. This raster is integrated with the solar plant layer through a spatial overlay operation, which makes it possible to extract and assign a potential value to each installation. This value is useful for feasibility studies and energy planning at the regional and national levels.

2.5. Data Integration

This stage is crucial for consolidating diverse information into a single coherent and consistent structure. This integration process required the unification of variables from different data categories: identification, location, topography, logistics, climate, and operational parameters.

One of the main challenges was to verify that the location points corresponded to actual solar installations. To this end, a cross-validation was performed between GSA and GEM, using ArcGIS and QGIS to overlay the points onto irradiance layers. This process was useful in identifying inconsistencies, such as incorrect coordinates or sites located in

areas unsuitable for PV generation. In addition, illegible or missing GEM names were replaced with the corresponding GSA identifier, when available, and records that remained unresolved after this step were flagged rather than discarded. As a result, it ensured that only actual, correctly georeferenced solar plants were included.

Another challenge was to precisely delineate each solar installation, which is useful for calculating its respective area. First, the spatial and geometric verification of the location was carried out using ArcGIS satellite maps as a cartographic reference. The contour of each plant was then digitized manually (polygonization) using an appropriate zoom level that allowed the geometries to be delineated with sufficient precision. The area was then calculated using ArcMap's geometric calculator.

To ensure scientific rigor and transition the dataset from a descriptive registry to a validated scientific contribution, the area and distance-to-road variables were validated through manual measurements in GIS, using a random sample of 100 cases to verify the precision of the values obtained. The results of the quantitative validation of the area and distance-to-road variables (mean absolute percentage difference between manual and automated measurements) are 3% and 0.1%, respectively. Furthermore, to formalize the limitations inherent in the use of OpenStreetMap, we explicitly acknowledge that the completeness of OSM data and positional accuracy vary substantially across countries and regions due to the heterogeneous nature of crowdsourcing, which can lead to local distortions in areas with limited map coverage.

During the digitization phase, multiple records were identified for the same physical location of solar panels, which resulted in redundancy. For example, in the case of Brazil, a solar installation appeared four times with identical geographic coordinates and descriptive parameters. The identification of this spatial issue was based on the matching of latitude and longitude attributes. Data cleaning was performed using the attribute-based duplicate detection. In addition, spatial outliers outside the coverage area of the rasters were removed using distribution metrics such as "Average Nearest Neighbor".

Once the data had been cleaned, the coordinate system was standardized to WGS 84 by reprojecting the original layers, ensuring spatial compatibility. In addition, the spatial resolution was standardized to 9 arc-seconds. Since the solar resource layers (global, direct, and diffuse irradiance) have this nominal resolution, the complementary layers (photovoltaic potential, terrain elevation, and climate variables) were resampled to match that resolution.

As a complementary step, the solar aptitude map was validated to ensure its spatial coherence and consistency with solar radiation patterns and terrain characteristics. To this end, a comparison was made with the locations of existing solar power plants, revealing that they are concentrated in areas with high solar aptitude.

The actions described above enabled us to compile a single, structured, and consistent dataset with high spatial accuracy and high data quality, which better reflects the characteristics of solar power plants. Table 4 provides a comparison of existing photovoltaic datasets.

Table 4. Comparison of existing PV plant datasets and the proposed dataset.

Dataset	Number of Plants/Sites	Variables Included	Geographic Coverage	Advantage/Limitation
Global Inventory of PV Units [41]	68,661 facilities	Facility location, estimated capacity, installation date, historic land-cover classification, proximity to vulnerable areas	Global	Large-scale global asset inventory. Focuses on commercial, industrial, and utility-scale facilities, not on extensive multidimensional integration of logistical and topographical attributes
Global PV Solar Panel Dataset (2019–2022) [42]	Not reported	Annual distribution of PV panels at 20 m; map showing potential locations, GE imagery, Sentinel-2 imagery, and land cover data	Global	Recent annual update and global spatial detail. Geared toward panel mapping, not toward multidimensional characterization of plants with rich operational or logistical attributes
UK Open PV Installations [43]	Over 260,000 installations	Geographic location, power output, system geometry and metadata, capacity, orientation, tilt, and commissioning date	United Kingdom	Very high coverage, especially of small residential systems. Focused on a single country
China PV Power Plants Map [44]	Not reported; total mapped area 2917 km ²	Polygon boundaries, area, elevation, annual mean temperature, yearly precipitation, population density, EVI, land-cover type	China	Detailed national public dataset on spatial extent. Coverage restricted to China
India AI Solar Locations [45]	1363 solar PV farms	Geospatial Location of Solar Farms and Associated Land Cover	India	Significantly improves the comprehensiveness compared to previous surveys in India. National coverage, not global
GM-SEUS dataset [46]	Over 15,000 arrays and 2.92 million panel-rows	Year, azimuth, technology, panel-row area and dimensions, inter-row spacing, ground cover ratio, tilt, capacity	United States	Rich design metadata at array and panel-row level; limited to ground-mounted US systems
Rooftop PV Hong Kong [47]	60 stations	Inverter-level generation and electrical variables, irradiation, temperature, humidity, visibility, pressure, wind, rain	Hong Kong	High operational and meteorological resolution, useful for benchmarking, fault detection, and degradation monitoring. Local coverage and rooftop systems; no global inventory of plants
UNISOLAR [48]	42 PV sites	PV generation, solar irradiance, weather data, geographical placement, engineering specifications	Victoria, Australia	High level of detail and usefulness for forecasting and benchmarking. Small institutional scale
Proposed dataset	58,978 records	27 attributes covering geography, topography, logistics, climate, and power	Global	Combines multidimensional plant descriptors rarely available together in a single source, with standardized integration, highly reproducible methodology, and global scope

A review of the literature shows that existing PV datasets focus on specific needs. Some prioritize high-resolution operational and meteorological data but cover only a few dozen local sites (Hong Kong and UNISOLAR). Others maximize geographic coverage or the number of installations (U.K. inventory, Kruitwagen’s global inventory, and GM-SEUS) but do not simultaneously integrate a broad set of topographic, logistical, climatic, and power variables into a single, harmonized framework. Global and national mapping efforts also exhibit recurring limitations: a lack of updates or coverage in previous datasets, a focus on a single country, or an emphasis on geometry and spatial distribution rather than on multidimensional plant characterization. In this context, the main contribution of the proposed dataset is that it offers global coverage, 58,978 records, and 27 integrated attributes related to geography, topography, logistics, climate, and energy, thereby overcoming the fragmentation noted in the literature [49].

3. Statistical Analysis of the Dataset

This section presents a statistical characterization of the geographic, topographic, climatic, logistics, and energy-related variables in the created dataset. Identification variables (code and name) are excluded, as these attributes serve only as labels to distinguish each PV plant. In addition, the numerical variables of area, slope, curvature, and aspect were used to generate the corresponding categorical variables: size, slope type, curvature type, and aspect type. This enables more meaningful and understandable statistical comparisons. The statistical analysis is based on descriptive plots and indicators generated using R v4.6.1 and RStudio v.2026.

3.1. Geographical Variables

It is essential to know where solar power plants are located and to understand their geographic distribution. Globally, a small number of countries account for a large proportion of solar power plants. China tops the list with nearly a quarter of all plants, while Japan, the United States, and Germany also account for a significant share (Figure 6a). This trend is mirrored on a continental scale, where Asia, Europe, and North America are the leading regions in terms of PV solar power (Figure 6b). The distribution of longitude is clearly multimodal (Figure 6d), identifying three specific regions for PV development: the European axis (0° to 30° E), the Asian axis (100° to 150° E), and the North American axis (-120° to -70° W). In contrast, latitude (Figure 6c) reveals a massive concentration of data in the Northern Hemisphere. With an average of 34.86° N, most plants are located in the industrial “Solar Belt” (China, the United States, and Southern Europe).

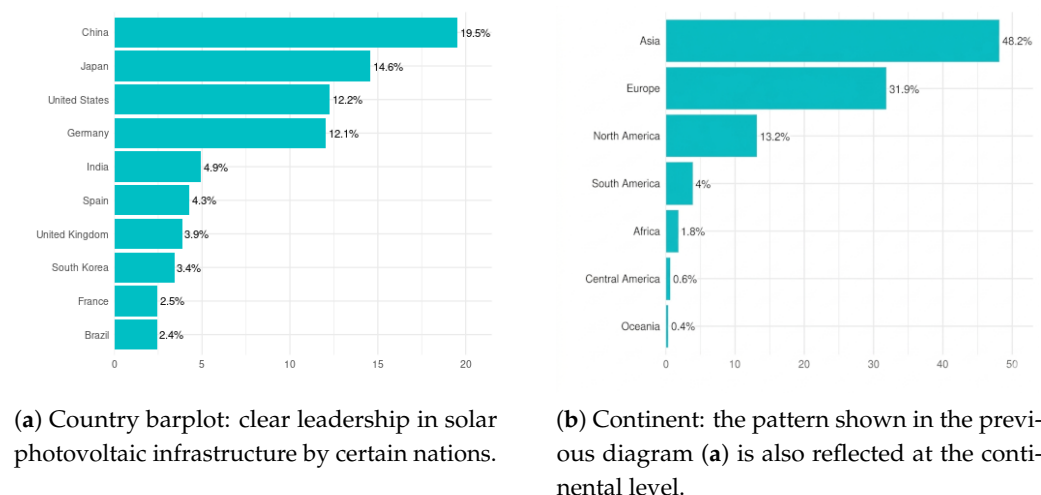
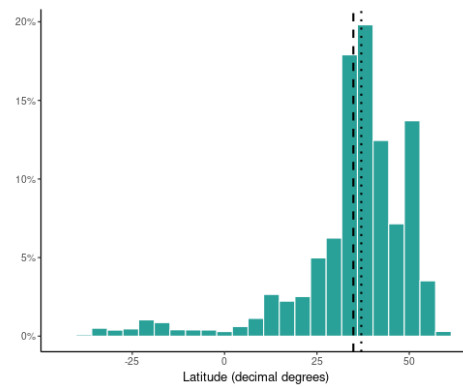
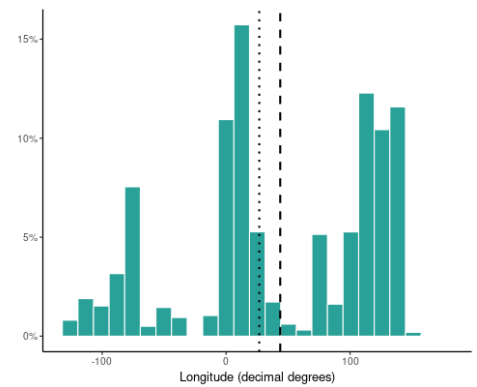


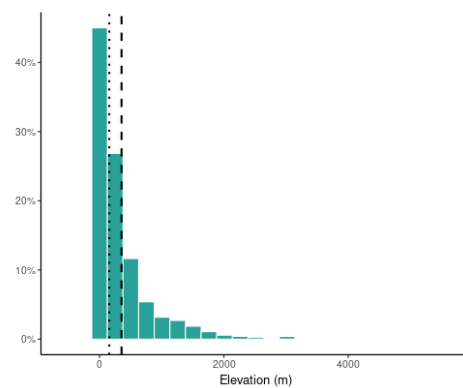
Figure 6. Cont.



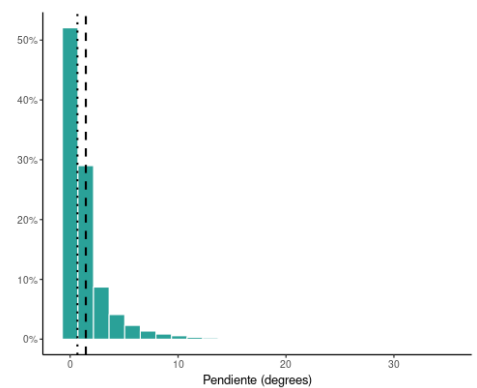
(c) Latitude histogram: min: -41.61 ; max: 59.94 ; mean: 34.86 ; median: 37.01 ; std. dev.: 16.01 .



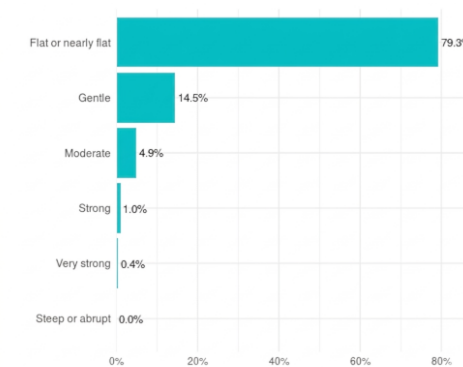
(d) Longitude: min: -124.01 ; max: 177.98 ; mean: 43.68 ; median: 26.76 ; std. dev.: 77.31 .



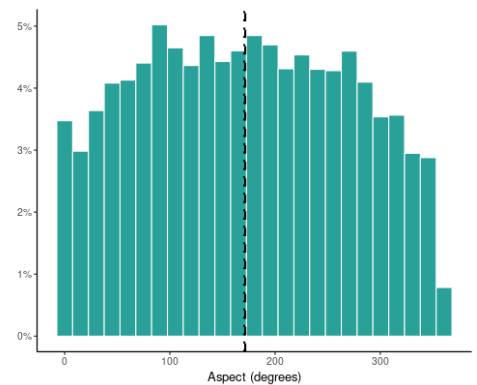
(e) Elevation: min: -378 ; max: 5656 ; mean: 353.62 ; median: 156 ; std. dev.: 526.64 .



(f) Slope: min: 0 ; max: 34.68 ; mean: 1.45 ; median: 0.67 ; std. dev.: 2.26 .

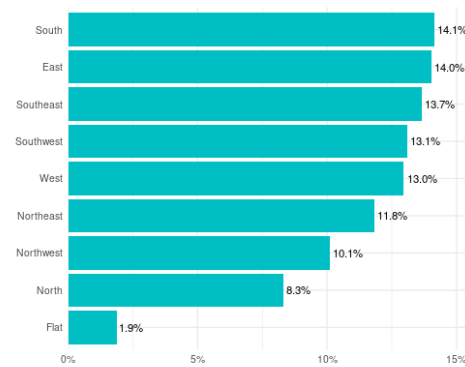


(g) Slope type: a strong preference for flat or slightly sloping terrain for the installation of solar panels.

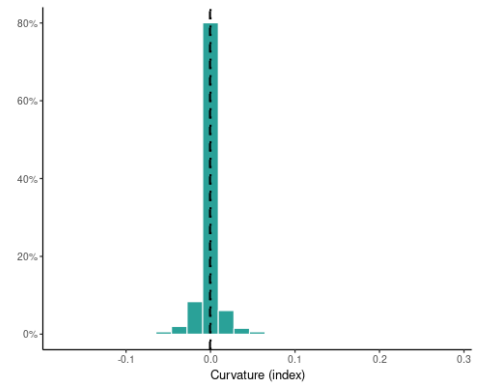


(h) Aspect: min: -1.00 ; max: 359.61 ; mean: 171.59 ; median: 170.67 ; std. dev.: 98.54 .

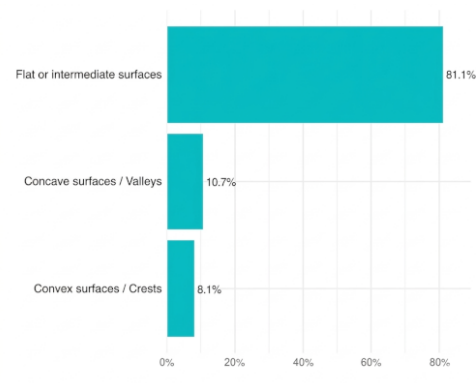
Figure 6. *Cont.*



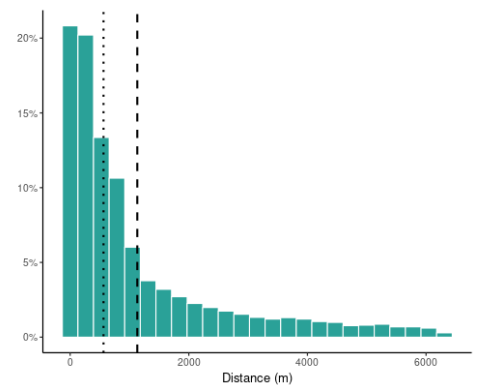
(i) Aspect type: predominantly plants on slopes facing south, east, and southeast.



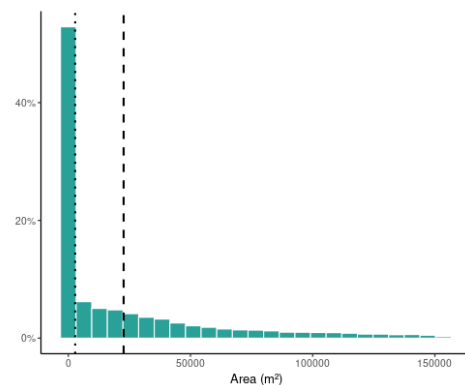
(j) Curvature: min: -0.17 ; max: 0.27 ; mean: -0.0006 ; median: 0.0001 ; std. dev.: 0.014 .



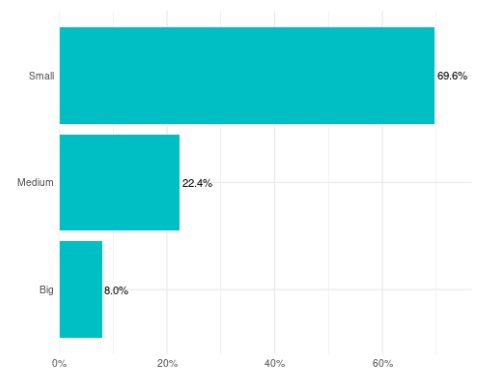
(k) Curvature type: a clear preference for terrain with little or no curvature.



(l) Distance to road: min: 0.00 ; max: 6309.67 ; mean: 1131.61 ; median: 562.41 ; std. dev.: 1412.82 .

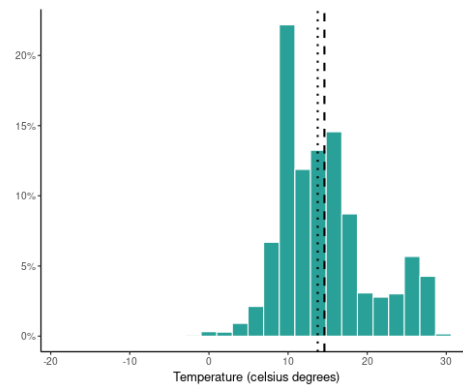


(m) Area: min: 0.013 ; max: 153488 ; mean: 22641.16 ; median: 2813.24 ; std. dev.: 34540.73 .

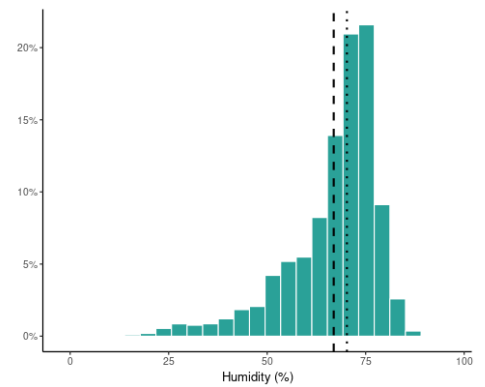


(n) Size: indicates a global trend toward small-scale solar power plants.

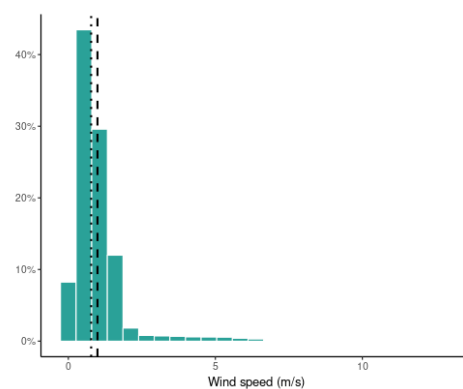
Figure 6. Cont.



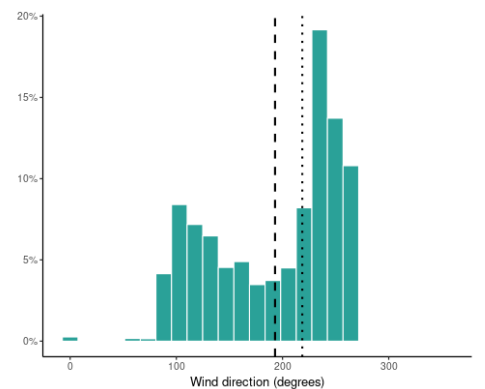
(o) Temperature: min: -17.06 ; max: 30.34 ; mean: 14.59 ; median: 13.74 ; std. dev.: 5.77 .



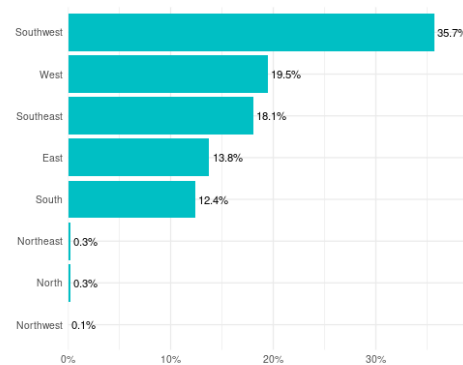
(p) Humidity: min: 0 ; max: 95 ; mean: 66.89 ; median: 70.25 ; std. dev.: 11.64 .



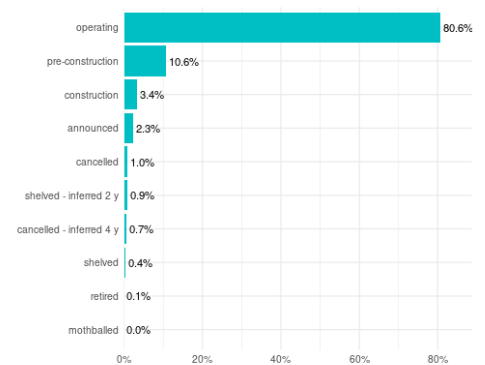
(q) Wind speed: min: 0.00 ; max: 12.75 ; mean: 0.99 ; median: 0.78 ; std. dev.: 0.93 .



(r) Wind direction: min: 0.00 ; max: 352.50 ; mean: 192.97 ; median: 218.60 ; std. dev.: 59.19 .



(s) Wind direction categories: predominantly southerly winds combined with easterly and westerly winds.



(t) Operational status: the vast majority of solar power plants are currently generating electricity.

Figure 6. Cont.

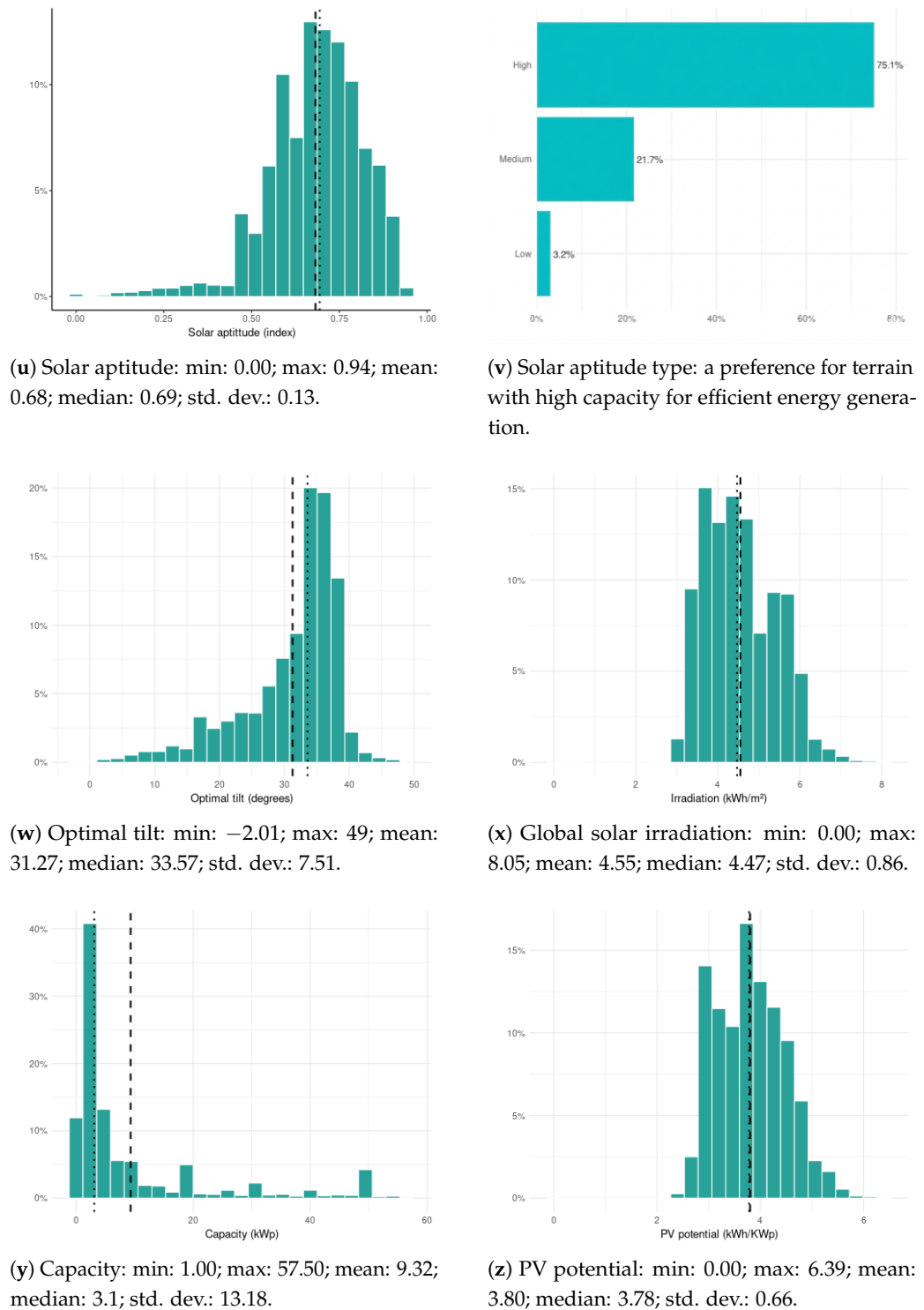


Figure 6. Distribution of the values of the variables in the dataset.

3.2. Topographical Variables

These variables enable an analysis of the soil or terrain conditions where the solar infrastructure is located. The elevation data show that more than 50% of solar power plants are located in lowland areas (Figure 6e), with an average elevation of 353.61 m above sea level, which is advantageous from a logistical and economic standpoint (proximity to cities and to electrical and road infrastructure). The vast majority of solar power plants (approximately 95%) are located on land with little or no slope, falling into the categories

of flat, nearly flat, and slightly sloped (Figure 6f,g). This makes sense both technically and economically, as it facilitates installation, simplifies operation and maintenance, reduces costs, and improves efficiency. The maximum value (34.68°) sets the upper limit of technological adaptability for plants on steep terrain. With regard to aspect or orientation (the direction the land faces), the average value is 171.58° (Figure 6h), very close to 180° . The majority of installations are located on slopes facing south, east, and southeast, accounting for more than 40% (Figure 6i). This distribution confirms that most plants are located in the Northern Hemisphere, with their leaves oriented toward the equator. More than 80% of the data show a curvature close to 0, indicating a clear preference for areas with low or no curvature (Figure 6j,k), where the terrain is geomorphologically stable and conducive to the efficient development of PV infrastructure.

3.3. Logistic Variables

These variables are what can determine the feasibility of building a solar energy project. A key logistic metric is the distance to the road (Figure 6l). Nearly 60% of solar power plants are located less than 1 km from an existing road, including those immediately adjacent to the road network (minimum of 0). This demonstrates that access to transportation infrastructure is a critical factor, so it is important to select sites that already have transportation links for personnel, panels, and heavy machinery. There are also plants located more than 6 km from the main road lines. These are typically megaprojects in desert or remote areas, where abundant solar resources justify the massive investment. One of the factors that influences the planning of resources, transportation, and implementation of a solar project is the size of the plant. In solar engineering, size determines operational complexity. As for the size of solar power plants (Figure 6m), there is considerable variation, with thousands of small PV plants coexisting alongside a few “giants” that generate the majority of the energy. More than 50% of the plants are concentrated in the range of 0 to 10,000 m², suggesting a significant presence of industrial and commercial self-consumption facilities or small distributed generation parks. The above dimensions were grouped into size categories. There are a large number of small plants, a moderate number of medium-sized plants, and a few large plants (Figure 6n). This is consistent with the global trend toward distributed solar power generation, that is, the sharp increase in residential and commercial installations in contrast to large-scale projects, which require greater investment and are more complex.

3.4. Climate Variables

Regarding the climatic conditions at the sites of solar power plants, one of the most critical climatic variables is temperature (Figure 6o). This variable represents a loss factor: as the cell temperature rises, the panel's efficiency decreases. The data distribution shows a clear predominance in temperate and warm climates, but the average of 14.59°C indicates that most plants operate under conditions where the air helps keep the panels at temperatures ideal for performance. Most plants operate in areas with moderate to high humidity levels (Figure 6p), with an average of 66.90%. Plant design must take into account corrosion processes in metal structures and material degradation caused by the constant presence of water vapor in the atmosphere. Wind acts as a thermal regulator and a structural factor. The wind speed (Figure 6q) indicates a predominantly low-turbulence operating environment, with an average of 0.99 m/s. Although this characteristic is somewhat detrimental to panel cooling and efficiency, it is highly beneficial for the long-term structural stability of PV arrays. The wind direction distribution is clearly bimodal (Figure 6r), with two main wind directions for most solar plants. A large concentration of plants is exposed to winds coming from between 220° and 260° (southwest), while a secondary area is exposed to winds coming from between 100° and 130° (southeast). A strategic correlation is observed

between wind direction (Figure 6s) and the orientation of solar plants. Winds blowing in a certain direction are most effective at cooling panels facing that same direction, since they strike the front surface of the panels directly.

3.5. Power Variables

These variables enable analysis of how solar photovoltaic plants are designed and how much energy they generate. The operational status is related to the life cycle of solar power plants (Figure 6t). The vast majority of plants are currently in operation; a smaller proportion consists of projects in the pre-construction and construction phases, and there is a low incidence of failed or withdrawn projects. This is a positive development, as it reflects maturity in electricity generation and the ongoing expansion of solar capacity. As for how “suitable” a site is for power generation, most values fall between 0.6 and 0.85 (Figure 6u), confirming that these sites are not only “suitable” but are very close to their theoretical maximum potential. This suggests that site selection criteria have prioritized areas with maximum theoretical efficiency. Approximately 75% of solar plants fall into the “High” solar suitability category (Figure 6v). This shows that the global solar industry does not install systems at random; sites are selected by prioritizing locations where solar energy is naturally optimal. Installation in an area with moderate or low suitability is typically driven by specific reasons, critical needs, or regions with high energy demand. In the context of a global dataset, these are outliers or atypical cases, whose presence adds diversity, allowing us to analyze different realities and identify patterns of efficiency outside of ideal conditions. The optimal tilt angle corresponds to 45% of the plants falling within the range of 30° to 40°, with an average of 31.27° (Figure 6w). This confirms that the design follows a rule of thumb for maximizing energy efficiency: the optimal tilt angle is typically approximately equal to the latitude of the location (mean of 34.86°). Irradiance is an indicator of available solar energy (Figure 6x). Their values are not scattered randomly; rather, they cluster around radiation values known to be optimal for the industry (similar to a Gaussian bell curve), but with a bias toward locations with high solar resources. The capacity diagram shows a high degree of heterogeneity (Figure 6y), demonstrating that it covers the entire spectrum of the photovoltaic industry. However, nearly 60% of the plants have an installed capacity of less than 10 MWp. This suggests that medium-sized solar farms and distributed generation projects have reached maturity. Finally, PV potential (kWh/kWp) is the most important performance metric; it tells us how much energy can actually be converted into electricity (Figure 6z). The vast majority of plants fall within the high-efficiency range (3.5 to 4.2 kWh/kWp), with an average of 3.80 kWh/kWp and a low standard deviation (0.65). This demonstrates that the global infrastructure is located in areas of high energy productivity.

4. Discussion

4.1. Plant Size Thresholds

The categorization of PV plants into small, medium, and large sizes in this study was adapted from standard industrial benchmarks through an empirical evaluation of the global dataset distribution. While traditional commercial frameworks often employ fixed thresholds based strictly on financial or regulatory segments, our approach reflects the actual operational geometries and capacities logged in global registries. This empirical classification preserves comparability across studies by providing a clear mathematical justification based on the distribution of our dataset and the visual validation of the plants using Google Earth, thereby minimizing subjectivity and adapting nominal industrial classifications to the actual geometries and operational scale of current PV plants.

4.2. Capacity Thresholds

A critical aspect of multi-source database integration is the standardization of plant categorization. To classify solar installations into small, medium, and large scales, this study adjusted standard industrial baselines. While standard industrial reports often apply rigid financial or grid-connection boundaries, our refined thresholds were derived directly from an empirical evaluation of the dataset's global power distribution (kW_p). This data-driven classification minimizes the subjectivity inherent in purely qualitative or single-region "visual surveys". By utilizing the operational distribution of over 58,000 installations, our thresholds establish statistically distinct categories that guaranty global comparability across highly heterogeneous energy landscapes—ranging from fragmented European rooftop clusters to vast, contiguous desert installations in Asia and North America. This categorization preserves both technical engineering relevance and computational consistency during large-scale downscaling and regional clustering analyses.

4.3. Geographic Distribution and Imbalances

A prominent characteristic of the compiled dataset is its significant geographic concentration in specific clusters, notably China, Japan, the United States, and Germany. This distribution accurately reflects the macroeconomic realities of global photovoltaic deployment, driven by aggressive national subsidies, regulatory incentives, and large-scale infrastructure investments. However, this concentration introduces an inherent geographic representation bias. Users of this dataset must consider that the derived statistical insights, spatial correlations, and multi-criteria relationships are heavily weighted toward the regional conditions of these leading countries, which might mask distinct operational or environmental patterns relevant to emerging markets in under-represented equatorial and southern hemisphere territories. Furthermore, this asymmetry is intrinsically related to variations in data density within our secondary geographic databases maintained by volunteers, such as OpenStreetMap (OSM). Regions with very active mapping communities exhibit considerably higher completeness rates for infrastructure polygons compared to developing regions, representing a macrospatial uncertainty that constitutes a recognized limitation of open-source mapping at the global scale. Consequently, although the dataset offers unprecedented unified global coverage, regional variations in density must be contextualized by taking into account local limitations in data contribution before conducting intercontinental density modeling.

4.4. Statistical Insights and Advanced Characterization

The multidimensional integration of topographic, logistical, climatic, and power attributes allows for the identification of latent patterns behind plant installations. For instance, the intersection between distance-to-road vectors and topographic constraints explains the trade-offs between the optimization of solar yield and logistics accessibility. Future research iterations can leverage this consolidated structure to implement advanced spatial statistics, such as regional clustering and multi-variable correlation matrices, confirming that global database integration serves as a predictable baseline for localized solar performance assessments. The non-linear interactions observed between the solar aptitude index and the actual spatial layout of the installations demonstrate that land-use optimization is heavily constrained by secondary morphological factors, such as hillshade and terrain curvature, rather than by incident solar radiation alone. By comparing these empirical results with the existing literature, this dataset represents an advance in the current state of the art, moving from isolated regional inventories to a comprehensive and dynamic tool capable of supporting simulations of the global energy transition and spatial optimization workflows.

5. Conclusions

This study presents a comprehensive and multidimensional global dataset of photovoltaic power plants, comprising 58,978 records enriched with 27 geographic, topographic, logistical, climatic, and power-related attributes integrated from multiple heterogeneous data sources. Supported by a largely documented and reproducible data integration methodology, the dataset constitutes a reliable and comprehensive resource that complements existing solar energy databases by providing broader attribute integration, improved data consistency, and extensive global coverage.

The selection of dataset attributes was guided by a comprehensive review of the scientific literature and existing photovoltaic data platforms, ensuring that all included variables were relevant and supported by current research. Compared with existing datasets, the proposed dataset incorporates additional high-value attributes, such as plant area and proximity to road infrastructure, which are essential for infrastructure characterization and subsequent analytical applications. Furthermore, the proposed workflow integrates data collection, processing, missing-data generation, cleaning, validation, and integration into a transparent, systematic, and largely reproducible methodology that can be readily adapted and extended to support future renewable energy data initiatives.

The integration of Geographic Information Systems (GISs) and Python-based automation enabled the efficient execution of complex geospatial and data engineering tasks required for dataset construction and the generation of derived attributes. By combining heterogeneous data sources within a unified and reproducible workflow, the proposed methodology reduces fragmentation and inconsistencies commonly found in existing photovoltaic datasets, resulting in a more comprehensive and reliable characterization of global photovoltaic infrastructure.

Descriptive statistical analysis of the 58,978 photovoltaic power plant records revealed consistent global patterns in the distribution and characteristics of existing installations. Operational and small-scale plants constitute the majority of the dataset, while notable geographic concentrations were observed in countries such as China and the United States. The analysis also identified common site-selection trends associated with terrain slope, surface curvature, accessibility to road infrastructure, prevailing wind direction, and panel orientation, reflecting the technical, environmental, and logistical factors that influence photovoltaic plant deployment. Collectively, these findings demonstrate the dataset's capacity to capture meaningful spatial and operational patterns, highlighting its value as a reliable resource for photovoltaic infrastructure analysis and data-driven decision-making.

Future work will focus on expanding the range of integrated attributes, implementing periodic updates to maintain data currency, and further validating derived variables to continuously improve the dataset's quality, completeness, and reliability. These enhancements will extend its applicability to emerging research challenges, including energy planning, site suitability assessment, predictive modeling, and intelligent decision-support systems [50–52]. By providing an open, reproducible, and extensible data resource, this work lays the foundation for the next generation of data-driven photovoltaic research and supports the global transition toward more sustainable energy systems.

Author Contributions: The conceptualization, study design, research, methodology, data collection and processing, validation, data analysis and interpretation, writing, and revision were carried out equally by all authors. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data and code that support the findings of this study are openly available in the official GitHub repository (<https://github.com/cimejia/solarPV>, accessed on 3 June 2026) and are permanently archived in Zenodo at <https://doi.org/10.5281/zenodo.21149809>, corresponding to version 1.0 (accessed on 2 July 2026).

Acknowledgments: The authors would like to thank the support team formed by Fernando Tarco, Karen Santamaria, Martin Sarmiento, Fernando Neira, Carlos Ramirez, Juleydi Miles, and Shelby Gualan. All of them are young students at the Central University of Ecuador, whose contributions have been essential to the completion of this project.

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

Appendix A. Supplementary Material

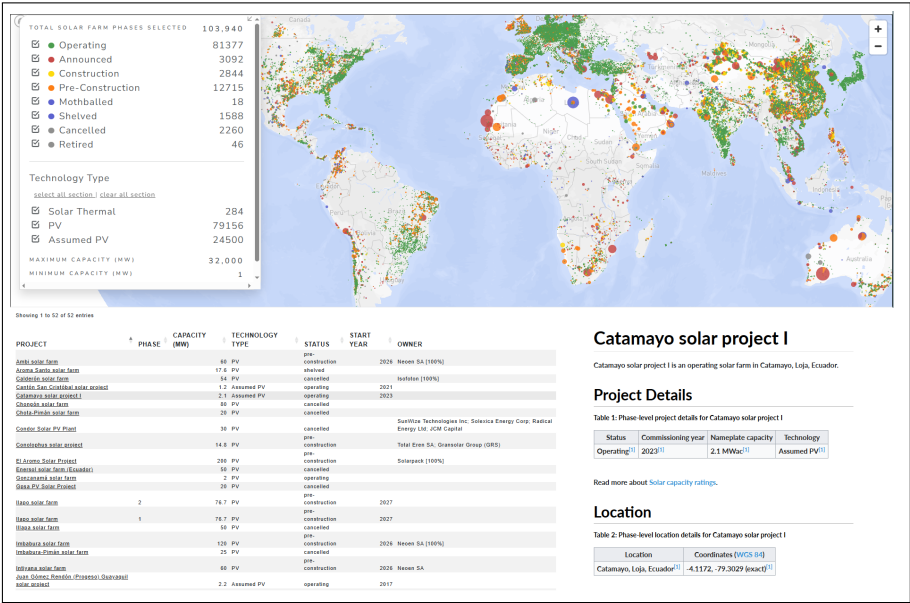


Figure A1. Global Energy Monitor (GEM) website.

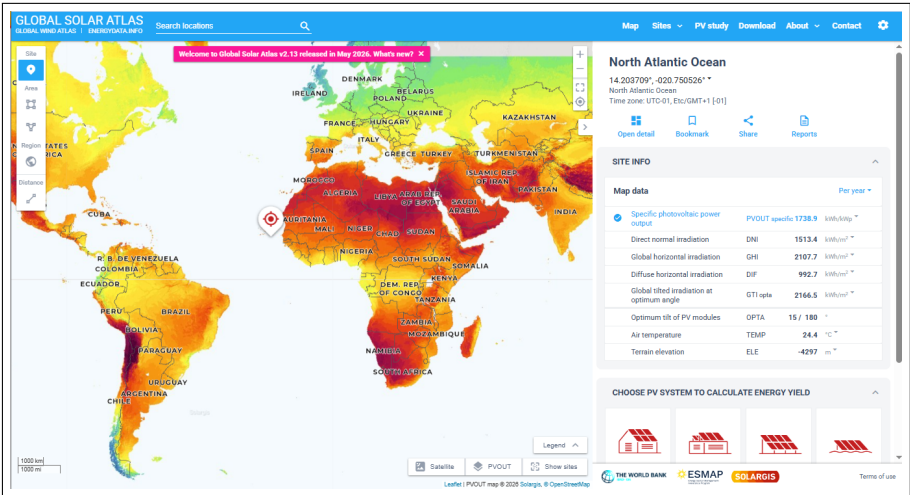


Figure A2. Global Solar Atlas (GSA) website.

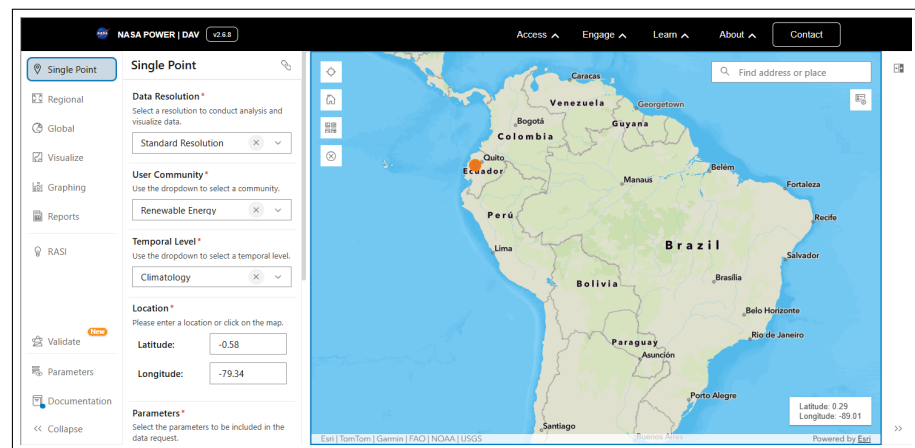


Figure A3. NASA POWER website.

References

1. Lijiang, L.W. A systematic review of photovoltaic power plant site selection approaches. *Energy Rep.* **2025**, *14*, 1993–2014.
2. Mekhilef, S.S. A review on solar energy use in industries. *Renew. Sustain. Energy Rev.* **2011**, *15*, 1777–1790.
3. Betak, J.; Caltik, M.; Cebecauer, T.; Chrkavy, D.; Erdelyi, B.; Rosina, K.; Suri, M.; Suriova, N. *Global Photovoltaic Power Potential by Country*; World Bank Group, ESMAP: Washington, DC, USA, 2020.
4. Fang, H.L. Sustainable site selection for photovoltaic power plant. *Energy Convers. Manag.* **2018**, *174*, 755–768.
5. Chen, X.; Li, B.; Braid, J.L.; Byford, B.; Colvin, D.J.; Glaws, A.; Jost, N.; Pierce, B.; Rabade, S.; Springer, M.; et al. Open data sets for assessing photovoltaic system reliability. *Appl. Energy* **2025**, *395*, 126132. <https://doi.org/10.1016/J.APENERGY.2025.126132>.
6. Ghiyas, S.; Sikarudi, D.; Jafari, M. A novel integrated GIS-AI framework for optimal CSP plant site selection. *Adv. Space Res.* **2025**, *76*, 5146–5167. <https://doi.org/10.1016/J.ASR.2025.08.028>.
7. Karadol, I.; Yildirim, R. Location choice in solar power plants by applying meteorological data to multi-criteria decision-making method. *Eng. Appl. Artif. Intell.* **2025**, *152*, 110766. <https://doi.org/10.1016/J.ENGAPPAL.2025.110766>.
8. Singh, D.; Elgeberi, N.; Aljaidi, M.; Kumar, R.; Al Mamlook, R.E.; Singla, M.K. Optimal Location of Renewable Energy Generators in Transmission and Distribution System of Deregulated Power Sector: A Review. *Energy Eng.* **2025**, *122*, 823–859. <https://doi.org/10.32604/EE.2025.059309>.
9. Rane, N.L. GIS-based multi-influencing factor application for optimal site selection. *Environ. Sci. Eur.* **2024**, *36*, 5.
10. Hadi, M.; Elbouchikhi, E.; Zhou, Z.; Saim, A.; Shafie-khah, M.; Siano, P.; Rahbarimaghani, H.; Colom, P.M. Artificial intelligence for microgrids design, control, and maintenance: A comprehensive review and prospects. *Energy Convers. Manag. X* **2025**, *27*, 101056. <https://doi.org/10.1016/J.ECMX.2025.101056>.
11. Kurukuru, V.H. A Review on Artificial Intelligence Applications for Grid-Connected Solar Photovoltaic Systems. *Energies* **2021**, *14*, 4690.
12. Cook, D.; Petursson, J.G. The role of GIS mapping in multi-criteria decision analysis in informing the location and design of renewable energy projects: A systematic review. *Energy Strategy Rev.* **2025**, *59*, 101765. <https://doi.org/10.1016/J.ESR.2025.101765>.
13. Goncalves, G.L.; Gaudencio, L.M.; de Oliveira, R.; do Nascimento, R.S. Sustainable planning of solar plants: Bibliometry and integration of GIS-multicriteria methods. *Sustain. Futur.* **2025**, *10*, 100993. <https://doi.org/10.1016/J.SFTR.2025.100993>.
14. Sharma, T.; Sarin, A. Multi-criteria decision making for solar site selection in Punjab, India. *Results Eng.* **2025**, *27*, 106288. <https://doi.org/10.1016/J.RINENG.2025.106288>.
15. Shao, Y.; Jiang, Y.; Kanell, T.; Xu, P.; Khattab, O.; Lam, M. *STORM: Synthesis of Topic Outlines Through Retrieval and Multi-Perspective Question Asking*; Stanford Open Virtual Assistant Lab, Stanford University: Stanford, CA, USA, 2024. Available online: <https://github.com/stanford-oval/storm> (accessed on 1 March 2026).
16. Consensus. Consensus: AI-Powered Academic Search Engine. 2024. Available online: <https://consensus.app> (accessed on 1 March 2026).
17. NASA POWER Project. *NASA POWER Data Access Viewer*; NASA Langley Research Center: Hampton, VA, USA, 2024. Available online: <https://power.larc.nasa.gov/> (accessed on 1 March 2026).
18. European Commission, Joint Research Centre. *PVGIS: Photovoltaic Geographical Information System*; European Commission: Brussels, Belgium, 2024. Available online: <https://re.jrc.ec.europa.eu/pvgis/> (accessed on 1 March 2026).
19. World Bank Group; Solargis. *Global Solar Atlas 2.0*. 2024. Available online: <https://globalsolaratlas.info> (accessed on 1 March 2026).

20. Solargis. SolarGIS: Solar Resource Data and Photovoltaic Performance Modeling. 2024. Available online: <https://solargis.com> (accessed on 1 March 2026).
21. Meteotest. Meteonorm: Global Meteorological Database for Solar Energy and Applied Climatology, Version 8.0. 2023. Available online: <https://meteonorm.com> (accessed on 1 March 2026).
22. Copernicus Climate Change Service. ERA5: Fifth Generation of ECMWF Atmospheric Reanalysis of the Global Climate. 2024. Available online: <https://cds.climate.copernicus.eu> (accessed on 1 March 2026).
23. Global Modeling and Assimilation Office (GMAO). MERRA-2: Modern-Era Retrospective Analysis for Research and Applications, Version 2. 2015. Available online: <https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/> (accessed on 1 March 2026).
24. NASA Langley Research Center. CERES: Clouds and the Earth's Radiant Energy System Data Products. 2024. Available online: <https://ceres.larc.nasa.gov> (accessed on 1 March 2026).
25. MINES Paris—PSL and Transvalor. HelioClim Solar Radiation Data. 2024. Available online: <https://www.soda-pro.com> (accessed on 1 March 2026).
26. World Climate Research Programme. Baseline Surface Radiation Network (BSRN). 2024. Available online: <https://bsrn.awi.de> (accessed on 1 March 2026).
27. NOAA Global Monitoring Laboratory. SURFRAD: Surface Radiation Budget Network. 2014. Available online: <https://gml.noaa.gov/grad/surfrad/> (accessed on 1 March 2026).
28. Global Energy Monitor. Global Energy Monitor. 2026. Available online: <https://globalenergymonitor.org/> (accessed on 16 March 2026).
29. Global Energy Monitor. Global Solar Power Tracker, February 2026 Release. Dataset. 2026. Available online: <https://globalenergymonitor.org/projects/global-solar-power-tracker/download-data/> (accessed on 16 March 2026).
30. FAO. *A Framework for Land Evaluation*; Food and Agriculture Organization of the United Nations: Rome, Italy, 1976.
31. Verheye, W.; Koohafkan, A.P.; Nachtergaele, F. The FAO guidelines for land evaluation. In *Encyclopedia of Land Use, Land Cover and Soil Sciences*; EOLSS Publishers: Oxford, UK, 2009; Volume 2, pp. 78–100.
32. Wilson, J.P.; Gallant, J.C. Digital Terrain Analysis. In *Terrain Analysis: Principles and Applications*; Wilson, J.P., Gallant, J.C., Eds.; John Wiley & Sons, Inc.: New York, NY, USA, 2000; pp. 1–27.
33. Geofabrik GmbH. OpenStreetMap Data Extracts. 2026. Available online: <https://download.geofabrik.de/> (accessed on 14 March 2026).
34. OpenStreetMap Contributors. OpenStreetMap. 2026. Available online: <https://www.openstreetmap.org> (accessed on 14 March 2026).
35. Xpert.Digital. Este es el Espacio que Necesita el sol: ¿Cuál es el Tamaño Mínimo de un Parque Solar para que Sea Rentable? 2024. Available online: <https://xpert.digital/es/tamano-minimo-del-parque-solar-rentable/> (accessed on 17 March 2026). (In Spanish)
36. Adjiski, V.; Serafimovski, D. GIS- and AHP-based Decision Systems for Evaluating Optimal Locations of Photovoltaic Power Plants: Case Study of Republic of North Macedonia. *Geomat. Environ. Eng.* **2024**, *18*, 51–82. <https://doi.org/10.7494/geom.2024.18.1.51>.
37. Turan, E.S.; Balci, B.; Erkurt, F.E. Optimization in Site Selection of Solar Power Plant by Using Geographical Information Systems and Analytical Hierarchy Process for the City of Adana/Türkiye. *SSRN* **2025**, preprint. <https://doi.org/10.2139/ssrn.5558040>.
38. Saaty, T.L. *The Analytic Hierarchy Process*; McGraw-Hill: New York, NY, USA, 1980.
39. Arslan, M.; Cunkas, M. An experimental study on determination of optimal tilt and orientation angles in photovoltaic systems. *J. Eng. Res.* **2025**, *13*, 2689–2701. <https://doi.org/10.1016/J.JER.2024.07.015>.
40. Global Energy Monitor. GEM.wiki. 2026. Available online: <https://www.gem.wiki/> (accessed on 16 March 2026).
41. Kruitwagen, L.; Story, K.T.; Friedrich, J.; Byers, L.; Skillman, S.; Hepburn, C. A global inventory of photovoltaic solar energy generating units. *Nature* **2021**, *598*, 604–610.
42. Li, A.; Liu, L.; Li, S.; Cui, X.; Chen, X.; Cao, X. Global photovoltaic solar panel dataset from 2019 to 2022. *Sci. Data* **2025**, *12*, 637.
43. Stowell, D.; Kelly, J.; Tanner, D.; Taylor, J.; Jones, E.; Geddes, J.; Chalstrey, E. A harmonised, high-coverage, open dataset of solar photovoltaic installations in the UK. *Sci. Data* **2020**, *7*, 394.
44. Zhang, X.; Xu, M.; Wang, S.; Huang, Y.; Xie, Z. Mapping photovoltaic power plants in China using Landsat, Random Forest, and Google Earth Engine. *Earth Syst. Sci. Data* **2022**, *14*, 3743–3755.
45. Ortiz, A.; Negandhi, D.; Mysorekar, S.R.; Nagaraju, S.K.; Kiesecker, J.; Robinson, C.; Bhatia, P.; Khurana, A.; Wang, J.; Oviedo, F.; et al. An Artificial Intelligence Dataset for Solar Energy Locations in India. *Sci. Data* **2022**, *9*, 497.
46. Stid, J.T.; Kendall, A.D.; Anctil, A.; Rapp, J.; Bingaman, J.C.; Hyndman, D.W. A harmonized dataset of ground-mounted solar energy in the US with enhanced metadata. *Sci. Data* **2025**, *12*, 1586.
47. Lin, Z.; Zhou, Q.; Wang, Z.; Wang, C.; Bookhart, D.B.; Leung-Shea, M. A high-resolution three-year dataset supporting rooftop photovoltaics (PV) generation analytics. *Sci. Data* **2025**, *12*, 63.
48. Wimalaratne, S.; Haputhanthri, D.; Kahawala, S.; Gamage, G.; Alahakoon, D.; Jennings, A. UNISOLAR: An Open Dataset of Photovoltaic Solar Energy Generation in a Large Multi-Campus University Setting. In Proceedings of the 2022 15th International Conference on Human System Interaction (HSI), Melbourne, Australia, 28–31 July 2022.

49. Sun, Y.; Li, Y.; Wang, R.; Ma, R. Dynamic geospatial modeling of solar energy expansion potential in China. *Energy* **2025**, *333*, 137433. <https://doi.org/10.1016/J.ENERGY.2025.137433>.
50. Shobanke, M.; Bhatt, M.; Shittu, E. Advancements and future outlook of Artificial Intelligence in energy and climate change modeling. *Adv. Appl. Energy* **2025**, *17*, 100211. <https://doi.org/10.1016/J.ADAPEN.2025.100211>.
51. Wang, Q.; Li, Y.; Li, R. Integrating artificial intelligence in energy transition: A comprehensive review. *Energy Strategy Rev.* **2025**, *57*, 101600. <https://doi.org/10.1016/J.ESR.2024.101600>.
52. Ukoba, K.; Olatunji, K.O.; Adeoye, E.; Jen, T.-C.; Madyira, D.M. Optimizing renewable energy systems through artificial intelligence. *Energy Environ.* **2024**, *35*, 3833–3879.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.