data

MDPI

Data Descriptor

Historical Hourly Information of Four European Wind Farms for Wind Energy Forecasting and Maintenance

Javier Sánchez-Soriano [1], Pedro Jose Paniagua-Falo [1] and Carlos Quiterio Gómez Muñoz [2,*]

[1] Escuela Politécnica Superior, Universidad Francisco de Vitoria, 28223 Pozuelo de Alarcón, Spain; javier.sanchez@ufv.es (J.S.-S.); pedrojosepaniagua01@gmail.com (P.J.P.-F.)
[2] HCTLab Research Group, Universidad Autónoma de Madrid, 28049 Madrid, Spain
* Correspondence: carlosq.gomez@uam.es

Abstract: For an electric company, having an accurate forecast of the expected electrical production and maintenance from its wind farms is crucial. This information is essential for operating in various existing markets, such as the Iberian Energy Market Operator—Spanish Hub (OMIE in its Spanish acronym), the Portuguese Hub (OMIP in its Spanish acronym), and the Iberian electricity market between the Kingdom of Spain and the Portuguese Republic (MIBEL in its Spanish acronym), among others. The accuracy of these forecasts is vital for estimating the costs and benefits of handling electricity. This article explains the process of creating the complete dataset, which includes the acquisition of the hourly information of four European wind farms as well as a description of the structure and content of the dataset, which amounts to 2 years of hourly information. The wind farms are in three countries: Auvergne-Rhône-Alpes (France), Aragon (Spain), and the Piemonte region (Italy). The dataset was built and validated following the CRISP-DM methodology, ensuring a structured and replicable approach to data processing and preparation. To confirm its reliability, the dataset was tested using a basic predictive model, demonstrating its suitability for wind energy forecasting and maintenance optimization. The dataset presented is available and accessible for improving the forecasting and management of wind farms, especially for the detection of faults and the elaboration of a preventive maintenance plan.

Dataset: DOI: 10.5281/zenodo.10846754. URL: https://doi.org/10.5281/zenodo.10846754 (accessed on 17 March 2025).

Keywords: wind power; renewable energy; forecasting; faults; preventive maintenance

Academic Editor: Giuseppe Ciaburro

Received: 11 February 2025
Revised: 13 March 2025
Accepted: 18 March 2025
Published: 19 March 2025

Citation: Sánchez-Soriano, J.; Paniagua-Falo, P.J.; Gómez Muñoz, C.Q. Historical Hourly Information of Four European Wind Farms for Wind Energy Forecasting and Maintenance. *Data* **2025**, *10*, 38. https://doi.org/10.3390/data10030038

Correction Statement: This article has been republished with a minor change. The change does not affect the scientific content of the article and further details are available within the backmatter of the website version of this article.

1. Introduction

In the rapidly evolving field of renewable energy, accurate prediction of wind energy production is pivotal for efficient market operation and strategic planning. The dataset underpinning this research encapsulates comprehensive meteorological data and wind park operational metrics, offering a rich foundation for predictive modeling using machine learning (ML) [1,2] and deep neural networks (DNNs).

Wind power has experienced significant growth in recent decades, becoming a crucial source of renewable energy. However, its intermittent nature poses unique challenges for grid integration and energy market management. Accurate forecasting of wind power production is essential for optimizing power system operation, reducing balancing costs, improving electricity market integration, and facilitating preventive maintenance planning [3–5].

Forecasting models vary according to the time horizon [6–9]. Recent studies have highlighted that XGBoost and Random Forest are effective in the medium and long term, while KNN with Min-Max performs better in the short term [10]. A method based on seasonality analysis improves long-term forecasting, with a MAPE error of 1–7%, compared to 6–10% for traditional methods. Valuation of models using metrics such as MAE [7] and RMSE indicates that XGBoost and Random Forest are more robust to changes in data scaling [10].

In all cases, access to reliable data is a challenge. Meteorological models such as ERA5 (1950–present, 30 km resolution) provide key information, albeit with biases in mountainous regions [6]. PLUSWIND, CHUWD-H [11], and others [12] provide hourly data for the USA, but fewer resources exist for Europe.

Maintenance costs are high, especially offshore [13]. Methods such as MILP optimize planning, but underestimate costs due to weather uncertainty [14]. Heuristics have been developed to improve estimation [14]. Data-driven diagnostics and hardware-in-the-loop have demonstrated reliability in early failure detection [15]. Blade erosion reduces efficiency, and protection systems (LEP) optimize blade life [16].

The primary objective of this research is to enhance the accuracy of wind energy predictions by leveraging advanced machine learning techniques, including deep learning and support vector machines (SVMs) [17,18]. Secondary objectives include the following:

- Developing robust predictive models capable of handling the inherent variability of wind energy production.
- Identifying key patterns and factors that influence wind energy generation.
- Providing tools for early fault detection and optimizing preventive maintenance.
- Contributing to the scientific understanding of renewable energy forecasting.

By offering a diverse, high-quality dataset spanning multiple geographical locations, this study enables comparative analyses and model generalization research. Additionally, it demonstrates the practical application of advanced AI techniques in renewable energy management, addressing real-world challenges. The research establishes a framework for collaboration between academia and industry, fostering joint innovation in renewable energy studies. Furthermore, it facilitates progress in fault detection and preventive maintenance for wind farms, which are crucial for operational efficiency.

The study tackles several key challenges in wind energy forecasting. One of the primary difficulties is the inherent variability and non-linearity of wind energy production, which complicates traditional prediction models. Another major challenge is the integration of diverse data sources, including meteorological and operational information, to enhance forecasting accuracy. Additionally, optimizing preventive maintenance strategies is essential for minimizing downtime and maximizing energy output. Finally, predictive models must be adaptable to different geographical contexts and operational conditions to ensure broad applicability.

Collected in collaboration with ENGIE-GEMS, this dataset provides real-world insights into wind energy dynamics, demonstrating the application of cutting-edge AI techniques to real-world challenges in the energy sector [19] and bridging the gap between theoretical models and practical energy management solutions.

The predictive model's development aligns with ongoing efforts to optimize energy production, emphasizing the strategic importance of accurate forecasting in energy markets.

This dataset is being actively used to develop new research lines, focusing on methodology, model development, and validation processes. It is being actively used to develop new research lines, focusing on methodology, model development, and validation processes. The findings contribute to the broader scientific dialogue on renewable energy management, offering insights into the potential of ML and AI [20] to revolutionize energy forecasting.

The public release of this dataset and its documented applications reflect a commitment to advancing scientific knowledge and fostering innovation in renewable energy. By sharing

this resource, the research community can explore new predictive models, validate existing theories, and contribute to the sustainable management of wind energy resources. This academia–industry collaboration sets a precedent for future studies in renewable energy forecasting and management.

2. Data Description

The dataset consists of two different tables, "Metadata" and "Data_wind". The first contains information about the codification and location of the different Eolic plants, which are those that allow the use of wind as an energy resource and are conditioned by the variability of this atmospheric phenomenon, although they benefit from the minimum technical requirements necessary for the operation of the installations. The second contains information about the specific energy produced, as well as climatological data.

In summary, the dataset collects information from four wind farms located in three different countries: two in Auvergne-Rhône-Alpes (France), one in Aragon (Spain), and one in the Piemonte region (Italy). In these parks, the number of wind turbines varies between 6 and 20, and their production capacities range between 17,400 and 38,000 kWh. More specifically, there are 58,496 hourly records distributed as follows among the farms: 14,879 for ID-1, 14,859 for ID-2, 14,537 for ID-3, and 14,221 for ID-4. The data have were collected between 15:00 on 14 April 2021 and 23:00 on 29 November 2023.

2.1. Metadata

The metadata dataset offers a detailed description of wind energy production sites in CSV format, encompassing a variety of geographical locations and capacities. Comprising data from four distinct sites, each entry in the dataset is enriched with comprehensive details that include the site's name, the number of wind generators, total capacity in kilowatt-hours (kWh), and geographical coordinates (latitude and longitude). These sites are strategically located across three countries, with unique environmental and infrastructural characteristics.

Each site is meticulously cataloged with its corresponding "Wind_generator_number", indicating the operational capacity through the number of generators installed at each site, and "Capacity kWh", showcasing the energy production potential. The geographical diversity is highlighted by the sites' spread across different regions and states, such as Haute-Savoie in France and Zaragoza in Spain, offering insights into the wind energy landscape across various European locales. The data are exemplified in Table 1. The variables in the dataset are as follows:

- Site: A textual identifier for each wind energy production site, serving as a unique name or designation that distinguishes each location.
- Wind_generator_number: The total number of wind generators (turbines) installed at each site. This variable indicates the scale of the wind energy operation at each location, providing insight into the site's capacity to generate electricity from wind.
- Capacity kWh: The total capacity of the wind energy site, measured in kilowatt-hours (kWh). This figure gives an indication of the maximum amount of electrical energy the site can produce under optimal conditions within a specific timeframe, highlighting the site's contribution to the energy grid.
- ID: An additional numerical identifier for each wind energy site, likely used for internal tracking or database management purposes. It serves as an alternative reference to the site variable.
- Latitude: The geographical latitude of the wind energy site, expressed in decimal degrees. Latitude is a critical factor in determining the amount of solar exposure and potentially influences wind patterns at the site.

- Longitude: The geographical longitude of the wind energy site, also in decimal degrees. Longitude, along with latitude, helps to precisely locate the site on the globe, facilitating spatial analysis and the assessment of geographical influences on wind energy production.
- City: The name of the nearest city to the wind energy site. This variable provides a local context, helping to associate each site with a nearby urban area for logistical, administrative, and social considerations.
- State: The state, province, or regional administrative division where the wind energy site is located. This variable further specifies the site's location within a country, offering insights into regional policies, wind energy incentives, and environmental conditions that might affect the site.
- Region: A broader geographical categorization that encompasses the site, often reflecting ecological, climatic, or administrative commonalities among sites within the same area.
- Country_iso: The ISO country code of the nation where the wind energy site is situated. ISO codes provide a standardized short-form representation of country names, facilitating data analysis and international comparisons.
- Country: The full name of the country hosting the wind energy site. This variable situates each site within a national context, highlighting the global distribution of wind energy operations and allowing for country-specific analyses of wind energy production.

Table 1. Metadata variables.

Site	Wind_Generator _Number	Capacity kWh	ID	Latitude	Longitude	City	State	Region	Country_iso	Country
Canacoloma	7	20,800	1	45.9993	6.6405	Magland	Haute-Savoie	Auvergne-Rhône-Alpes	FR	France
ElSasoG	6	17,400	2	46.0845	6.7098	Samoëns	Haute-Savoie	Auvergne-Rhône-Alpes	FR	France
LasMajas	20	38,000	3	41.97745	−0.94812	Castejón de Valdejasa	Zaragoza	Aragón	ES	Spain
SierradeLuna	8	21,400	4	45.6784	8.12859	Valdilana	Biella	Piemonte	IT	Italy

As demonstrated in Table 1, the following observations can be made with respect to the initial entry in the dataset. It corresponds to the Canacoloma wind energy site, which has seven wind turbines with an installed capacity of 20,800 kWh. This site is in the town of Magland, in the Haute-Savoie department, in the Auvergne-Rhône-Alpes region of France (FR). Its geographical coordinates are 45.9993 latitude and 6.6405 longitude. The information provided for this site makes it possible to analyze wind energy production according to its location and operational characteristics.

2.2. Data_Wind Dataset

The "Data_Wind.csv" dataset encapsulates a comprehensive collection of wind energy production data, spanning 58,496 instances from 14 April 2021, recorded over different timestamps from four unique wind energy production sites.

Figure 1 presents the distribution of several operational variables of a wind turbine, including the energy generated, ambient temperature (AmbientTemperature), nacelle angle (NacelleAngle_value), rotor speed (RotorSpeed_value), wind direction (WindDirection_value), and wind speed (WindSpeed_value). It is observed that energy exhibits a right-skewed distribution, with a high concentration of low values and a long tail extending toward higher values. The ambient temperature follows an approximately normal distribution, while both the nacelle angle and rotor speed display multimodal patterns, suggesting different operational states of the wind turbine. Conversely, wind direction shows a relatively uniform distribution, which is expected given the variability of wind flow, and wind speed also presents a right-skewed distribution, indicating that lower values are more

frequent. These patterns reflect the operational conditions of the wind turbine and provide insights into the relationship between environmental factors and system performance.

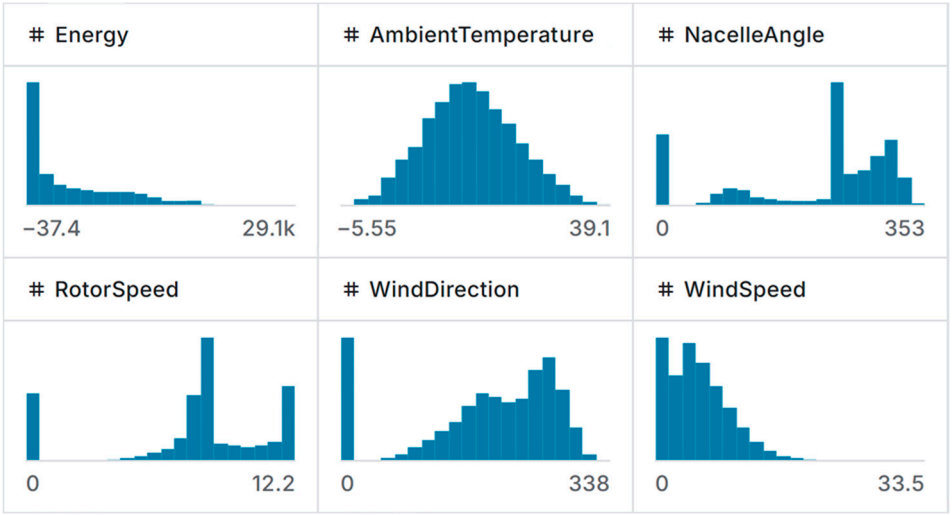

Figure 1. Histograms of the wind turbine's operational variables. The distributions are displayed, reflecting different operating states and environmental conditions.

This extensive dataset, structured in a time series format, offers detailed insights into the operational metrics and environmental conditions affecting wind energy generation. These values can be observed in Table 2. Key attributes of the dataset include the following:

- Timestamps (index): The dataset features 15,253 unique timestamps, capturing data at hourly intervals. Each timestamp follows the format YYYY-MM-DDTHH:MM:SS.000Z, providing a granular view of wind energy production dynamics over time.
- Site Identification (site): Data are aggregated from four distinct wind energy sites, each assigned a numerical identifier, allowing for site-specific analyses and comparisons.
- Energy Output: These values represent the amount of energy produced over each hourly period, ranging from −37.40 kWh to 29,148.28 kWh. Negative values may indicate periods of low wind activity or technical downtimes.
- Ambient Temperature (AmbientTemperature_value): Ambient temperatures vary significantly, from −5.54 °C to 39.06 °C, highlighting the diverse environmental conditions across the sites.
- Nacelle Angle (NacelleAngle_value): The angle of the nacelle is the structure at the top of the tower that houses all the internal components, such as the transmission system and the power generator. The value of this parameter is a crucial factor in optimizing wind energy capture. It ranges from 4.88° to 355°, reflecting the adjustments made to align with wind direction.
- Rotor Speed (RotorSpeed_value): Rotor speeds, an indicator of turbine activity, vary from 0 revolutions per minute or RPM (indicating standstill conditions) to 12.24 RPM, showcasing the operational speeds necessary for energy production, measured in rpm.
- Wind Direction (WindDirection_value): Wind directions recorded in the dataset range from 21.66° to 337.56°, providing insights into prevailing wind patterns at the sites, as measured in angles.
- Wind Speed (WindSpeed_value): Wind speeds are recorded for 14,439 instances, with measurements ranging from 0 m/s (calm conditions) to 27.17 m/s (indicating strong wind conditions), underscoring the variability of wind resources.

Table 2. Illustration of content in Data_Wind.

Index	Site	Energy_kWh	Ambient-Temperature_Value	Nacelle Angle_Value	RotorSpeed _Value	WindDirection _Value	WindSpeed _Value
14 April 2021 15:00:00,000	1	370.159	14.122	102	7.408	195.625	12.2
14 April 2021 15:00:00,000	2	207.683	14.921	86.291	7.484	162.667	10.06
14 April 2021 15:00:00,000	3	376.644	15.668	100.212	7.075	193.450	3.92
12 May 2021 10:00:00,000	4	9313.650	11.806	232.311	8.298	168.891	10.1

As demonstrated in Table 2, the following observations can be made with respect to the initial entry in the dataset. The initial entry in the dataset, dated 14 April 2021 at 15:00:00, documented operational data for site 1 (Canacoloma from Metadata). The energy production was recorded as 370.159 kWh at that juncture. The ambient temperature surrounding the wind turbine was measured at 14.122 °C. The nacelle, which houses the main components of the wind turbine, was oriented at an angle of 102°. The rotor was spinning at a speed of 7.408 revolutions per minute, while the wind direction recorded was 195.625°. The wind speed at that time was 12.2 m/s. These values reflect the environmental and operational conditions of the wind turbine at that time.

3. Methods

In data mining and analysis projects, there are several methodologies, such as KDD (Knowledge Discovery in Databases), SEMMA (Sample, Explore, Modify, Model, Assess), and CRISP-DM (Cross-Industry Standard Process for Data Mining) [21]. KDD is an approach that encompasses the entire knowledge discovery process, from the selection and preprocessing of data to the interpretation of results. It focuses on identifying useful patterns in large volumes of data and integrating statistical and machine learning techniques [22,23]. On the other hand, SEMMA is a method developed by SAS that focuses on data exploration and modeling using an iterative approach. Its structure facilitates experimentation with different models and preprocessing techniques. Finally, CRISP-DM (Cross-Industry Standard Process for Data Mining) is a methodology structured in six phases (business understanding, data understanding, data preparation, modeling, evaluation, and deployment) [24–26]. This cycle is illustrated in Figure 2.

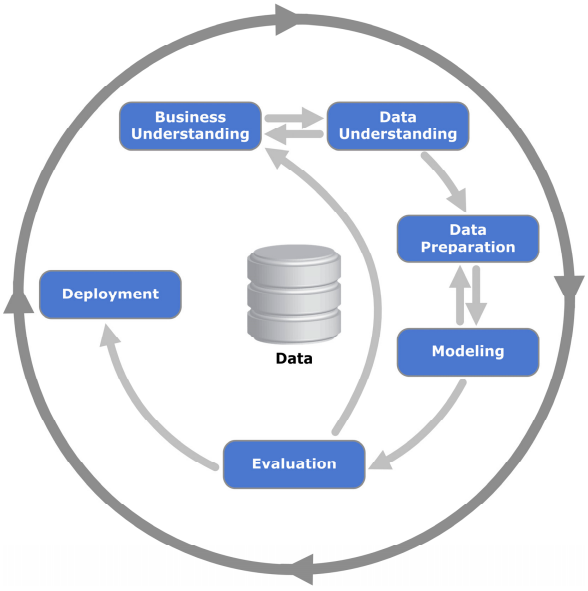

Figure 2. CRISP-DM methodology diagram [27].

The development of the dataset described in this paper, which contains information on electric wind turbines with data from the ENGIE-GEMS company, was achieved by utilizing the CRISP-DM methodology due to its structured and adaptable approach. This methodology facilitates a seamless integration of business objectives with data analysis, ensuring that the generated dataset aligns with specific business needs and is both interesting and useful for the scientific community. Furthermore, the emphasis on understanding and preparing the data is fundamental in a project where the quality and structuring of the information are key to guaranteeing its usefulness in future analyses and applications. The development of a model is beyond the scope of this data descriptor. This was done simply to validate the usability of the data as well as to explore possible new uses in subsequent work. Nevertheless, Section 4, "User Notes", shows some possible uses of the data in the basic models created to validate the CRISP-DM cycle.

3.1. Data Understanding

The first step in this phase is gaining a comprehensive understanding of the dataset by analyzing two critical tables: "Data_Wind" and "Metadata". The "Data_Wind" dataset consists of time series data capturing the operational parameters of wind turbines, such as power output, wind speed, wind direction, and environmental conditions at various timestamps. This dataset enables the identification of temporal patterns in wind energy production and the assessment of how environmental factors influence turbine performance, as shown in Table 1.

On the other hand, the "Metadata" dataset provides essential information about wind farms, including geographical locations, turbine specifications, capacity, and other site-specific characteristics. Together, these datasets offer a comprehensive overview of both the micro (turbine-level operations) and macro (wind farm characteristics) aspects of wind energy production; this can be seen in Table 2.

3.2. Data Preparation

During the data preparation phase of the CRISP-DM methodology for this work, a meticulous process was undertaken to refine and structure the dataset for the development of a possible predictive model for wind energy forecasting and maintenance. It should be noted that the preprocessing and preparation of the data to make it reusable and interesting for the scientific community was tailored to the information systems available in the company, so that the replicability of the process would be possible in similar circumstances and configurations. Briefly, the data preprocessing included several key phases:

- First, data ingestion was performed by retrieving the datasets from the company's storage systems, ensuring that all available records were loaded into the pipeline. The data sources consisted of SCADA system logs and maintenance records, which were imported into CSV and SQL formats.
- Next, the two datasets were consolidated through an inner join operation in Dataiku, ensuring temporal alignment of the measurements. The join was performed on the "Timestamp" field, which was first standardized to a common format to avoid mismatches due to time zone differences or different timestamp representations.
- A data integrity check was then conducted to identify missing values and inconsistencies. Missing values in critical fields (such as wind speed or power output) were handled using forward-fill or interpolation techniques, while categorical missing values (such as turbine status) were replaced using the mode of the corresponding category.
- Unnecessary data were filtered out, and the units of measurement were standardized by converting MWh to kWh to ensure consistency. It was decided to convert the measurements from MWh to kWh by multiplying by 1000 for consistency, given that there were fewer measurements in MWh.

- To further enhance consistency, all numerical values were converted to float64 format, avoiding potential issues with integer division in subsequent processing steps.
- A pivot operation was applied to structure the data by turbine and date, facilitating analysis, and irrelevant columns were removed.
- Additionally, an outlier detection step was implemented using the interquartile range (IQR) method, where data points falling outside 1.5 times the IQR were flagged for review. This step was crucial in eliminating erroneous readings that could distort the predictive model.
- Categorical variables, such as turbine operational modes, were one-hot encoded to allow for their integration into machine learning models.
- The final dataset structure was optimized for understanding and manipulating the electric production metrics in relation to the specific characteristics of each wind turbine. Unnecessary columns were discarded to improve data quality
- Finally, the cleaned and structured dataset was exported in both CSV and Parquet formats, ensuring compatibility with a wide range of analytical tools.

It is important to note that this process was applied to information from the four sites available at ENGIE-GEMS, and it would be reproducible if new sites were available in the system.

3.3. Modeling

A wide range of models can be implemented using this dataset, each with the potential to unlock significant insights and value in the realm of renewable energy, particularly in wind power generation. To perform a simple validation of the data and identify possible errors or shortcomings, a wind speed and direction prediction model was trained. The results and metrics are not the focus of this paper, but the usefulness of such a model as part of the CRISP-DM methodology cycle was demonstrated. Some potential examples of models that could be trained using this dataset are shown in Section 4, "User Notes".

3.4. Evaluation

The last step involved dividing the dataset into training and testing parts, using November data to validate the model, and ensuring that the dataset contained data from specific dates. This comprehensive data preparation process not only involved meticulous cleaning and structuring but also highlighted the necessity of applying a tree reduction method. By conducting 100 reductions with a depth of 10 leaves to minimize noise, this method proved advantageous over principal component analysis (PCA) or Pearson correlation index, particularly in handling the complexity and variability inherent in meteorological and operational data from wind parks. Decision trees effectively decompose the dataset into smaller subsets based on conditional decisions, enhancing the model without sacrificing precision. This approach simplifies the model by identifying significant patterns and eliminating outliers or noisy data that could distort predictions, illustrating a tailored and thoughtful approach to data preparation within the CRISP-DM framework.

A wide range of evaluation metrics can assess the performance of a model. Each metric captures specific aspects of performance, thus contributing to a comprehensive and detailed understanding of the accuracy and efficacy of the predictive model. In accordance with the following dataset, some possible metrics could be R^2—coefficient of determination [28], MSE—mean squared error, RMSE—root mean squared error [29], MAE—mean absolute error [30], MAPE—mean absolute percentage error [31], and EVS—explained variance score, among others. Some of these metrics were considered to validate the model and, in turn, the data constructed.

A simple model was trained with a range of hyperparameter settings, varying the number of hidden layers (HL), the number of units per layer (Units), the learning rate (LR), the batch size, the dropout rate, and the regularization coefficients L2 and L1. The models were evaluated using metrics such as evs, mae, mape, rmse, and r^2, as shown in Table 3. The findings indicate that the optimal configuration, yielding the lowest values for MAE and RMSE, was achieved with a model comprising five hidden layers and 180 units per layer, in conjunction with a learning rate of 0.0001. The configuration was found to be optimal, with a dropout rate of 0.2 and regularization coefficients L2 = 0.4 and L1 = 0.2, resulting in an EVS of 0.969 and an R^2 of 0.956.

Table 3. Results obtained from training a model with different hyperparameter settings.

HL	Units	LR	Batch Size	Dropout	L2	L1	evs	mae	mape	rmse	r^2
3	140	0.001	64	0.4	0.8	0.2	0.968	134.116	0.558	277.864	0.957
5	180	0.0001	64	0.2	0.4	0.2	0.969	133.649	0.602	276.398	0.956
7	100	0.001	64	0.4	0.4	0.2	0.956	148.897	0.592	288.345	0.954
9	100	0.0001	64	0.4	0.4	0.2	0.956	140.325	0.592	282.813	0.955

Figure 3 presents a line graph that compares the actual and predicted values of power generation in kilowatt-hours (kWh) over the course of the day on 11 November 2023. The horizontal axis of the graph represents the hours of the day, ranging from 0 to 23, while the vertical axis displays the power values. The graph illustrates the red line as a representation of the actual values, while the black line signifies the model predictions. It is evident that both lines demonstrate variations throughout the day, exhibiting peaks and troughs at varying times. There are instances when the two curves coincide, and there are also instances when there are discrepancies between the actual and predicted values.

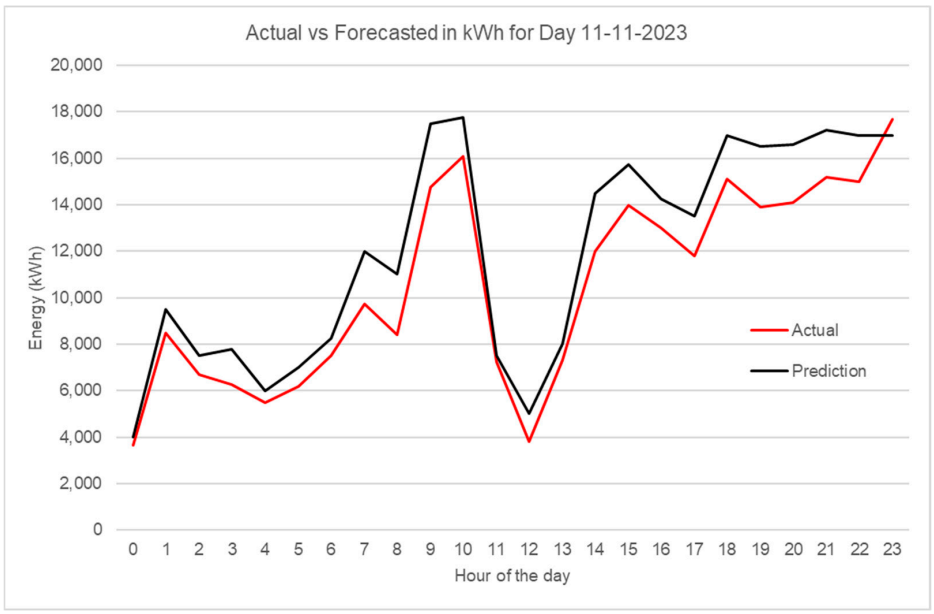

Figure 3. Actual and predicted values of energy generation in kWh throughout the day.

4. User Notes

All data from both tables (Metadata and Data_wind) are stored as two CSV files (comma-separated values) suitable for importing to virtually any spreadsheet program. All files can be accessed without logging in at https://zenodo.org/doi/10.5281/zenodo.10846754 (accessed on 17 March 2025). The semicolon (;) has been used as a data separator, leaving the period (.) to determine the decimal part of the values contained in the dataset.

Some possible uses of this dataset are proposed below, delving into areas such as advanced forecasting, optimization, and real-time adaptive control systems:

- Optimal Turbine Control Strategy. Developing a model to determine the optimal settings for a wind turbine's nacelle angle and rotor speed involves using advanced machine learning techniques like reinforcement learning or supervised learning algorithms (e.g., decision trees or gradient boosting machines). Trained on historical data, this model identifies the correlation between wind conditions and optimal turbine settings to maximize power output, ensuring that the turbine operates at peak efficiency, as we can see in Figure 4.
- Wind Speed and Direction Forecasting. Forecasting wind speed and direction at turbine sites is critical for short-term power generation, grid stability, and financial planning. Using time series forecasting models, such as Autoregressive Integrated Moving Average (ARIMA) or Long Short-Term Memory (LSTM), the system leverages historical wind data to predict future conditions accurately. This model helps optimize energy production and maintain a stable power supply. To aid in understanding, an interactive wind map is proposed, showcasing forecasted wind speeds and directions at various turbine locations. This tool supports informed decision-making by highlighting potential wind condition shifts and merging predictive insights with spatial and temporal data for proactive wind farm management.
- Energy Yield Optimization. This model analyzes data to find patterns and correlations between environmental factors (wind speed, direction, temperature) and energy yield. Using regression models or deep learning networks, it aims to identify the most efficient operating conditions for turbines. Advanced visualization techniques like scatter plots and heatmaps illustrate the optimal conditions for maximum energy production, providing actionable insights to enhance turbine efficiency. Figure 5 showcases the intricate relationship between power production, wind direction, and the angle of the turbine's nacelle.
- Geospatial Analysis for Optimal Wind Farm Location. Geospatial analysis, potentially enhanced by machine learning, is used to evaluate historical wind data, terrain characteristics, and other environmental factors to find optimal wind farm locations. This analysis ensures strategic turbine placement for maximum efficiency and energy production. Maps serve as dynamic visualization tools, marked with color coding or symbols to denote site suitability. These maps, featuring wind patterns, terrain, and infrastructure overlays, offer a comprehensive view of each site's advantages and limitations, facilitating informed decision-making for new wind farm installations that are efficient and sustainable.

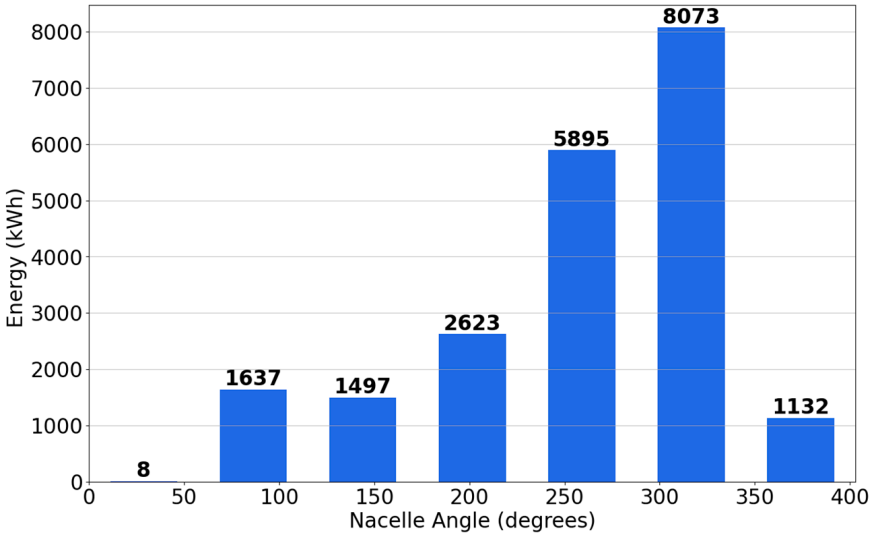

Figure 4. Distribution of energy values by nacelle angle.

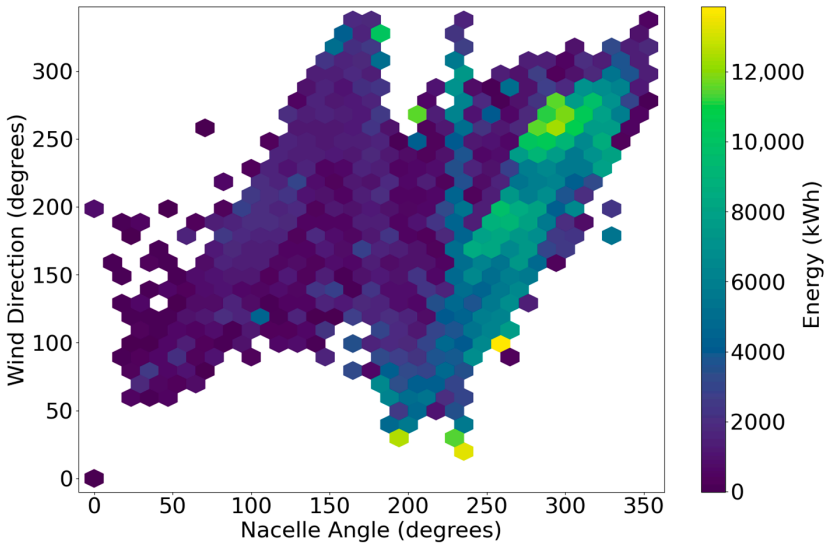

Figure 5. Hexagon plot showing casing production values in relation to the direction and nacelle angle.

5. Conclusions

The importance of this dataset extends beyond its immediate practical applications, serving as a critical tool for researchers and professionals in the energy sector. It offers a unique opportunity to delve into the nuances of wind energy dynamics, providing real hourly data from four European wind farms over a two-year period. This level of granularity and geographical coverage is scarce in the current literature on research in Europe. This wealth of information allows for the exploration and validation of advanced predictive models for production optimization and preventive maintenance planning. Collaboration with ENGIE-GEMS not only underscores the dataset's relevance but also highlights the synergy between academic research and industry needs, fostering a better understanding of the complexities of wind energy production.

Furthermore, the dataset's accessibility to the broader scientific community plays a vital role in advancing renewable energy research. It enables the exploration of new predictive models, with excellent metrics, as can be seen in Table 3; the validation of existing theories; and the development of innovative strategies for the sustainable management of wind energy resources. To ensure its usability, the dataset was validated using a basic predictive model, demonstrating its reliability and potential for wind energy forecasting. The rigorous application of the CRISP-DM methodology in the creation and validation of this dataset ensures its structure, quality, and replicability, providing a reference framework for future studies in the field. The results confirm that the dataset is well structured and suitable for training machine learning models, reinforcing its value for further research.

Ultimately, this data descriptor contributes significantly to the field by providing a valuable and accessible resource that drives innovation in wind farm prediction and management, enabling comparative analysis across geographically diverse locations and the generalization of advanced artificial intelligence models. This open access to valuable data underscores the commitment to advancing scientific knowledge and innovation in renewable energy, setting a precedent for future studies in wind energy forecasting and management.

Author Contributions: Conceptualization, J.S.-S.; methodology, P.J.P.-F.; software, P.J.P.-F.; validation, P.J.P.-F. and J.S.-S.; formal analysis, J.S.-S. and C.Q.G.M.; investigation, J.S.-S. and P.J.P.-F.; resources, C.Q.G.M. and J.S.-S.; data curation, P.J.P.-F.; writing—original draft preparation, P.J.P.-F.; writing—review, C.Q.G.M. and J.S.-S.; writing—editing, J.S.-S. and C.Q.G.M.; visualization, P.J.P.-F. and J.S.-S.; project administration, C.Q.G.M.; funding acquisition, C.Q.G.M. All authors have read and agreed to the published version of the manuscript.

Funding: The project, titled "ICARUS—Inspección y Control Automatizado con Redes neuronales y UAVs en Sistemas Fotovoltaicos" (ref. SI4/PJI/2024-00233), is funded by the Comunidad de Madrid through a direct grant agreement for the promotion of research and technology transfer at the Universidad Autónoma de Madrid.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data can be accessed without logging in at https://zenodo.org/doi/10.5281/zenodo.10846754 (accessed on 17 March 2025).

Acknowledgments: The authors would like to thank the Universidad Francisco de Vitoria and the Universidad Autonoma de Madrid for their support. In addition, special thanks are due to ENGIE-GEMS for allowing us to use their historical data for the dataset construction and authorize its use for research purposes.

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

References

1. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* **2021**, *8*, 53. [CrossRef] [PubMed]
2. Muñoz, C.Q.G.; Márquez, F.P.G. Wind energy power prospective. In *Renewable Energies*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 83–95.
3. Habib, A.; Hossain, M.J. Revolutionizing Wind Power Prediction—The Future of Energy Forecasting with Advanced Deep Learning and Strategic Feature Engineering. *Energies* **2024**, *17*, 1215. [CrossRef]
4. Karaman, Ö.A. Prediction of Wind Power with Machine Learning Models. *Appl. Sci.* **2023**, *13*, 11455. [CrossRef]
5. Rosende, S.B.; Sánchez-Soriano, J.; Muñoz, C.Q.G.; Andrés, J.F. Remote Management Architecture of UAV Fleets for Maintenance, Surveillance, and Security Tasks in Solar Power Plants. *Energies* **2020**, *13*, 5712. [CrossRef]
6. Bloomfield, H.C.; Brayshaw, D.J.; Deakin, M.; Greenwood, D. Hourly historical and near-future weather and climate variables for energy system modelling. *Earth Syst. Sci. Data* **2022**, *14*, 2749–2766. [CrossRef]
7. Borunda, M.; Ramírez, A.; Garduno, R.; García-Beltrán, C.; Mijarez, R. Enhancing Long-Term Wind Power Forecasting by Using an Intelligent Statistical Treatment for Wind Resource Data. *Energies* **2023**, *16*, 7915. [CrossRef]
8. Muñoz, C.Q.G.; Márquez, F.P.G. Future maintenance management in renewable energies. In *Renewable Energies*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 149–159.
9. Gómez, C.Q.; García, F.P.; Villegas, M.A.; Pedregal, D.J. Big Data and Web Intelligence for Condition Monitoring: A Case Study on Wind Turbines. In *Handbook of Research on Trends and Future Directions in Big Data and Web Intelligence*; IGI Global Publishers: Hershey, PA, USA, 2015. [CrossRef]
10. Ekinci, G.; Ozturk, H.K. Forecasting Wind Farm Production in the Short, Medium, and Long Terms Using Various Machine Learning Algorithms. *Energies* **2025**, *18*, 1125. [CrossRef]
11. Wang, C.; Deng, C.; Horsey, H.; Reyna, J.L.; Liu, D.; Feron, S.; Cordero, R.R.; Song, J.; Jackson, R.B. CHUWD-H v1.0: A comprehensive historical hourly weather database for U.S. urban energy system modeling. *Sci. Data* **2024**, *11*, 1383. [CrossRef]
12. Millstein, D.; Jeong, S.; Ancell, A.; Wiser, R. A database of hourly wind speed and modeled generation for US wind plants based on three meteorological models. *Sci. Data* **2023**, *10*, 883. [CrossRef]
13. Stock-Williams, C.; Swamy, S.K. Automated daily maintenance planning for offshore wind farms. *Renew. Energy* **2019**, *133*, 1393–1403. [CrossRef]
14. Carlos, S.; Sánchez, A.; Martorell, S.; Marton, I. Onshore wind farms maintenance optimization using a stochastic model. *Math. Comput. Model.* **2013**, *57*, 1884–1890. [CrossRef]
15. Simani, S.; Farsoni, S. Fault diagnosis and sustainable control of wind turbines: Robust data-driven and model-based strategies. In *Fault Diagnosis and Sustainable Control of Wind Turbines: Robust Data-Driven and Model-Based Strategies*; Elsevier: Amsterdam, The Netherlands, 2018.
16. Hoksbergen, N.; Akkerman, R.; Baran, I. The Springer Model for Lifetime Prediction of Wind Turbine Blade Leading Edge Protection Systems: A Review and Sensitivity Study. *Materials* **2022**, *15*, 1170. [CrossRef] [PubMed]
17. Lu, P.; Ye, L.; Zhao, Y.; Dai, B.; Pei, M.; Tang, Y. Review of meta-heuristic algorithms for wind power prediction: Methodologies, applications and challenges. *Appl. Energy* **2021**, *301*, 117446. [CrossRef]

18. Jiménez, A.A.; Muñoz, C.Q.G.; Márquez, F.P.G. Machine Learning and Neural Network for Maintenance Management. In *Lecture Notes on Multidisciplinary Industrial Engineering*; Springer: Berlin/Heidelberg, Germany, 2018. [CrossRef]
19. Marugán, A.P.; Márquez, F.P.G.; Perez, J.M.P.; Ruiz-Hernández, D. A survey of artificial neural network in wind energy systems. *Appl. Energy* **2018**, *228*, 1822–1836. [CrossRef]
20. Wang, S.; Qin, C.; Feng, Q.; Javadpour, F.; Rui, Z. A framework for predicting the production performance of unconventional resources using deep learning. *Appl. Energy* **2021**, *295*, 117016. [CrossRef]
21. Azevedo, A.; Santos, M.F. KDD, SEMMA and CRISP-DM: A Parallel Overview. 2008. Available online: http://hdl.handle.net/10400.22/136 (accessed on 6 March 2025).
22. Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. The KDD process for extracting useful knowledge from volumes of data. *Commun. ACM* **1996**, *39*, 27–34. [CrossRef]
23. Shaaban, A.G.; Khafagy, M.H.; Elmasry, M.A.; El-Beih, H.; Ibrahim, M.H. Knowledge discovery in manufacturing datasets using data mining techniques to improve business performance. *J. Electr. Eng. Comput. Sci.* **2022**, *26*, 1736–1746. Available online: https://www.researchgate.net/profile/Amani-Shaaban/publication/361086019_Knowledge_discovery_in_manufacturing_datasets_using_data_mining_techniques_to_improve_business_performance/links/63330630165ca2278778589a/Knowledge-discovery-in-manufacturing-datasets-using-data-mining-techniques-to-improve-business-performance.pdf (accessed on 6 March 2025). [CrossRef]
24. Solano, J.A.; Cuesta, D.J.L.; Ibáñez, S.F.U.; Coronado-Hernández, J.R. Predictive models assessment based on CRISP-DM methodology for students performance in Colombia—Saber 11 Test. *Procedia Comput. Sci.* **2022**, *198*, 512–517. [CrossRef]
25. Martinez-Plumed, F.; Contreras-Ochando, L.; Ferri, C.; Hernandez-Orallo, J.; Kull, M.; Lachiche, N.; Ramirez-Quintana, M.J.; Flach, P. CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Trans. Knowl. Data Eng.* **2019**, *33*, 3048–3061. [CrossRef]
26. IBM SPSS Modeler CRISP-DM Guide. Available online: https://www.ibm.com/docs/it/SS3RA7_18.3.0/pdf/ModelerCRISPDM.pdf (accessed on 17 March 2025).
27. Jensen, K. Crisp-dm Ilustrration. Available online: https://es.m.wikipedia.org/wiki/Archivo:CRISP-DM_Process_Diagram.png (accessed on 17 March 2025).
28. Saunders, L.J.; Russell, R.A.; Crabb, D.P. The Coefficient of Determination: What Determines a Useful R^2 Statistic? *Investig. Opthalmology Vis. Sci.* **2012**, *53*, 6830–6832. [CrossRef]
29. Wang, Z.; Bovik, A.C. Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures. *IEEE Signal Process. Mag.* **2009**, *26*, 98–117. [CrossRef]
30. Pelanek, R. Metrics for Evaluation of Student Models. *J. Educ. Data Min.* **2015**, *7*, 1–19.
31. Ballı, S. Data analysis of Covid-19 pandemic and short-term cumulative case forecasting using machine learning time series methods. *Chaos Solitons Fractals* **2020**, *142*, 110512. [CrossRef] [PubMed]