energies

MDPI

Article

Wind Farm and Resource Datasets: A Comprehensive Survey and Overview

Diogo Menezes [1], Mateus Mendes [1,2,*], Jorge Alexandre Almeida [1,3] and Torres Farinha [1,4]

[1] Polythecnic Institute of Coimbra—ISEC, 3030-199 Coimbra, Portugal; a21270934@isec.pt (D.M.); jorge.alexandre.almeida@ubi.pt (J.A.A.); tfarinha@isec.pt (T.F.)
[2] Institute of Systems and Robotics, University of Coimbra—ISR, DEEC, 3030-290 Coimbra, Portugal
[3] Electromechatronic Systems Research Centre, University Beira Interior, 6201-001 Covilhã, Portugal
[4] Centre for Mechanical Engineering, Materials and Processes, 3030-788 Coimbra, Portugal
* Correspondence: mmendes@isr.uc.pt

Received: 16 July 2020; Accepted: 1 September 2020; Published: 9 September 2020

Abstract: The use of clean and renewable energy sources is increasingly important, for economic and environmental reasons. Wind plays a key role among renewable energy sources. Hence, the location, monitoring and maintenance of wind turbines are areas that have received more and more attention in recent years. The paper presents a survey of datasets of wind resources, wind farm installed capacity and wind farm operation, which contain generous amounts of data. Those datasets are important tools, freely available for analysis of wind resources and study of the performance of wind turbines. A short analysis of one of the datasets is also presented, identifying different operational regions, and the ones more likely to aggregate failures. Principal Component Analysis (PCA) is used to study wind turbines' behavior.

Keywords: wind farm; wind resources; wind turbine maintenance

1. Introduction

Energy supply chains are fundamental for modern human societies, which require large amounts of energy per capita. In the last centuries, fossil fuels were the main energy source for electricity production, transportation and other economic sectors. The trend, however, is not sustainable because of the environmental footprint and rising exploration costs for oil. To overcome or minimize the limitations of non-renewable energy sources, different renewable sources were proposed, such as biomass, hydropower and wave, geothermal, solar and wind energy. Although energy supply chains have been changing for a renewable reality, there are still many barriers to renewable energy development, such as the conversion cost and efficiency, location selection and distribution network, among others. Wee et al. [1] report the performance of these new renewable energy supply chains, the existing barriers and how to surpass them.

Wind energy took a key position in the new trend and wind turbines have been widely deployed to generate electric power. Numerous, large and small, wind farms were created in recent decades, in different countries, as part of the global effort to expand production of clean energies from renewable energy sources. Many of the wind farms are offshore, in order to collect higher wind power values and have a lower environmental impact on land usage.

Planning a wind farm is a difficult task. Among other challenges, it is necessary to choose an adequate location. A good location will have good wind power density, stable wind speed, a small environmental footprint and easy access for maintenance, among other requirements.

For optimal performance of a wind farm, it is also important to monitor several variables such as wind speed and temperature of key parts of the wind turbines. Good monitoring and adequate maintenance of wind turbines allow to optimize production and prevent malfunctions that could lead to downtime or even endanger people and property. It can also considerably reduce the maintenance costs of the turbines and support infrastructure. Hau & Erich [2], and Letcher [3], offer a comprehensive overview and insight into aspects of wind energy, from historical background to the fundamental science behind the modern industry, covering technical and economic aspects.

Predictive maintenance has been considered to be the best answer for maintenance of wind farms. It allows the extension of components' lifetime, the maximization of energy output and the reduction of maintenance costs, leading to the performance of corrective maintenance just before equipment failure [4]. For determining the expected point of failure, it is important to monitor wind turbines and analyze the data collected, using data analysis techniques.

The present paper proposes a survey of the state-of-the-art datasets of wind resources and wind farms. Most of the datasets are available for public use, and they offer a wealth of information which can and must be analyzed for optimal decisions in the process of planning wind farms and optimizing maintenance plans.

The remainder of the paper is organized as follows. Section 2 presents a literature review. Section 3 describes some fundamentals about wind as an energy resource and wind turbines' technology. Section 4 discusses the characteristics of good datasets. Sections 5–7 report important open datasets related to wind turbine capacity and wind farm projects, wind measurements and wind turbine/farm monitoring Supervisory Control and Data Acquisition (SCADA) systems, respectively. Section 8 gathers other available datasets and a discussion about the existence of data, its quality and comparison of all datasets. Section 9 presents a deeper overview of a specific dataset. Section 10 reports the main contributions of the present work to the state of art. Section 11 draws some conclusions and proposes future work.

2. Literature Review

Numerous research projects analyze existing wind resources and wind turbine monitoring datasets, advancing the state of the art soon after the datasets are available for public use.

González-Aparicio et al. [5] propose a methodology to capture local geographical information and generate meteorological derived wind power time series, allowing better understanding of the wind resource at wind farms. The study followed up to develop a European wind power generation dataset called European Meteorological derived HIgh RESolution (EMHIRES) [6]. Both studies mention several sources of wind farm and wind resource databases, such as the Wind Power, Global Wind Atlas and Merra dataset, operational forecast wind speed datasets, European Centre for Medium-Range Weather Forecasts (ECMWF) dataset and wind statistics reports from different countries around the world. Diffendorfer et al. analyze onshore wind turbine locations for the United States [7], with the purpose of creating a free, centralized, national, turbine-level geospatial dataset, for scientific research, land and resource management. In 2017, the USWTDB (United States Wind Turbine Database) [8] was created, a national turbine capacity database. Van Vuuren & Vermeulen also report about investigation of wind speed profiles for renewable energy development zones in South Africa [9].

Predicting the output of a wind farm is an important goal for wind energy industry, and one of the most, if not the most, important variable to look at when deciding to move forward with a wind farm project. Therefore, it is very important to develop performance models. However, it may represent a challenge, since wind turbines' power is essentially determined by variables which are hard to predict with good accuracy, due to their stochastic nature. Kusiak et al. [10] examine time series models to predict wind speed and power at different time scales, namely ten minutes and one hour long. They use the wind speed as an input to compute an integrated k-nearest neighbors model, for prediction of wind farm output. The author uses five different algorithms to construct the time series models, in order to select the most suitable for the task. The algorithms used include Support Vector Machine regression algorithm [11,12], Multilayer Perceptron [13], Reduced Error Pruning tree [14], M5P Tree [13,15,16] and the Bagging Tree [13,17,18]. Research used data generated at a wind farm and collected by a SCADA system, resulting in 4455 recorded instances for wind speed and power at 10 min intervals. The dataset was divided into a 3568 observations training dataset and an 887 observations test dataset.

Computation Fluid Dynamics (CFD) models have also been applied recently [19]. Most of the models are related to prediction of wind turbines wakes and dynamic behavior, associated with overall power generation. Wu et al. [20] use Large-Eddy Simulation (LES) to explore the effect of turbine array configurations on the turbine wake characteristics, as well as the power extraction efficiency. The paper associates the impact of the turbines' hub arrangements and the wind farm power generation. Lin & Porté-Agel [21] also use LES model to study wind turbine wakes, comparing the prediction results between the two different yaw models. Li & Yang [22] use the Actuator Disk (AD) model to simulate wind turbine wakes. They also present a study on AD and Actuator Surface (AS) models and their capability to predict dynamic behavior, on utility-scale turbines, for both uniform and turbulent non-uniform conditions. Uchida [23] also studies the wake characteristics of wind turbines, by predicting them with LES models and parallel computation based on a hybrid LES/actuator line (AL) model. The accuracy from both models is compared and the effects of inflow shear on the wake characteristics is investigated.

When planning a wind farm project, another very important decision is the choice of the turbine with the most suitable characteristics for the place and operating conditions. Pessanha et al. [24] propose one methodology to analyze anemometer data and evaluate wind potential, in order to help to identify which turbine characteristics should be chosen for maximum profit. The authors use Weibull distribution as a wind speed frequency distribution model at 25 m and 50 m height. Knowing wind speed values at two different heights facilitates estimating wind speed values for other heights, using Equation (1).

$$v = v_{measured} \times (\frac{h_{measured}}{h})^{\alpha} \tag{1}$$

In Equation (1), v is the wind velocity desired at height h, $v_{measured}$ is the wind velocity measured at height $h_{measured}$, and α is the power exponent. After wind power velocity, the authors calculate the average power, from the turbine's power curve and the Weibull distribution. From average power values it is possible to calculate the Capacity Factor (CF), given by Equation (2).

$$CF = \frac{E_{actual}}{E_{ideal}} = \frac{Time \times P_{average}}{Time \times P_N} \tag{2}$$

Using data from different wind turbine models with different characteristics, the energy production models are computed for each one. The data used are from Sistema de Organização Nacional de Dados Ambientais (SONDA), a Brazilian project to implement infrastructures to survey wind and solar energy resources, with 10 min sampling period, as described in Section 6.3.

The benefits of monitoring wind turbines are quantified in [4], where the authors show some of the costs for different maintenance plans. Almost all data collections referenced in the present work have 10 min sampling period. That can be seen as a negative aspect, due to the possible loss of information. Higher frequency data sampling offers more accurate information [25], although at the cost of using additional computing power. Among other techniques, Principal Component Analysis (PCA) is a useful statistical technique that is applied for data reduction with minimal loss of information [26,27]. It is often used in complement with machine learning algorithms [28,29].

3. Wind Power and Wind Turbines Fundamentals

3.1. Wind Power

Wind is atmospheric air in motion. Depending on the speed of the moving air it is possible to determine the strength of the wind and estimate the amount of energy on it. The fundamental equation of wind power is given by Equation (3), in which ρ is the air density, v is the wind velocity and A is the area covered by rotor's blades.

$$P = \frac{1}{2} \times \rho \times A \times v^3 \tag{3}$$

Wind Power Density (WPD) is also used to compare wind resources, independently of the turbine's rotor size. It is the quantitative basis for the standard classification for wind resources. WPD is given by Equation (4).

$$WPD = \frac{P}{A} \tag{4}$$

The wind power class can be classified in several different levels of potential resource, according to wind power density. Wind's energy massively depends on its speed and mass of air. However, not all the power in the wind is available for use and the Power Coefficient (C_p) quantifies the ratio of power extracted by the turbine and the total wind power, according to Equation (5).

$$C_p = \frac{P_T}{P_{Wind}} \tag{5}$$

The Power Coefficient value is a percentage of the power that can be extracted. According to the Betz Limit, there is a theoretical upper limit for a wind farm or wind turbine. According to this theory, the maximum power coefficient is 59% [3]. Wind Power efficiency can be specified by another variable, called capacity factor (CF). CF is represented by the ratio of actual generated energy to the energy that could potentially be generated by the system in ideal environmental conditions. CF may also be regarded as the fraction of the year the turbine generator is operating at its nominal capacity. This nominal capacity is not overflown to avoid mechanical damage and parts wearing.

Usually, a realistic wind farm project has a 30% CF, but with good wind resources it could reach values up to 50%. Weather conditions might be the primary driver for the CP, but in a long-term period, it is always a design/economic decision. Over the 20–30 years lifetime of a wind farm, weather conditions will average out leaving the wind farm developer trade-offs between the cost of the blades, mechanics and electronics that compose the nacelle [30].

3.2. Wind Turbine Technology

Wind turbines are mechanical systems that capture wind energy and transform it in electricity, using complex technology for maximum conversion efficiency. They involve different technical areas including aerodynamics, mechanics, structure dynamics, meteorology and electrical engineering. Wind turbine technology has evolved very fast since the 80s. The main differences from the modern technology to the past technologies are in electrical design and control. Presently, wind turbines have variable speed and active control. They can be installed onshore or offshore, particularly in constant wind zones. Their operation can be summarized in three important steps:

1. Wind force against the blades causes them to rotate and propel the rotor. Connected to the main shaft, the rotor is responsible for moving the generator;
2. Inside the turbine there is a speed multiplier, with capacity to spin at 1500 RPM (Rotations Per Minute), allowing the generator to transform mechanical energy into electrical energy;
3. The electricity is conducted through the interior of the tower to the outside power lines.

A wind turbine will start working when the wind reaches the cut-in speed. There is no justified energy conversion below the cut-in speed. The turbine's power is also limited to its rated power and whenever it reaches the cut-out wind speed, it stops working with the purpose of not doing extra mechanical efforts and preserving mechanical quality. The wind turbine power curves are calculated using the cut-in and cut-out wind speeds and the theoretical power outcome for different wind speeds.

They show how the turbines will perform in function of the wind. Wind turbine technology mainly consists of the following components [3]:

- Rotor—The rotor is the first element in the chain of functional elements of a wind turbine. It captures the power from the blades and converts it to kinetic mechanical power. Typically, it has two or three blades. The most popular design in wind technology today is the horizontal axis rotor;
- Transmission System—Comprises the rotor shaft, mechanical brake(s) and a gearbox. The mechanical brakes are used as a backup system for the aerodynamic braking system. The gearbox acts as a rotational speed auger, converting the slow high torque rotation of the rotor into a faster rotation;
- Generator—Electromechanical component that converts the mechanical power into electrical power. There are two main types of generator used in the industry, which are synchronous and asynchronous:
  - The synchronous generator operates at the synchronous speed, dictated by the connected grid frequency, regardless of the applied torque's magnitude. It is more expensive and mechanically more complicated than an asynchronous generator of a similar size. It has one significant advantage compared to the alternative, specifically, it does not need power compensation equipment.
  - The asynchronous/induction generator has several robustness advantages, mechanical simplicity and it is produced in large series for a low price. However, the major disadvantage is that the stator needs a reactive magnetizing current. The asynchronous generator consumes reactive power to get its excitation, which may be supplied by the grid or by power electronics. The interaction of the associated magnetic field of the rotor with the stator field results in a torque acting on the rotor;
- Power Electronic Interface—The electrical power produced by the generator is fed into the power grid through the power electronic interface. It is placed between the generator and the power grid, satisfying both component requirements. The interface assures that the turbine's speed rotation is adjusted to extract maximum power from the wind and route it on to the grid, controlling active and reactive power, frequency and voltage;
- Control System—Assures a proper operation of the wind turbine under all operational conditions. It keeps the wind turbine within its normal operating range by passive or active means, maximizing the power production and lifespan and reducing structural loads on mechanical components and thus their costs.

4. Data Classification and Characteristics of a Good Dataset

The following sections list and describe important wind energy related datasets, which aggregate substantial and important amounts of data. These data are fundamental for modern machine learning applications and Big Data algorithms. There are different definitions of Big Data in the literature. Although some authors focus on the ontological characteristics of the data, others focus on the computational difficulties of processing data. According to Kitchin & McArdle [31], the concept of Big Data is still being defined, but Big Data datasets must abide by a majority of general traits, such as: (i) Volume (space required to storage data); (ii) velocity (considered a key attribute, it represents the frequency of generation, handling, recording or publishing); (iii) Variety (weakest characteristic attribute); (iv) Exhaustivity (seeking entire population within a system); (v) Resolution; (vi) Relationality; (vii) Extensionality (flexibility of data generation, where a highly flexible data system has a strong extensionality).

The quality of a dataset is directly related to the organization that creates it, and data quality is often related to its value and accuracy. However, data quality has other dimensions, such as uniqueness, completeness, validity and consistency. Also, a good quality dataset must not have errors due to incomplete data, as well as syntactic or semantic errors.

In summary, good open datasets offer a wealth of information that is easily available and should comply with the characteristics referenced above and the following requirements:

1. Available on the web, in an open format and under a license that permits use for research and other uses;
2. Available as machine-readable structured data, such as Comma Separated Value (CSV) files or other common data format;
3. Available in non-proprietary formats, such as CSV or eXtensive Markup Language (XML);
4. Complies with common open standards and main international standards for the World Wide Web;
5. Contains enough information about where the data were collected, or link the data to a context;
6. Contains data points in sufficient quantity and quality for use in data mining, machine learning or other computational methods. The more sensors are monitored the better;
7. Sampling frequency must be high enough to capture and describe the most important variables;
8. Ideally there are no gaps in the data, or the gaps are short enough not disrupting the patterns.

The most common definition of data quality is that which determines that the data can fulfill the function for which it was collected.

5. Open Datasets of Wind Turbine Capacity and Wind Farm Projects

5.1. The Wind Power Database

The Wind Power database, with free access at www.thewindpower.net (accessed on 17 August 2020), is a comprehensive database of detailed raw statistics on the rapidly growing sphere of wind energy and its supporting markets. Data are regularly updated. The database contains data from a variety of players in the worldwide wind industry, such as wind farm developers, operators and owners and turbine manufacturers. Also, it provides direct and immediate access to information about regions, countries, types and number of turbines with their relative hub height, nominal power and the capacity factor in which the operator lean on. It is not fully available for public use. A license must be purchased to navigate and have access to all the data. The free access model represents a particular complete data base in terms of labeling nominal power capacity around the world and the existing wind farms, with mention to their manufacturers and owners. The data collected in this data base come from different sources, mainly external:

- Developers, operators, and investors;
- Insurers and legal experts;
- Parts manufacturers, service providers, subcontractors;
- Heads of strategy, development, and R & D departments;
- Analysts, cartographers and meteorologists;
- Public organizations, universities, and research institutes;
- Professional associations.

However, some authors propose that this worldwide database may have a significant number of gaps, inconsistencies and inaccuracies [5]. Table 1 shows the wind power capacity installed in each country, according to this database. As the table shows, the country with larger capacity installed is China, with 133,799 GW installed. The country with more farms is Germany, with a total of 5253 wind farms.

Table 1. List of wind power capacity installed in each country, according to The Wind Power database

Country	Continent	Wind Farms	Wind Power (GW)	Country	Continent	Wind Farms	Wind Power (GW)
China	Asia	1848	133,799	Thailand	Asia	25	1053
USA	North America	1351	110,072	Egypt	Africa	10	1048
Germany	Europe	5253	62,777	Vietnam	Asia	26	1039
United Kingdom	Europe	1008	30,085	Croatia	Europe	25	912
India	Asia	624	29,983	Russia	Asia	18	892
Spain	Europe	1001	24,026	New-Zealand	Oceania	21	811
France	Europe	1251	17,157	Bulgaria	Europe	47	645
Brazil	South America	527	16,649	Pakistan	Asia	9	637
Canada	North America	280	13,827	Serbia	Europe	9	604
Italy	Europe	401	10,871	Lithuania	Europe	65	536
Australia	Oceania	102	10,212	Jordan	Asia	7	470
Sweden	Europe	943	9048	Philippines	Oceania	11	457
Turkey	Asia	195	8186	Costa Rica	North America	18	414
Denmark	Europe	1368	6717	Estonia	Europe	28	412
Mexico	North America	65	6533	Hungary	Europe	35	385
Netherlands	Europe	546	5991	Peru	South America	6	373
Poland	Europe	277	5916	Dominican Republic	North America	7	366
Portugal	Europe	255	5469	Kenya	Africa	2	336
Belgium	Europe	176	4196	Panama	North America	3	336
Ireland	Europe	236	3905	Czech Republic	Europe	65	326
South Africa	Africa	37	3428	Ethiopia	Africa	3	325
Argentina	South America	64	3211	Tanzania	Africa	1	300
Romania	Europe	68	2982	Iran	Asia	13	284
Norway	Europe	46	2956	Kazakhstan	Asia	7	252
Austria	Europe	256	2879	Tunisia	Africa	3	243
Greece	Europe	164	2858	Cyprus	Europe	6	189
Japan	Asia	245	2813	Nicaragua	North America	5	187
Chile	South America	41	2805	Honduras	North America	3	180
Finland	Europe	181	2382	Senegal	Africa	1	159
Uruguay	South America	46	1572	Mongolia	Asia	7	156
Ukraine	Europe	41	1502	Luxembourg	Europe	19	151
Morocco	Africa	15	1283	Albania	Europe	1	150
South Korea	Asia	59	1159	Indonesia	Oceania	2	147
Taiwan	Asia	29	1142	Mauritania	Africa	3	137

5.2. United States Wind Turbine Database

China is currently the country that produces more electricity from wind power with a total of approximately 199.50 GW per year. The United States of America is currently the second world producer of wind energy, producing approximately 133.28 GW of wind power per year, according to The Wind Power database. The USA is also one of the regions in the world where it is easier to find information about wind energy, wind farm capacity and wind resources along the 50 states. One of the most important databases available is the United States Wind Turbine Database (USWTDB) [8]. The USWTDB currently contains information of nearly 60,000 turbines that go from 30 m high and 70 kW capacity to turbines towering 181 m high with 6 MW capacity. It covers onshore and offshore installations. The US Department of Energy, in partnership with Lawrence Berkeley National Laboratory (LBNL), United States Geological Survey and the America Wind Energy Association (AWEA) developed the USWTDB in 2017, creating a comprehensive, accurate and regularly updated wind turbine dataset. It includes not just the location of the turbines, but also the characteristics of each turbine, such as the model, total and hub height, rotor diameter, year of installation and rated capacity. All the technical specifications are listed in the Federal Aviation Administration and Digital Obstacle File and collected via AWEA, LBNL and turbine manufacturers website. As new data become available, the USWTDB is updated and can be accessed by researchers and public via its online portal and in a variety of downloading file formats:

- Geographic Information System (GIS)—The shape file format is a popular geospatial vector data format, compatible with a variety of GIS software;
- Tabular Data—CSV format of all the information that is provided in the USWTDB;
- Metadata, XML format—Background information which describes the content, quality, condition and other appropriate data characteristics.

5.3. United Kingdom Wind Energy Database

The United Kingdom Wind Energy Database (UKWED) [32] contains data about operational onshore and offshore wind projects, which can be searched or browsed, with projects shown as a list or on a dynamic map. Although UKWED only holds information on 100 kW and larger projects, statistics on micro, small and medium wind turbines are published annually in an annual market report. It is free to use for any purpose and the data creator is RenewableUK, a business group focused on building a future energy system, powered by clean electricity, by ensuring increasing amounts of renewable electricity that are developed across the United Kingdom and access markets to export all over the world. Information on Project Status, Project Intelligence Hub and Wind Energy Maps also exists, but it is available only for RenewableUK workers and members. The database consists of a big list of projects, giving information about each wind farm technical specifications. The datasets are not downloadable. However, there are summary reports about wind operations in RenewableUK site's publications section at https://www.renewableuk.com/search/all.asp?bst= (accessed on 17 August 2020).

6. Wind Resource

The wind speed and direction are the most important variables that affect a wind turbine's output. Hence, accurate predictions and measurements of the wind behavior are fundamental for wind farm planning and management. Among other important decisions, the turbine model must be adequate for the wind available in the place, for maximum efficiency and life cycle.

6.1. OpenEI Dataset

OpenEI dataset [33] is a trusted source of energy data, specifically for renewable energy and energy efficiency. The information provided is aimed at helping to make informed decisions on energy, market investment and technology development. The data can be viewed, edited and added by the users after the evaluation of content by experts. Open data is part of the core mission for OpenEI and for that purpose, most accessed data on OpenEI comes from several different resources, such as Department of Energy Open Data Catalog (DOE Data), International Utility Rate Database (IURDB) and United States Utility Rate Database (URDB). OpenEI offers information about wind resources, rather than wind farms characteristics. The contents include wind maps, meteorological data and wind power maps, among other variables. In total there are 216 files available in the wind sector. Some of the data used in the present research were extracted from OpenEI.

6.2. Native American Anemometer Loan Program

The Native American Anemometer Loan Program (ALP) was conducted by the U.S. Department of Energy (DOE) and an initiative from Wind Powering America (WPA). The purpose of the ALP was to provide native American tribes a low-cost, low-risk means of quantifying their wind resource, since there were no data to make wind project production and economic performance estimations with precision. The validation process is based on a quality control strategy adopted by the Baseline Surface Radiation Network. By providing native American tribes a low-cost way of quantifying the wind resource in their lands, it was expected that they would be encouraged to pursue wind development, leading to the installation of wind turbines. The program was launched in 2000 and by the end of 2011, 90 towers had been installed over 10 states.

The ALP's anemometer towers record information about the wind, such as its speed, direction and turbulence, with 10 min sampling periods. All the information was forwarded to the National Renewable Energy Laboratory (NREL) of the USA and analyzed. Free access is given to 11 of the 144 locations that conducted this activity.

Table 2 shows some of the regions with free access to the raw data, available at OpenEI From the regions mentioned in the table, the Navajo Indian Reservation has some missing data, due to malfunction of one of the sensors. Almost all the anemometer towers are 20 m high, because the 20 m high towers were considered adequate for wind turbine projects up to ∼100 kW. All the data of the dataset are provided by NREL. The users are granted the right, without any fee or other cost, to copy, modify, enhance and distribute the data, as long as the users agrees to credit the NREL. The dataset contains monthly average values from more than 700,000 wind observations. The location with less observations available is the Navajo India Reservation, with just 35,661 samples collected in 2004 and 2005. There is a gap in the data, due to a malfunction in the direction vane, which lasted from 18 November 2004 to 22 February 2005. The place with more observations recorded is Northern Cheyenne India Reservation, with a total of 90,891 samples recorded in 2003 and 2004. The Pine Ridge Indian Reservation does not mention how much observations it took to develop the monthly wind speed, direction and turbulence average values.

Table 2. Regions included in the Native American Anemometer Loan monitoring program, with observation towers at 20 m and 120 m.

Region	State	Observations	Monitoring Period	Height (m)
Keweenaw Bay India Reservation	Michigan	54,025	7 June 2007 to 16 June 2008	20
Navajo India Reservation	Arizona	35,661	13 April 2003 to 25 March 2005	120
Bethel	Alaska	60,036	23 February 2003 to 15 April 2004	20
Wind River Indian Reservation	Wyoming	56,147	7 March 2002 to 28 July 2003	20
Ugashik Traditional Village	Alaska	60,455	6 June 2001 to 31 July 2002	20
Tanana Village	Alaska	55,167	20 September 2001 to 13 October 2002	20
Table Bluff India Reservation	California	70,983	23 September 2002 to 29 January 2004	20
Pine Ridge Indian Reservation	S. Dakota	-	29 October 2001 to 22 October 2002	20
North. Cheyenne India Reservation	Montana	90,891	19 February 2003 to 11 November 2004	20
Fort Belknap India Reservation	Montana	83,440	7 February 2001 to 6 November 2002	20
Potawatomi Indian Reservation	Oklahoma	63,648	8 November 2004 to 21 January 2006	20

6.3. SONDA

The SONDA network was born from a Brazilian project to install resources that could track data about wind and solar energy resources in Brazil. Every group of data available passed through a validation process to ensure their reliability, since there are numerous factors that can affect the reliability of the data. Project SONDA [34] provides wind speed, direction and air temperature at 25 m and 50 m height, with a 10 min sampling frequency. The network has three different stations. Table 3 summarizes the monitoring period of each survey station of the SONDA dataset. The dataset contains almost six years of good quality data about wind resource on the Brazilian west coast, up to a total of 385,488 observations, Table 4.

Table 3. Data period for each survey station of the SONDA wind monitoring project, in Brazil.

Station	Location	2004	2005	2006	2007	2008	2009
BJD	08° 22′ 02″ S 36° 25′ 46″ O	Jul.–Dec.	Jan.–Aug.				
SCR	07° 22′ 54″ S 36° 31′ 38″ O			Jan.–Dec.	Jan.–Apr. Jun-Dec.	Jan.–Dec.	Jan.–Set.
TRI	07° 49′ 38″ S 38° 07′ 20″ O	Jul.–Dec.	Jan.–Aug.	Jan.–Dec.	Jan.–Apr.		

Table 4. Number of observations from each station.

Station	Observations
Belo Jardim (BJD)	62,488
São João do Cariri (SCR)	192,672
Triunfo (TRI)	131,328
Total	385,488

6.4. Ethiopia Wind Measurement Data

Ethiopia Wind Measurement Data is a data repository for measurements collected from 17 wind poles, placed in different regions of Ethiopia [35]. Data are updated in batches, monthly. Daily wind speed, wind direction, air pressure, relative humidity and temperature reports are recorded.

The wind measurement campaign that generated the data was commissioned by The World Bank with funding from the Energy Sector Management Assistance Program (ESMAP). It is available under The World Bank's open data policy. Each wind pole contains six different sensors, at different heights, from six to eighty meters, as shown in Table 5.

Table 5. Height of the sensors installed for collecting data for the Ethiopia Wind Measurement.

Sensor Type	Height (m)
Anemometer	80
Anemometer	80
Wind Vane	78
Thermometer	77
Anemometer	60
Wind Vane	58
Anemometer	40
Anemometer	20
Thermometer	10
Barometer	6
Relative Humidity	6

A predictive wind resource map can be built using the data collected by the sensors installed at different locations. Table 6 shows the location of each data collection pole. The sampling frequency is 1 Hz. However, data recorded are the average of 10 min of data samples. Until now there are no known missing data samples. Hence, the available data should be of high quality for mining and studying the variables during the recording period, which in some cases is more than one year and a half.

Sensors installed in different poles may come from different manufacturers. Data for each pole are aggregated in a metadata file that also contains information about the pole and the sensors' manufacturer, model and serial number.

Table 6. Location of the data collection poles of the Ethiopia Wind Measurement project.

Region	Latitude	Longitude	Start Date
Somali	10.434	42.231	26 December 2018
Gumuz	9.876	34.683	11 June 2019
Somali	10.823	42.503	3 June 2019
Somali	10.772	42.578	14 April 2018
Afar	11.882	41.567	5 May 2019
Somali	9.722	42.010	3 December 2018
Somali	9.753	41.883	12 December 2018
Somali	5.578	43.340	5 April 2019
Somali	8.973	43.250	15 April 2019
Tigray	13.560	39.563	16 June 2019
Oromia	4.339	37.792	16 June 2019
Amhara	9.949	39.630	26 April 2019
Somali	9.583	41.553	7 December 2018
Somali	9.652	42.719	18 April 2019
Somali	9.688	42.768	18 January 2019
Oromia	7.875	38.700	18 May 2019

6.5. Global Wind Atlas

The Global Wind Atlas (GWA) is a free, web-based application developed as an aid for policymakers, planners, and investors, to help identify high-wind areas for wind power generation, possibly anywhere in the world [36]. It is available at https://globalwindatlas.info (accessed on 17 August 2020).

GWA provides an online service where users can search through different queries. It also provides free downloadable datasets, grouped by different sections, based on the latest input data and modeling methodologies. Table 7 shows a summary of the downloadable sections in the Global Wind Atlas.

Table 7. Downloadable sections in the Global Wind Atlas. From [37].

Section	Description
Maps	Map View for global or specific region, with different output layers. Provides wind energy class, wind speed, and power density and terrain surface layers
High resolution poster map	Selection of wind speed potential and power density potential maps. Provides an estimate of mean wind power density at 100 m above surface level. The map is derived from high-resolution wind speed distributions-based on a chain of models, which downscale winds from global models (≈70 km), to mesoscale (9 km) and to microscale (150 m)

In the GWA it is also possible to download high-resolution maps of the wind resource potential at a global and country level for 10 m to 200 m height. This wind resource database is maintained in partnership by the Department of Wind Energy at the Technical University of Denmark and the World Bank group. The mesoscale model mentioned on Table 7 uses ECMWF ERA-5 reanalysis data for atmospheric sampling for the period 1998–2017. ERA5 provides hourly estimates of many atmospheric, land and oceanic climate variables. The data cover the Earth on a 70 km grid and resolve the atmosphere using 137 levels from the surface up to a height of 80 km. ERA5 includes information about uncertainties for all variables at reduced spatial and temporal resolutions.

The output at 3 km resolution is generalized and downscaled further using WAsP software plus terrain elevation data at 150 m resolution, and roughness data at 300 m resolution. The WAsP software suite is the industry-standard for wind resource assessment, siting and energy yield calculation for wind turbines and wind farms. Finally, the microscale is sampled on calculation nodes every 150 m. However, this modeling process becomes more uncertain, most likely leading to an overestimation of mean wind power values. Table 8 shows the spatial scales for different length scales of the GWA, according to [3].

Table 8. Spatial scales and wind types found in the Global Wind Atlas. Reprinted with permission [3]; Elsevier, 2020.

Spatial Scales	Wind Types	Length Scale
Planetary scale	Global circulation	10,000 km
Synoptic scale	Weather systems	1000 km
Meso-scale	Regional orographic or thermally induced circulations	10–100 km
Microscale	Local flow modulation, boundary layer turbulent gusts	100–1000 m

7. Wind Farm Monitoring

The uncertainty on revenue of the existing wind farm installations pressures operation and maintenance departments to reduce their costs, since they might come up to 30% in offshore environments [4]. Due to factors like the ones mentioned, monitoring systems focused on wind farms' behavior and main components have been increasingly installed to optimize maintenance planning. Their economic benefit has been investigated and proven to exist [4].

7.1. ENGIE, La Houte Bourne Wind Farm

ENGIE is a company that produces and distributes energy from different sources, including renewable sources. It is a major player in the production of green electricity, being the 1st wind power producer with installed capacity of 1730 MW, representing approximately 8% of France wind power production, according to power production in Table 9.

Table 9. ENGIE Turbines Technical Information, for the La Haute Borne wind farm, Vaudeville-le-Haut, France.

Wind Turbine Name	Manufacturer	Model	Rated Power (kW)
R80711	Senvion	MM82	2050
R80721	Senvion	MM82	2050
R80736	Senvion	MM82	2050
R80790	Senvion	MM82	2050

According to the company's strategy, ENGIE decided to open up, for public use, data of the La Houte Borne wind farm, which is operated by ENGIE Green. The farm's four wind turbines are all from the same model and manufacturer, and have been providing electricity to the equivalent of 7300 people since 2009, avoiding 12,000 metric tons of CO_2 emission per year. The data set is composed by two very big and complete spreadsheet files, one from 2013 to 2016 and another from 2017 to 2020. The files contain information about each wind turbine's components, such as rotor speed, nacelle temperature, mechanical information like the torque and, finally, wind speed, direction and pitch angle. All the variables are cataloged and described in another file named "Data Descriptions". Finally, ENGIE provides the static information about each turbine: ID number, manufacturer, model, rated power, hub height, rotor diameter and precise location, as described in Table 10. All the data are recorded with a ten-minute period, and there is a total of 1,057,868 observations.

Data are in a file with SCADA data about component control variables and meteorological mast. Data description is provided, with every variable abbreviation explained and units. Static information contains some of the turbine's technical characteristics.

Table 10. Information Available in the ENGIE Renewables open La Haute Borne dataset.

File	Variables

Data	Pitch Angle; Converter Torque; Power Factor; Generator Speed and Temp;
	Gearbox Bearing/Oil Temp; Nacelle Angle; Grid Freq/Voltage;
	Active/Reactive Power; Rotor Speed and Bearing Temp; Wind Speed (2 sensors)

Data Description	Every variable in La Houte Bourne Data file The file's name clarifies its purpose.
Static Information	Wind Turbine's Name; ID; Model; Manufacturer; Rated Power; Rotor Diameter; GPS Location

7.2. Sotavento Wind Farm

Sotavento wind farm [38], was put into operation in 2001 by Sotavento Galicia, S.A, in Xermade, Lugo, Spain, after the Galician Government had decided to increase investment in renewable energy, especially wind-based energy. Composed by 24 onshore turbines with a total nominal power of 17,560 kW it has an average annual generation of 33 MWh and produces the equivalent consumption to 1051 families, avoiding 0.36 MT of CO_2 emissions per hour and consumption of 0.68 barrels of petroleum.

Sotavento's platform is a reliable source of wind resource and wind farm output production. With a high-quality monitoring program, the database provides real time data for wind speed and direction, turbines' production and capacity factor, and finally temperature and density of the air.

All the information is given with a 10 s period. Additionally, it is possible to search for historical data, allowing researchers to study wind farm and meteorological mast behavior.

Table 11 shows a summary of the turbines in use in the wind farm. The historical data is logged with 10 min intervals, hourly and daily.

Table 11. Sotavento Turbines Technical Information.

Technology	Model	Rated Power (kW)	Number of Units
Ecotecnia	44/640	640	4
Gamesa	G-47	660	4
Izar-Bouns	MK-IV 600	600	4
Izar-Bonus	1.3	1300	1
Made	AE-46/I	600	4
Made	Serie AE-52	800	1
Made	AE-61	1300	1
Neg Micon	NM 48/750	750	4
Neg Micon	52/900	750	1

7.3. EDP Wind Farm

EDP (Energias de Portugal) is an important player in the energy sector, especially in the Iberian Peninsula, where it produces and distributes a large share of electricity. The dataset available provides two years of SCADA records from five offshore wind turbines located in the West African Gulf of Guinea [39]. The dataset consists of different files that give information about failure logs and technical information about some of the main turbine's components, such as the gearbox, generator and rotor. Additional information includes meteorological data, namely wind speed and direction, air pressure, humidity, temperature and component signals, namely generator RPMs and oil temperature in the hydraulic group. All files available are summarized in Table 12, including listing of all variables logged. The training set is from 2016 (all year) and the testing set consists of nine months of data, from 2017 (1 January 2017 to 1 September 2017). Meteorological mast data and component signals are recorded with a 10 min period and there is a total of 69,962 observations.

The Wind Turbine Characteristics file contains wind turbine main characteristics, Table 13. Also, it supplies wind turbine's power curve, a defining variable, at a 1.225 kg/m^3 air density.

The meteorological mast file logs important meteorological signals, namely: Anemometer sensors 1 and 2 are at a 80 m and 77 m height; Weather vanes are located at 77 m and 40 m height; and temperature and pressure sensors at 75 m and 100 m height.

The component signals file includes SCADA signals for each wind turbine's most important components and production values.

The failure logs file is an historical failure logbook for the wind farm. It logs replacement and repaired processes, errors, high signal values and component failures.

Table 12. Information available in the EDP Open Dataset.

File	Variables
Wind Turbine Characteristics	Power; Rotor; Gearbox; Generator; Tower; Power Curve

Meteorological Mast	Wind Speed and direction (2 anemometer sensors);
	Ambient Temperature and Air Pressure (2 sensors);
	Humidity; Precipitation

Component Signals	Generator RPM and Temp;
	Gearbox Oil Temp; Nacelle Temp;
	Total active and Reactive Power; Pitch Angle

Failure Logs	Every component from the wind turbine

Table 13. Wind turbine technical information.

Power	
Rated power (kW)	2000
Cut-in wind speed (m/s)	4
Rated wind speed (m/s)	12
Cut-out wind speed (m/s)	25
Rotor	
Diameter	90
Number of blades	3
Max Rotor speed	14.9
Power density (W/m^2)	314.4
Gearbox	
Type	Planetary/sour
Stages	3
Generator	
Type	Asynchronous
Max Speed (rpm)	2016
Grid frequency (Hz)	50
Tower	
Hub height (m)	80
Type	Steel tube

7.4. Yalova Wind Turbine Dataset

Yalova is an onshore Turkish wind farm, located in west Turkey. It comprises 36 wind turbines and a total nominal power of 54,000 kW, with two different turbine models, according to http://www.tureb.com.tr/bilgi-bankasi/turkiye-res-durumu (accessed on 18 May 2020). Table 14 summarizes the turbine models and characteristics.

The Yalova wind farm has been operating since 2016. A SCADA system was used to measure and save wind turbine's data from one of the turbines—which model is monitored is not specified in the dataset. The SCADA system logged wind speed and direction, generated power and the theoretical power based on the turbine's power curve. Each new line of data is stored at 10 min intervals. However, there are a few gaps and some generated power is missing, which can be explained as a wind turbine's malfunction, maintenance or the wind speed being lower than the cut-in speed. The dataset is available in CSV format and it is for a one-year period, at https://www.kaggle.com/berkerisen/wind-turbine-scada-dataset (accessed on 17 August 2020). All the information about the data available is summarized on Table 15.

Table 14. Yalova wind farm turbine models, in Yalova, Turkey.

Turbine Manufacturer	Turbine Model	Turbine Capacity
SINOVEL	SL 1500/90	1.5 MW
SINOVEL	SL 1500/82	1.5 MW

Table 15. Yalova wind turbine dataset information, in Yalova, Turkey.

Author	Not specified
Variables	Active power; Theoretical power; Wind speed; Wind direction
Draft Frequency	10 min
Start Period	1 January 2018
End Period	31 December 2018

7.5. Wind Turbine SCADA dataset

In Kaggle website there is a data file, downloadable at https://www.kaggle.com/wasuratme96/turbine-fault-prediction (accessed on 17 August 2020). Kaggle is one of the world's largest data science communities. One of the community members released a massive SCADA dataset from an unknown turbine. The turbine's location or identification were not disclosed, but the data set seems to provide a wealth of information, including wind speed, power production, operating hours, component monitoring and different turbine status.

The data are divided in two files: one with wind and turbine's behavior and the other with turbine status, including status "under maintenance". The former file contains logs of wind forecast, turbine's generating power, component's temperature and available power in the wind. The latter file contains logs of maintenance periods and failures. Table 16 shows some details of the files available in the dataset.

Table 16. Details of the SCADA Kaggle dataset files.

Header	SCADA Data	Status Data
Draft Frequency	10 min	Variable
Observations	49,028	1849
Start Period	5 January 2014	24 April 2014
End Period	4 September 2015	28 April 2015

8. Other Datasets

8.1. Elia

Elia is a Belgium's high-voltage transmission system operator, operating over 19,271 km of power lines and underground cables throughout Belgium. The company plays a crucial role in the community by transporting electricity from generators to distribution systems and consumers. Due to their location, Elia is a key player in the energy market and electricity system in Belgium. The company sets up multiple initiatives promoting the development of an efficient, transparent and fair electricity market, according to https://www.elia.be/en/company (accessed on 13 May 2020).

Elia continuously tracks and forecasts wind power generation in different turbines. It provides monthly grid data for each onshore and offshore wind production. The data are made available online at https://www.elia.be/en/grid-data (accessed on 4 July 2020).

8.2. NREL Data Catalog

The NREL Data Catalog contains descriptive information and public data, resulting from funded research conducted by NREL researchers and analysts. Data are available at https://data.nrel.gov (accessed on 14 May 2020). The site, however, contains some dead links.

8.3. Discussion

Most wind turbine capacity and wind farm projects are cataloged in large global databases, of which there is a summary in Table 17. Wind energy is a growing energy source, and there are important wind farms all around the world. Data from the existing wind farms, as well as wind resources, is summarized in Table 18. It may be useful for improving wind turbine maintenance policies as well as planning new wind farms. Table 19 shows a summary of the datasets that contain monitoring data of the wind farms' turbines.

Table 17. Wind turbine capacity databases.

	The Wind Power	USWTDB	UKWTDB
Author	-	U.S. DoEnergy & LBNL & USGS & AWEA	RenewableUK
Location	Global	United States	United Kingdom
Information	Developers, operators, owners, manufacturers	Turbine capacity database	100 kW and larger projects
Turbines	20,838 (Wind Farms)	63,794	10,607
Start Period	-	2018	-
End Period	-	2049	-

Table 18. Wind resource datasets summary.

	ALP	SONDA	Ethiopia WMD	Global Wind Atlas
Author	U.S. DOE's & WPA	LABREN CCST & INPE	ESMAP	DWE & World Bank
Location Information	U.S Native Reservations Wind Monthly Average Speed Frequency Direction Turbulence	Brazil Wind Speed Direction Temperature at 25 m and 50 m Temperature Turbulence	Ethiopia Wind Speed Direction Air Pressure Relative Humidity Mean Wind Speed	Global Wind Power Density Maps Wind Frequency
Draft	10 min	10 min	10 min average	-
Frequency	-	-	from 1 Hz draft	-
Start period	2000	-	2018	-
End period	2011	-	-	-

Table 19. Wind farm/turbine monitoring datasets.

	La Haute Bourne ENGIE	Sotavento	EDP	Yalova	WT Data Set
Author	ENGIE Group	Sotavento, SA	EDP	Unknown	Unknown
Location	France	Spain	Guinea Gulf	Turkey	Unknown
Information	Wind turbine's component monitoring meteorological mast	Technical data Wind resource output production Wind farm and CF	SCADA records divided training and test set Failure Logs	Wind Speed and Direction Generated and Theoretical Power	Wind Speed Generated Wind Density Component's Temperature Turbine Status Maintenance Failures
Draft Frequency	10 min	10 min	10 min	10 min	10 min variable
WF Type	Onshore	Onshore	Offshore	Onshore	Unknown
WF Capacity	8 MW	17.56 MW	10 MW	54 MW	Unknown
Format File	CSV	CSV	CSV	CSV	CSV
Start	2013	-	2016	-	2014
End	2020	-	2017	-	2015

Developers and operators around the world are constantly making efforts to optimize the wind farms to their maximum potential. Supervising production and monitoring the state of the turbines, it is possible to develop a better insight into wind turbines' operation. It is important not only to monitor wind turbines' operation, but also the main resource for this energetic system, which is the wind. Since the wind is the main variable and input, a better understanding of that resource will allow the design of better wind farm projects, the discovery of patterns and the prediction of behaviors.

The meteorological mast is almost always monitored, due to its importance to the turbine' operation. Some datasets, namely EDP and ENGIE, focus on wind turbines' main components operational data, alarm logs and failures. Monitoring systems are increasingly installed in wind turbines to provide specific information that will help increase equipment availability [4]. That information is usually used for predictive maintenance plans, which means to do a prognostic and find patterns in raw data, preventing faults and failures before they occur, avoiding the costs of failure.

Deciding which components to monitor is always important, but priority should be given to the ones with higher failure rate and those which need more time to be repaired. The optimization of wind power systems has had a considerable progress during the last decade and the costs to maximize their production are viewed with growing concerns. Wind farms' production costs are not negligible, and the maintenance and optimization of the large power systems is very important, not only to make it a profitable power system, but also to create competitive advantages against fossil fuels.

Each dataset is different, but they all have similarities. When it comes to meteorological mast, wind speed and direction are monitored. All datasets have a 10 min sampling period. It is a normal procedure for SCADA systems that store values of parameters and characterize operating and environmental conditions. However, there are different opinions about which frequency should be used to store the data. According to [25], most data contributions that rely on 10 min averaged SCADA may be negatively affected due to rapid wind speed and power output fluctuations, leading to a non-efficient understanding of wind turbines' dynamic. On the other hand, some authors advocate that when using high-frequency data it will be highly affected by "noise" and 10 min intervals will smooth it. Using larger periods will result in loss of information and inaccurate data models. The study shows that 30 s intervals provide reasonable balance between resolution and dynamic response, but the mean absolute error shows lower values for 10 min training. Thus, highest accuracy is achieved with higher sampling rate, but it also achieved a higher variance error [25]. A 10 min rate for sampling SCADA systems is accepted.

9. A Deeper Overview of EDP Dataset

As mentioned in Section 7.3, EDP dataset is one of the most complete datasets available. Among other information, it is possible to search for relations between turbine failures and their behavior days before the failure.

9.1. Wind Turbine Operation Behavior

Wind Turbines have three main regions of operation, [40], as shown in Figure 1. The regions are:

- Region 1: Includes the time when the turbine is starting up;
- Region 2: Operational region in which it is desirable to seize as much wind power as possible;
- Region 3: Wind speeds are relatively high (rated wind speed) and force the turbine to limit the fraction of wind power captured, for electrical and mechanical safety.

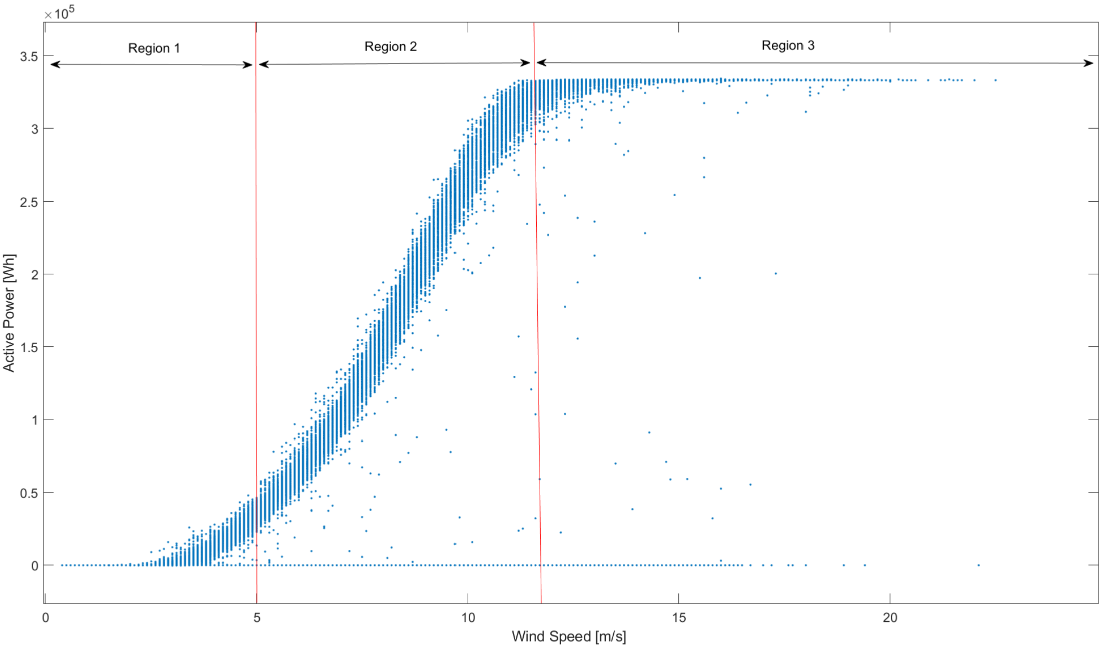


Figure 1. Wind turbine operation regions.

Table 20 shows the number of failures counted in the EDP dataset, grouped by turbine component group and sorted in descending order by number of failures counted. In the table it is possible to identify three groups of components more often affected by failures. They are: (i) Generator; (ii) Generator Bearing; and (iii) Hydraulic Group. Figure 2 shows that Turbine 06 and Turbine 09 are the two most affected turbines—the former counts seven failures, the latter counts five failures.

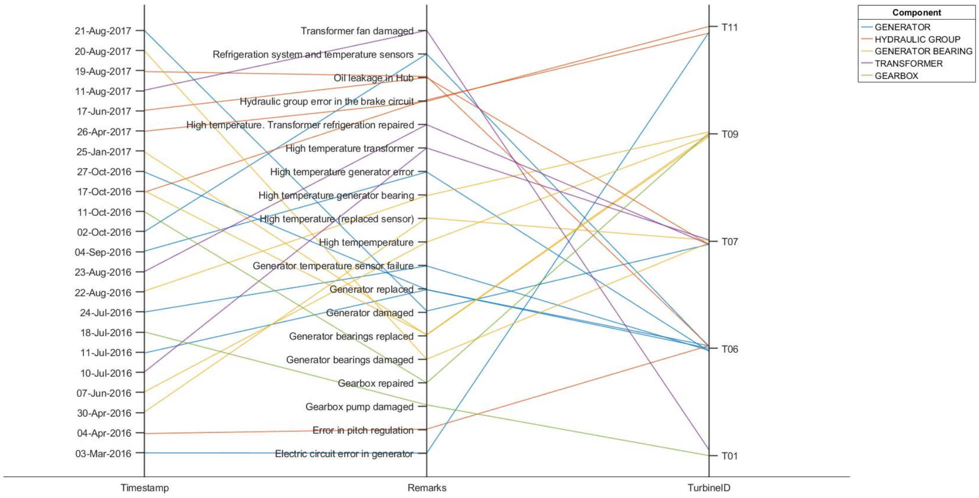
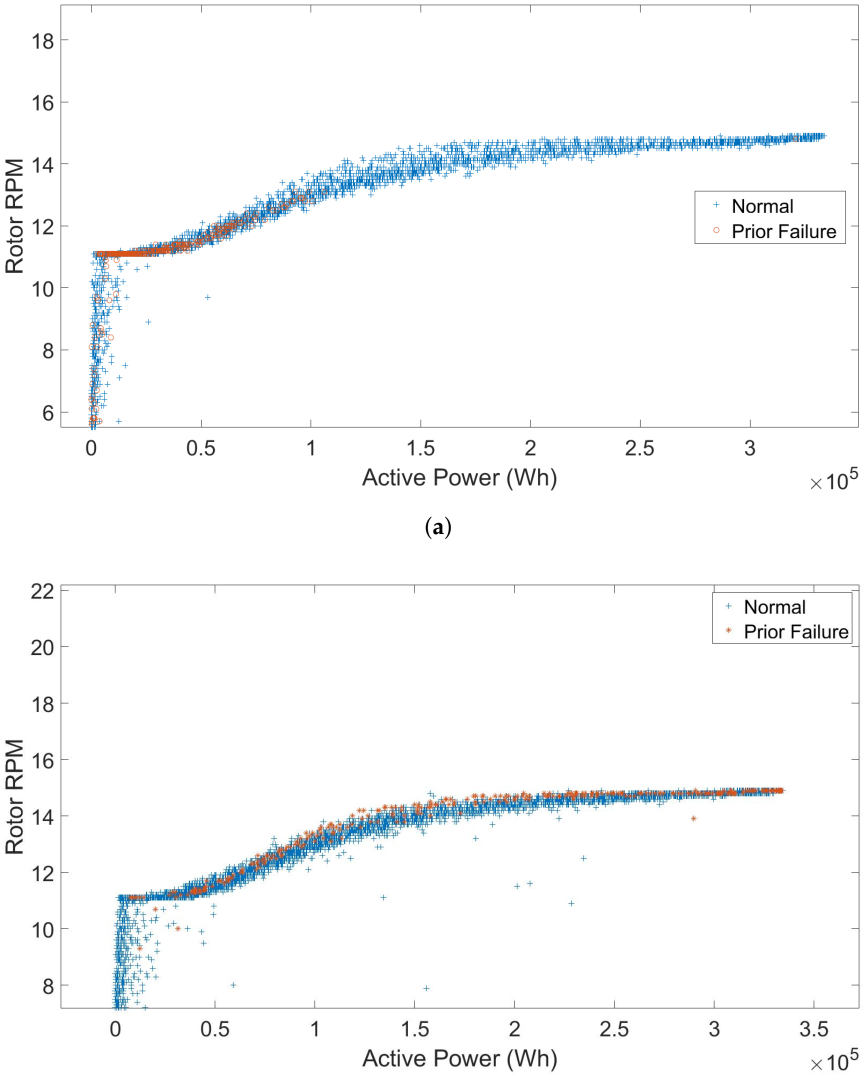


Figure 2. Chronological plot of the failures recorded in the EDP dataset.

Table 20. Number of failures for each component group, counted in EDP dataset.

Component's Group	Failures
Generator	7
Generator Bearing	6
Hydraulic Group	5
Transformer	3
Gearbox	2
Total	23

Figure 3a,b show plots that help identifying in which region the turbines were operating before failure. Prior to failure in turbines T06 and T11, Region 2, the desirable operational region of operation, is also identified as the one more prone to failure. Despite the fact that turbine T06 shows some prior to failure behavior near Region 3 of operation, most of the observations still tend to Region 2.



(**a**)

(**b**)

Figure 3. Charts of Active Power and Rotor Behavior. Observations of the turbines' normal behavior are plotted in plus symbols, observations prior to failure in star symbols (**a**) Normal observations (blue) and observations prior to generator failure (orange), for turbine T11; (**b**) Normal observations (plus symbols) and observations prior to generator failure (star symbols), for turbine T06.

9.2. Principal Component Analysis

PCA is a statistical method to identify patterns in data and express them in a way to highlight the similarities and differences, through a graphical representation. PCA is also a method to compress the size of the datasets, reducing the noise, removing outliers and simplifying data description.

Table 21 shows the correlations between different variables of the observations. The variables are: (i) Generator's rotations per minute; (ii) Generator's bear temperature; (iii) Hydraulic oil temperature; (iv) Gear oil temperature; (v) Nacelle temperature; (vi) Rotor's rotations per minute; (vii) Wind speed; (viii) Ambient temperature.

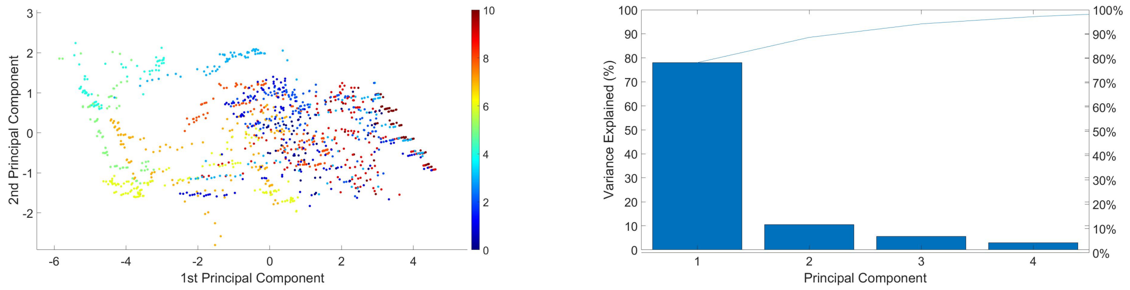
As the table shows, some of the variables have high percentage of correlation, which means their behavior is highly correlated. For example, the Generator RPM shows very high correlations ($R > 0.75$) with Rotor RPM, which is easily understood , since the behavior of one component is directly connected with the other. Nacelle temperature has high correlation with other temperature-related variables, less with the hydraulic oil temperature. Ambient temperature does not really correlate with any variable chosen. We might have thought otherwise, but this case shows no behavior correlation.

Table 21. Correlation between variables of the EDP dataset.

	Gen RPM	Gen Bear Temp	Hyd Oil Temp	Gear Oil Temp	Nac Temp	Rtr RPM	Wind Speed	Amb Temp
Gen RPM	1	0.6642	0.2672	0.7688	0.8678	0.9678	0.6727	0.1616
Gen Bear Temp	0.6642	1	0.6116	0.7802	0.8291	0.6929	0.8023	0.4522
Hyd Oil Temp	0.2672	0.6116	1	0.4590	0.4416	0.2793	0.5154	0.6592
Gear Oil Temp	0.7689	0.7802	0.4590	1	0.9555	0.7912	0.7488	0.3383
Nac Temp	0.8678	0.8291	0.4416	0.9555	1	0.8918	0.7935	0.3120
Rtr RPM	0.9678	0.6929	0.2793	0.7912	0.8918	1	0.7108	0.1530
Wind Speed	0.6727	0.8023	0.5154	0.7488	0.7935	0.7108	1	0.2616
Amb Temp	0.1616	0.4522	0.6592	0.3383	0.3120	0.1530	0.2616	1

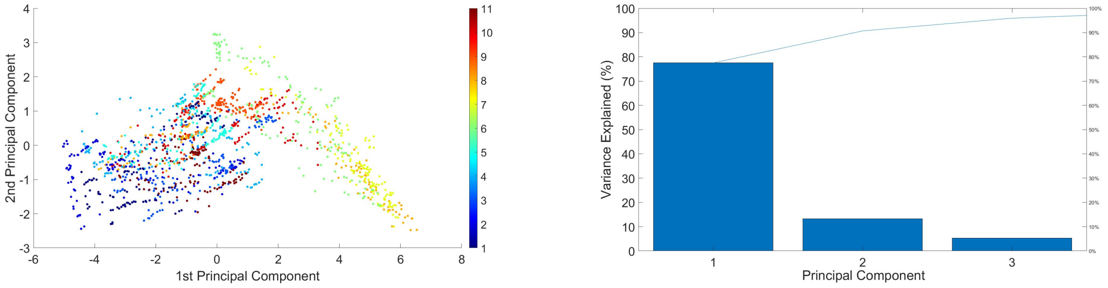
Since there are more than 200,000 observations, it is not possible to plot the raw SCADA records. An alternative is to compress the data to a lower number of dimensions, based on PCA approaches.

The generator component is the one with more failures recorded. Hence, it was chosen for this PCA implementation. Nine variables with high correlation levels were selected, from two distinct turbines, creating a 9-dimension plot. Using PCA it was possible to reduce the observations-variables plotting to only a 2-dimensional chart, without losing relevant levels of information. Figures 4b and 5b show that by only using the first two principal components, it is still possible to have over 80% of variance explained. PC plots in Figures 4a and 5a show that zero, one and two days before failure, turbine's behavior is clustered with approximately eight, nine and ten days before failure. This observation might indicate that it may be possible to predict a high probability of failure 10 days before it occurs.



(**a**) (**b**)

Figure 4. Principal Component Analysis for Turbine 06. (**a**) PCA for 10 days before T06 generator's failure; (**b**) Pareto's variance analysis for each PC used.

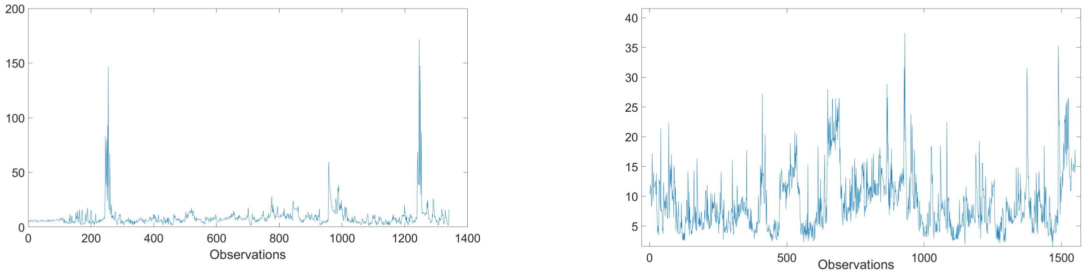


(**a**) (**b**)

Figure 5. Principal Component Analysis for Turbine 11. (**a**) PCA for 10 days before T11 Generator's failure; (**b**) Pareto's variance analysis for each PC used.

T^2 levels are often used for control charts, which give a statistical measure of the multivariate distance between each observation from the center of the dataset. Figure 6 shows the plot of T^2 for both turbines. As the plots show, T06 has a less erratic behavior and, as a first reading and impression, the process looks under control, with only two out of control moments. T11 has a more scattered T^2 plot, but still it is possible to identify some peaks that might surpass control chart's limits.

Although we do not have values for control charts, by looking at T^2 plots, Figure 6, we see similarities too and we can connect them it the PCA plot. The statistical measure has high peaks in the first observations (ten, nine and eight days before failure) and further close to failure.



(**a**) T^2 for T06 (**b**) T^2 for T11

Figure 6. Statistical measure of the multivariate distance of each observation from the dataset center.

10. Main Contributions

The present paper proposes some novel contributions to the state of the art. These can be highlighted:

- Survey of open datasets related to wind, wind energy and wind turbine's operation, which can be used for data analysis and knowledge extraction;
- Overview of correlations between variables deemed as more important for turbine monitoring;
- Insight into wind turbine´s behavior for different values of wind speed and identification of three main operating regions, such as the one that causes more failures;
- Identification of data clusters, using PCA;
- Use of statistical measures to identify out of control/failure behavior.

11. Conclusions

Wind energy has been the main renewable energy source in recent decades and it has potential to continue growing. Good quality open datasets of wind resources, wind farms and wind farm operation are fundamental for researchers, to extract knowledge and advance future research. The present paper proposes, therefore, a comprehensive survey of existing datasets, with their advantages and limitations. A total of 15 open datasets were analyzed, 13 of which have good quality for machine learning applications. The main characteristics of good quality datasets have been pointed out. This is an important contribution to facilitate future research in the field.

A deeper analysis of one of the most complete wind farm operation datasets available also provides these conclusions:

- The performance of component groups differs, and the faultiest behavior was identified;
- Wind turbine's region 2 of operation is where more failures occur, even though it requires less mechanical effort;
- Using PCA it may be possible to predict a turbine failure up to 10 days before the actual failure.

Future work includes the use of the knowledge extracted from the datasets to improve turbine maintenance plans, so that the number of failures, downtime and corrective maintenance costs may be reduced.

Author Contributions: All authors have contributed equally in all the steps of the present work. All authors have read and agreed to the published version of the manuscript.

Funding: The authors acknowledge Fundação para a Ciência e a Tecnologia (FCT) for the financial support to the project UIDB/00048/2020.

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

Abbreviations

The following abbreviations are used in this manuscript:

AD	Actuator Disk
AS	Actuator Surface
ALP	American Loan Program
AWEA	American Wind Energy Association
CF	Capacity Factor
CFD	Computational Fluid Mechanics
CSV	Comma Separated Value (file format)
CCST	Centro de Ciência do Sistema Terrestre/ Earth System Science Center
DOE's	Department of Energy's
DOE Data	Department of Energy Open Data Catalog DOE Data
DWE	Department of Wind Energy
ESMAP	Energy Sector Management Assistance Program
INPE	Instituto Nacional de Pesquisas Espaciais/ National Institute for Space Research
IURDB	International Utility Rate Database
LABREM	Laboratório de Modelagem e Estudos de Recursos Renováveis de Energia/ Laboratory for Modeling and Studies of Renewable Energy
LES	Large-Eddy Simulation
LBNL	Lawrence Berkeley National Laboratory
NREL	National Renewable Energy Laboratory
OEDI	Energy Data Initiative OEDI
PC	Principal Component
PCA	Principal Component Analysis
RPM	Rotations Per Minute
SCADA	Supervisory Control and Data Acquisition
SD	Standard Deviation
SONDA	National Organization System for Environment Data (Brazil)
UKWED	United Kingdom Wind Energy Data Base
URDB	United States Utility Rate Database
USGS	United States Geological Survey

USWTDB	United States Wind Turbine Data Base
WPA	Wind Powering America
WPD	Wind Power Density
WMR	Wind Measurement Data
XML	eXtensible Markup Language

References

1. Wee, H.M.; Yang, W.H.; Chou, C.W.; Padilan, M.V. Renewable energy supply chains, performance, application barriers, and strategies for further development. *Renew. Sustain. Energy Rev.* **2012**, *16*, 5451–5465. [CrossRef]
2. Hau, E. *Wind Turbines: Fundamentals, Technologies, Application, Economics*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2005; ISBN 978-3540242406.
3. Letcher, T.M. *Wind Energy Engineering: A Handbook for Onshore and Offshore Wind Turbines*; Elsevier: Amsterdam, The Netherlands, 2017; ISBN 978-0128094518.
4. González-Aparicio, I.; Monforti, F.; Volker, P.; Zucker, A.; Careri, F.; Huld, T.; Badger, J.L. Towards quantification of condition monitoring benefit for wind turbine generators. In Proceedings of the European Wind Energy Conference & Exhibition, Milan, Italy, 7–10 May 2007; pp. 1–11.
5. González-Aparicio, I.; Monforti, F.; Volker, P.; Zucker, A.; Careri, F.; Huld, T.; Badger, J.L. Simulating European wind power generation applying statistical downscaling to reanalysis data. *Appl. Energy* **2017**, *199*, 155–168. [CrossRef]
6. EMHIRES Dataset Part I: Wind Power Generation. Available online: https://setis.ec.europa.eu/publications/relevant-reports/emhires-dataset-part-i-wind-power-generation (accessed on 1 July 2020).
7. Diffendorfer, J.E.; Kramer, L.A.; Ancona, Z.H.; Garrity, C.P. Onshore industrial wind turbine locations for the United States up to March 2014. *Sci. Data* **2015**, *2*, 1–8. [CrossRef] [PubMed]
8. USGS; Berkeley Lab; AWEA. The U.S. Wind Turbine Database. Available online: https://eerscmap.usgs.gov/uswtdb/ (accessed on 17 June 2020).
9. Van Vuuren, C.J.; Vermeulen, H.J. Clustered wind resource domains for the South African renewable energy development zones. In Proceedings of the IEEE 2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA), Bloemfontein, South Africa, 28–30 January 2019; pp. 616–623.
10. Kusiak, A.; Zheng, H.; Song, Z. Short-term prediction of wind farm power: A data mining approach. *IEEE Trans. Energy Convers.* **2009**, *24*, 125–136. [CrossRef]
11. Shevade, S.K.; Keerthi, S.S.; Bhattacharyya, C.; Murthy, K.R.K. Improvements to the SMO algorithm for SVM regression. *IEEE Trans. Neural Netw.* **2000**, *11*, 1188–1193. [CrossRef] [PubMed]
12. Smola, A.J.; Schölkopf, B. A Tutorial on Support Vector Regression. *Stat. Comput.* **2004**, *14*, 199–222. Available online: https://alex.smola.org/papers/2004/SmoSch04.pdf (accessed on 17 August 2020). [CrossRef]
13. Witen, I.H.; Frank, E. Data mining: practical machine learning tools and techniques with Java implementations. *ACM Sigmod Rec.* **2002**, *31*, 76–77. [CrossRef]
14. Mining, D. *Practical Machine Learning Tools and Techniques*; Morgan Kaufmann: Burlington, MA, USA, 2005.
15. Wang, Y.; Witten, I.H. *Induction of Model Trees for Predicting Continuous Classes*; (Working Paper 96/23); University of Waikato, Department of Computer Science: Hamilton, New Zealand, 1996.
16. Frank, E.; Wang, Y.; Inglis, S.; Holmes, G.; Witten, I.H. Using model trees for classification. *Mach. Learn.* **1998**, *32*, 63–76. [CrossRef]
17. Hothorn, T.; Lausen, B. Bundling classifiers by bagging trees. *Comput. Stat. Data Anal.* **2005**, *49*, 1068–1078. [CrossRef]
18. Dietterich, T.G. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Mach. Learn.* **2000**, *40*, 139–157. [CrossRef]
19. Ti, Z.; Deng, X.W.; Yang, H. Wake modeling of wind turbines using machine learning. *Appl. Energy* **2020**, *257*, 114025. [CrossRef]
20. Wu, Y.T.; Liao, T.L.; Chen, C.K.; Lin, C.Y.; Chen, P.W. Power output efficiency in large wind farms with different hub heights and configurations. *Renew. Energy* **2019**, *132*, 941–949. [CrossRef]

21. Lin, M.; Porté-Agel, F. Large-Eddy Simulation of Yawed Wind-Turbine Wakes: Comparisons with Wind Tunnel Measurements and Analytical Wake Models. *Energies* **2019**, *12*, 4574. [CrossRef]
22. Li, Z.; Yang, X. Evaluation of Actuator Disk Model Relative to Actuator Surface Model for Predicting Utility-Scale Wind Turbine Wakes. *Energies* **2020**, *13*, 3574. [CrossRef]
23. Uchida, T. Effects of Inflow Shear on Wake Characteristics of Wind-Turbines over Flat Terrain. *Energies* **2020**, *13*, 3745. [CrossRef]
24. Pessanha, J.F.M.; Barcelos, G.F.B.; Faria, A.V.C.; Ferreira, V.M.F. Análise Estatística de Registros Anemométricos e Seleção de Turbinas Eólicas: Um Estudo de Caso. In Proceedings of the XLII Simpósio Brasileiro de Pesquisa Operacional, Bento Gonçalves, Brazil, 1–4 September 2009.
25. Gonzalez, E.; Stephen, B.; Infield, D.; Melero, J.J. Using high-frequency SCADA data for wind turbine performance monitoring: A sensitivity study. *Renew. Energy* **2019**, *131*, 841–853. [CrossRef]
26. Smith, L.I. A Tutorial on Principal Components Analysis. University of Montreal. 2002. Available online: http://www.iro.umontreal.ca/~{}pift6080/H09/documents/papers/pca_tutorial.pdf (accessed on 17 August 2020).
27. Abdi, H.; Williams, L.J. Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 433–459. [CrossRef]
28. RodRigues, J.; FARinhA, J.T.; Mendes, M.; MARgALho, L. Predicting motor oil condition using artificial neural networks and principal component analysis. *Eksploat. Niezawodn. Maint. Reliab.* **2020**, *21*, 440–448. [CrossRef]
29. Kim, K.; Parthasarathy, G.; Uluyol, O.; Foslien, W.; Sheng, S.; Fleming, P. Use of SCADA data for failure detection in wind turbines. In Proceedings of the ASME 2011 5th International Conference on Energy Sustainability, American Society of Mechanical Engineers Digital Collection, Washington, DC, USA, 7–10 August 2011; pp. 2071–2079.
30. Smith, A.Z.P. What does the Capacity Factor of Wind Mean? Available online: https://energynumbers.info/capacity-factor-of-wind (accessed on 24 June 2020).
31. Kitchin, R.; McArdle, G. What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data Soc.* **2016**, *3*, 2053951716631130. [CrossRef]
32. RenewableUK. Wind Energy Statistics. Available online: https://www.renewableuk.com/page/UKWEDhome (accessed on 15 June 2020).
33. United States Government; NREL; Alliance for Sustainable Energy. OpenEi Datasets. Available online: https://openei.org/datasets/dataset (accessed on 1 July 2020).
34. INPE; CCST. SONDA—Sistema de Organização Nacional de Dados Ambientais. Available online: http://sonda.ccst.inpe.br/index.html (accessed on 1 July 2020).
35. Ethiopia-Wind Measurement Data. Available online: https://energydata.info/dataset/ethiopia-wind-measurement-data (accessed on 15 May 2020).
36. World Bank Group; ESMAP; Technical University of Denmark; Vortex. Global Wind Atlas. Available online: https://globalwindatlas.info (accessed on 1 July 2020).
37. Wikimedia Commons. Available online: https://commons.wikimedia.org/w/index.php?title=File:Global_Map_of_Wind_Speed.png&oldid=401722013 (accessed on 3 September 2020).
38. Sotavento Galicia Foundation. Parque Eólico Experimental Sotavent. Available online: http://www.sotaventogalicia.com/en/ (accessed on 20 May 2020).
39. EDP Group. EDP Open Data. Available online: https://opendata.edp.com/explore/?refine.keyword=visible&sort=modified (accessed on 1 May 2020).
40. Johnson, K.E.; Pao, L.Y.; Balas, M.J.; Kulkami, V.; Fingersh, L.J. Stability analysis of an adaptive torque controller for variable speed wind turbines. In Proceedings of the 2004 43rd IEEE Conference on Decision and Control (CDC) (IEEE Cat. No. 04CH37601), Nassau, Bahamas, 14–17 December 2004; Volume 4, pp. 4087–4094.