Received: 16 February 2022 | Revised: 17 May 2022 | Accepted: 23 June 2022

DOI: 10.1002/we.2766

REVIEW ARTICLE

WILEY

A collection and categorization of open-source wind and wind power datasets

Nina Effenberger | Nicole Ludwig

Cluster of Excellence Machine Learning, University of Tübingen, Tübingen, Germany

Correspondence
Nina Effenberger, Cluster of Excellence Machine Learning, University of Tübingen, Tübingen, Germany.
Email: nina.effenberger@uni-tuebingen.de

Abstract

Wind power and other forms of renewable energy sources play an ever more important role in the energy supply of today's power grids. Forecasting renewable energy sources has therefore become essential in balancing the power grid. While a lot of focus is placed on new forecasting methods, little attention is given on how to compare, reproduce and transfer the methods to other use cases and data. One reason for this lack of attention is the limited availability of open-source datasets, as many currently used datasets are non-disclosed and make reproducibility of research impossible. This unavailability of open-source datasets is especially prevalent in commercially interesting fields such as wind power forecasting. However, with this paper, we want to enable researchers to compare their methods on publicly available datasets by providing the, to our knowledge, largest up-to-date overview of existing open-source wind power datasets, and a categorization into different groups of datasets that can be used for wind power forecasting. We show that there are publicly available datasets sufficient for wind power forecasting tasks and discuss the different data groups properties to enable researchers to choose appropriate open-source datasets and compare their methods on them.

KEYWORDS
open-source data, time series forecasting, wind power forecasting

1 | INTRODUCTION

Over the past few years, the absolute and relative amount of sustainable and renewable energy sources integrated into the power grid has grown. Among these energy sources, wind power is the most common in many areas such as the European Union.[1] Wind power is both renewable and sustainable as it has only a minor impact on the environment, and its main resource, the wind, is not exhaustible. However, wind power is less reliable than unsustainable energy sources such as gas or coal due to the stochasticity of wind and weather in general. Furthermore, wind power cannot be generated when there is no wind, and alternative sources of energy have to be included into the power system to balance out the wind fluctuations. Therefore, sustainable energy forecasting is essential when it comes to predicting times at which alternative energy sources or measures to change the demand behavior have to be taken in order to stabilize the grid.

Wind power forecasting is hence crucial for an efficient interplay between the different kinds of power and can be divided into different tasks. Among these tasks are predicting the actual power generation, variability of the wind or quick and large changes in the power generation.[2] Independent of the forecasting task, wind power forecasting can be performed on different time scales, ranging from very short (≤30 min) to

long-term (several days to months) and on different spatial scales, ranging from individual turbines to whole regions.[3] While much literature focuses on wind power forecasting tasks on different spatial and temporal scales, little attention is given to the datasets used as the basis of this research. However, the underlying datasets are crucial, as using different datasets for the same tasks does not allow properly comparing the models and keeping the datasets private, instead of publicly available, renders reproducibility impossible.

In order to tackle this lack of comparability, Kariniotakis et al[4] compare different models using different but fixed datasets. They reveal a dependence between the performance of the models and the complexity of the terrain on which the wind farm is situated. The authors show that depending on the turbine's environment, the normalized mean absolute errors (NMAEs) computed on the same time horizon can be more than three times higher (NMAE > 35% in contrast to NMAE < 10%) for turbines on complex terrain in contrast to flat terrain. Additionally to the environment of a turbine, other close-by turbines influence the power output of individual wind turbines.[5] These turbine–turbine interactions play a role in both offshore and onshore wind power forecasting (e.g.Barthelmie et al.,[6] McKay et al.[7]). Thus, different terrain complexities and turbine–turbine interactions make it hard or even impossible to compare methods tested and validated on different datasets. Additionally, prediction errors depend on the forecast horizon.[8] However, we argue that most issues can be overcome by comparing models using the same open-source data.

Previous work has examined several of these open-source datasets. For example, Menezes et al[9] present datasets with a focus on wind farms and resources, most of the datasets they describe are open-source and some of them are also part of this survey. While the authors cover a broader range of wind resource datasets, we focus on wind power data and cover various additional and new datasets. Clifton et al[10] provide an overview of energy-related wind (and solar) resource datasets. Such wind resource data aim to quantify the amount of wind available for conversion into wind power and can, for example, be used to identify potential construction sites for future wind farms. However, Clifton et al[10] do not include any real turbine level datasets.

The lack of literature giving an overview of currently available open-source datasets hinders open science. Confidential data limit the reproducibility and comparability of research and stand in contrast to the dogma of open research. In 2016, Wilkinson et al[11] proposed a set of principles and guidelines that should improve Findability, Accessibility, Interoperability, and Reusability of data. These so-called FAIR data principles are widely known, accepted and applied in various scientific research fields (e.g., El-Gebali et al.,[12] Sinaci et al.,[13] Vuong et al.,[14] Frank et al.[15]). Open-source data are not only seen as necessary in research; political institutions such as the Cabinet of Germany (Deutsche Bundesregierung)[16] and the European Union[17] also emphasize the importance of open data. Nevertheless, these scientific and political ambitions are often not implemented, and the known issue of the result's dependence on the data is rarely addressed.

Therefore, to bring forward open science in the wind power forecasting community, we present and categorize open-source datasets that can be used for wind power forecasting. Furthermore, our categorization and detailed descriptions simplify and motivate the use of non-confidential, open-source wind power data. The remainder of this paper is organized as follows: Section 2 is the main focus of this paper and describes and categorizes the open-source datasets that contain wind and wind power data. We then evaluate the datasets properties in Section 3 to help choose an appropriate dataset before discussing known systematic errors in wind data in Section 4 and data quality in Section 5. We discuss our approach and selection of datasets in Section 6 and conclude in Section 7.

2 | OPEN-SOURCE WIND (POWER) DATASETS

In this section, we give an overview of over forty datasets for wind power forecasting. We compiled the datasets listed in this paper in several different ways, mainly by searching online for datasets (e.g., on the webpages of the IEA 36[18] or the WRAG[19]), contacting researchers from different continents and looking into papers that work with disclosed data. We first introduce the groups into which we separate the datasets and give some insights into each group, before we present the groups separately. We mainly differentiate between two groups of data, namely, *wind power data* and *wind-based data*. We then further split the wind power data into three subgroups and the wind-based data into two subgroups, resulting in a total of five different groups of data. Figure 1 shows the connections between the five data groups. These five subgroups have different characteristics, which allow us to assign each of the found open-source datasets into one of the groups. All data groups include wind data, but we can differentiate them using different characteristics: whether they contain additional weather measurements or weather data derived by weather models, include real or synthetic wind power data, include turbine-specific control parameters, are measured on turbine level, and include turbine-level supervisory control and data acquisition (SCADA) information. Table 1 gives an overview over the five data groups.

This distinction into five groups forms the basic framework for this paper, and the rest of this section is therefore structured accordingly. We give some information on how our found datasets are distributed over the five data groups, the world and other descriptives in Section 2.1. We then introduce the datasets that we categorize as wind power data in Section 2.3, before introducing the remaining datasets categorized as wind-based data in Section 2.4. As we present the datasets in large tables, we also introduce how these tables are structured and what information they contain in Section 2.2.

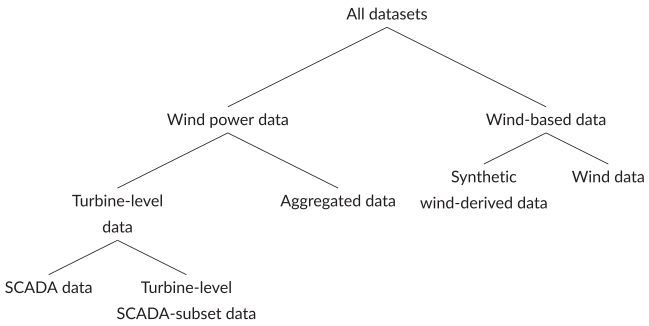

FIGURE 1 All of the presented datasets can be grouped into one of five different categories. The two supergroups are wind power data and wind-based data. The wind power datasets can be divided into the subgroups turbine-level and aggregated data. Turbine-level datasets contain measurements on turbine level and aggregated data are spatially aggregated on different levels from farm to country level. All of the turbine-level datasets include wind and wind power measurements, SCADA data contain more variables than turbine-level SCADA-subset data. The wind-based datasets contain either wind measurements or synthetic power data that is derived from numerical weather prediction models

2.1 | Descriptive statistics

The above described categorization into five groups is unique, meaning that each dataset is associated with exactly one group. In this subsection, we want to take a look at the distribution of the datasets over these five groups and further properties that separate the datasets. Figure 2 visualizes all properties and their composition within the entirety of the datasets.

Regarding the data groups, most of the datasets contain SCADA (11 datasets), SCADA-subset data (13 datasets), or aggregated data (nine datasets), while the smallest group consists of synthetic data (three datasets). More than 75% of the datasets contain real wind power data and more than 50% of these data were collected on turbine-level. Wind power and turbine-level data are most important for very short to short-term forecasting, for long-term forecasting location information becomes even more important as it allows to include additional weather data from weather models. Most datasets cover at least 1 year of data. This temporal coverage is important to investigate seasonal changes and include these into the model. However, due to yearly seasonality, it is generally assumed that 1 year of data is not sufficient to discover trends in wind speed or wind power. Vargas et al[20] investigate 145 different models and come to the conclusion that most models for long-term forecasting use hourly data (49%). Thirty-seven of the datasets that we present here provide or can be aggregated to this resolution. Most commonly in our selection of datasets, the data are provided in 10-min intervals. This resolution is generally assumed to be high enough for wind power prediction addressing grid integration but it too coarse to, for example, control the turbine.

While different problems and research questions require different datasets, two general statements can be made. First, location information is always beneficial as it allows to include variables from weather models. Second, due to the possibility of aggregating data, a finer temporal resolution on a large time span is also advantageous -- whether aggregated wind power data, pure wind data, synthetic data, or any other type of data are needed to answer a specific research question lies up to the researcher. If data are not readily available, it might then be worth thinking about either adapting the research question or shifting the spatial focus, for example to a location where more or higher resoluted data are publicly available.

The datasets stem from different locations as presented on the map in Figure 3. This map shows that there exist open-source datasets from all European countries, one achievement of the European association for the cooperation of transmission system operators for electricity (ENTSO-E).[21] Furthermore the datasets also cover Africa, North and South America, and Australia. However, among the datasets in this paper, none are from Asia or the Poles.

2.2 | Guide through the tables

We present all datasets in tables in their corresponding subsection. To make these tables easier to understand, we explain their structure in the following. The first column in each table, **Dataset**, gives an identifier of the dataset, which is, if possible, related to the location. We also provide a reference in this column, the first citation usually refers to the web page where the dataset is stored. If a paper exists, it is also referenced in this first column. Additional documentation or additional data can be found in the column **Information**. If accessible, the column **Location** provides the country were the data was collected or the region for which it was synthetically generated. The column **Coordinates** then gives the specific coordinates to this location, if available. The column **Time span** provides the time span the dataset covers and the column **Interval** the temporal distance between two successive data points. This interval is not always equal to the distance between two measurements, but measurements are averaged over this interval. Lastly, the column **Origin, License** gives an overview of the data providers and sources and we also present

TABLE 1 Overview of the characteristics of the five different data groups

Data group	Wind measurements	Weather measurements	Wind and weather data from weather models	Real wind power data	Synthetic wind power data	Turbine-specific control parameters	Turbine-level data	SCADA data	Key characteristics
SCADA data (Section 2.3)	✓	✓	-	✓	-	✓	✓	✓	Includes SCADA data
Turbine-level SCADA-subset data (Section 2.3)	✓	partly	-	✓	-	rarely	✓	-	Does not include full SCADA data but turbine-level measurements
Aggregated data (Section 2.3)	✓	partly	-	✓	-	-	-	-	Data are aggregated spatially
Meteorologically derived data (Section 2.4)	-	-	✓	-	✓	-	-	-	No wind power measurements are contained
Wind data (Section 2.4)	✓	partly	✓	-	-	-	-	-	No wind power data, only wind data

Note: All of the presented datasets can be assigned to one of five data groups. The fine-grained datasets are SCADA data and other turbine-level data. Aggregated datasets do also contain wind power data but weather data are not included and can usually not be applied directly. The two last data groups are based on wind data and contain either synthetic power data or pure wind measurements. While synthetic wind power data could theoretically be generated from real wind and weather measurements, all the datasets that we present in Table 6 are derived using data from publicly available weather models.

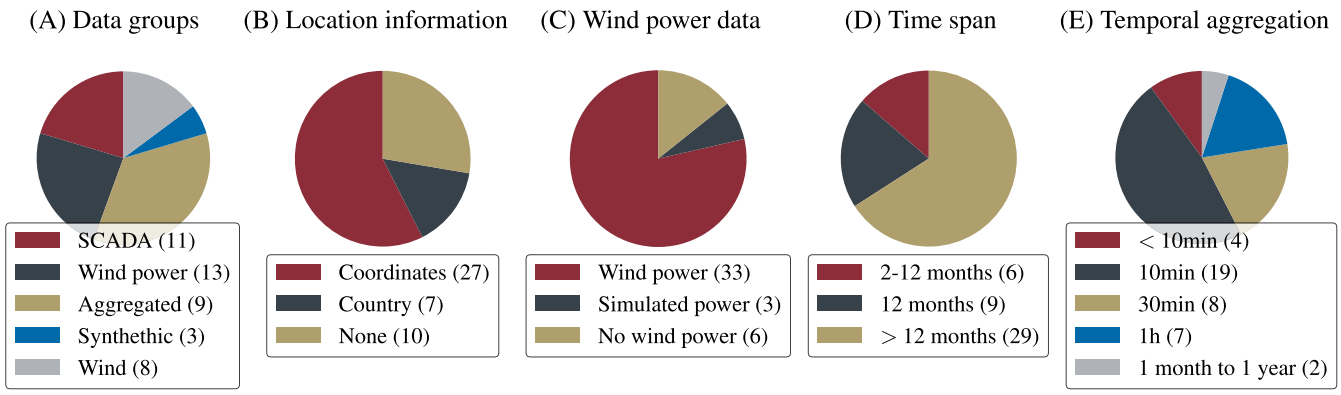

FIGURE 2 Descriptive statistics of the presented datasets. The numbers in the brackets display the absolute size of each group respectively. In (A)–(D), all of the five groups are contained, (E) does not take the wind datasets into account

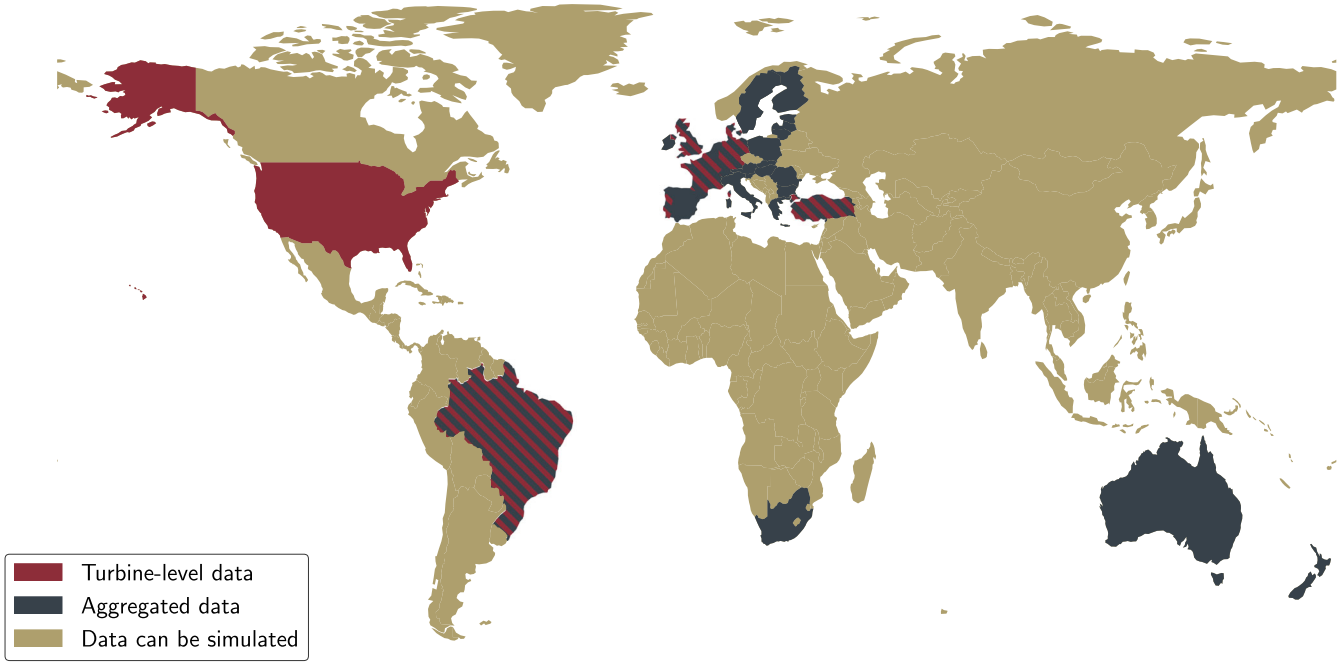

FIGURE 3 Country coverage of the datasets. Synthetic data can be simulated for all countries and locations given publicly available NWP data. For several countries turbine-level data or aggregated data are accessible. Countries where at least one dataset of aggregated and turbine-level data are accessible are colored in stripes

information regarding data policy, if applicable. Further, we provide the column **Additional information** which in general consists either of additional documentation, descriptions on how to add location data or variables that are included in the dataset. In the column **Operational data** additional information regarding operational data can be found. Additionally, if applicable, the number of turbines covered is given in the column **Turbines** and for the aggregated datasets the aggregation level is provided in the column **Aggregation level**. The column **Type** contains the type of wind data, which is either numerical weather prediction (NWP) model, reanalysis data or wind measurements.

2.3 | Wind power and turbine-level data

As stated above, we split the wind power data into three disjoint groups (see Figure 1): Two of them contain turbine-level data that cover different variables, while the third group contains wind power data aggregated on different spatial scales. In turbine-level data, weather variables such as wind direction are measured at hub height, the height on which wind power is generated. Having data at that height allows detecting height-specific turbulence patterns. Additionally, wind power forecasts can also rely on other weather variables such as temperature or humidity.[22] Bilal et al[23] show that including this additional weather information can increase the accuracy of the models. One reason behind this is icing of the

turbines, which was found to induce turbine power losses of up to 80%.[24] About 94% of turbines in Europe were found to be affected by icing.[25] Therefore, icing and its related weather variables such as precipitation and humidity can play a major role in the wind power forecast. Precipitation can also play an important role in areas where icing does not occur, mainly because rain cleans the turbine blades which increases power generation.[26] On the other hand, precipitation leads to turbine erosion on the leading edge of wind turbine blades which decreases wind power generation.[27] This knowledge can be beneficial in long-term forecasting. Summarized, taking into account weather variables besides wind speed and other wind characteristics can increase forecast accuracy. If these data are not provided but the turbine location is known, data from NWPs or close-by met masts can be included.

We further divide turbine-level data into two subgroups. The first group includes a large variety of supervisory control and data acquisition (SCADA) variables. SCADA data cover a set of environmental, operational, thermal and electrical measurements[28] usually recorded for maintenance reasons. We give an overview of existing datasets that include SCADA data in Table 2. The second data group covers SCADA-subset turbine-level data; an overview of the datasets is given in Table 3. The datasets in this SCADA-subset group only account for a limited amount of measured weather variables and do not contain any technical measurements. All of them contain wind speed and wind power data and most of them also contain information on the wind direction.

However, among the three groups of wind power data the third group of datasets, namely, aggregated datasets, are the ones that are most often publicly available. Aggregated data do not contain turbine-level measurements or turbine-specific power output but wind power data that is aggregated over different spatial regions, ranging from wind farms to whole countries. In contrast to turbine-level data, most of the datasets are of lower temporal resolution and contain hourly data. Nevertheless, the dataset in this category with the highest resolution is with measurements every 4 s the most finely resolved one in the whole collection. When the location of the wind farm is known, additional weather data from NWP models or close-by met masts can be taken into account. Several databases allow for a mapping of wind farms to their location. This mapping can for example be performed for the datasets provided by Elexon[29] and the transparency platform of ENTSO-E,[21] two of the largest aggregated datasets. Both are updated daily with a lag of approximately five days and wind power data can be found under "Actual Generation Output per Generation Unit." Table 4 gives more details of all aggregated datasets.

2.4 | Wind-based data

Wind power is generated by transforming the wind's kinetic energy into physical torque. As generated wind power is proportional to wind speed cubed,[30] the performance of an operating wind turbine is mainly determined by wind speed. Some research therefore focuses on wind speed forecasting in order to perform wind power forecasting; sometimes, this is also called indirect wind power forecasting.[31]

This dependence of wind power on wind speed is also exploited to generate meteorologically derived time series. These data are usually generated by transforming real or modeled wind and weather data into synthetic wind power data. By mapping wind intensity with a turbine-specific power curve to extracted wind power, data can be generated without taking any—potentially disclosed—wind power data of the turbines that are modeled into account. The wind datasets can be sub-divided into pure wind datasets and synthetic wind power datasets based on wind data. The datasets in Table 5 cover the former and contain either NWP models, reanalysis data or met mast measurements. Reanalysis data consists of measured, post-processed and interpolated data[32] while NWPs use mathematical models of the environment and its current state to predict future weather variables. Synthetic datasets that are based on wind data and can be derived from these can be found in Table 6. In general, all of the wind and wind power datasets are suitable for various wind power forecasting tasks, but not all of the datasets are suitable for every model. To find a suitable dataset for a specific model, the variable description in the dataset tables (Tables 2,3,4,5, and 6) should be considered.

2.5 | Providing the data

Having discussed the different types of (open-source) data used in wind power forecasting research, we now conclude with some suggestions for researchers who work with their data. The best way to proceed is to use data that can be published and re-used without limitations. The FAIR guiding principles[15] (see also Section 1) provide a good starting point for better reproducibility. Additionally, code, including scripts for data preprocessing, should be provided. However, due to constraints posed by the data owners, it is not always possible to publish the data. In this case, a comprehensive description of the dataset and its metadata, including farm size and location, can help put the research in context.

3 | EVALUATION OF THE DATASETS' PROPERTIES

Having introduced all datasets, we take a closer look at the properties of the datasets. We want to point out that dealing with live data for operational forecasts introduces additional difficulties that are not addressed in this work. With this section we aim to enable choosing a proper data

TABLE 2 Overview of the supervisory control and data acquisition (SCADA) datasets

Dataset	Turbines	Location	Coordinates	Time span	Interval	Operational data	Additional information	Origin and License
Beberibe[36]	32	Brazil	*Tabela* 20 of *anexo* B[37]	08.2013–07.2014	10 min	Not available	The Beberibe wind farm has an installed capacity of 25.6MW, with 32 Enercon E-48's installed at 75m a.g.l. It was used in Yoshiaki Sakagami's PhD thesis[37]	Supported by the Brazilian Electricity Regulatory Agency,[36] Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)
Pedra do Sal[36]	20	Brazil	*Tabela* 19 of *anexo* B[37]	08.2013–07.2014	10 min	Not available	The Pedra do Sal wind farm has an installed capacity of 18MW, with 20 Enercon E-44's installed at 55m a.g.l. It was used in Yoshiaki Sakagami's PhD thesis[37]	Supported by the Brazilian Electricity Regulatory Agency,[36] Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)
Penmanshiel[38]	14	United Kingdom	Can be found in the documentation[38]	01.01.2016–01.07.2021	10 min	Site substation/PMU meter data and site fiscal/grid meter data where available for the same period	The Penmanshiel wind farm has an installed capacity of 28.7MW, with 14 Senvion MM82's. Additional data is provided on Zenodo[38] including site substation/PMU meter data and site fiscal/grid meter data.	Cubico Sustainable Investments Ltd,[39] Attribution 4.0 International (CC BY 4.0)
Delabole[40]	10	Great Britain	Can be found in the documentation[41]	01.05.1993–30.04.1994	10 min	Not available	The Delabole wind farm has an installed capacity of 4MW, with 10 Windane 34's, data of 8 turbines is available. Additional information. Further documentation is available.[41]	DTU Database,[42] Attribution 4.0 International (CC BY 4.0)
Kelmarsh[43]	6	United Kingdom	Can be found in the documentation[43]	01.01.2016–01.07.2021	10 min	Site substation/PMU meter data and site fiscal/grid meter data where available for the same period	The Kelmarsh wind farm has an installed capacity of 12.3MW, with 6 Senvion MM92's. Additional data is provided on Zenodo[43] including site substation/PMU meter data and site fiscal/grid meter data.	Cubico Sustainable Investments Ltd,[39] Attribution 4.0 International (CC BY 4.0)
La Haute Borne[44]	4	France	See static information[44]	01.01.2013–13.01.2018	10 min	Not available	The La Haute Borne wind farm has an installed capacity of 8.2MW, with 4 Senvion MM82's.	Engie Renewables,[45] Open License v2.0 (Etalab)

(Continues)

TABLE 2 (Continued)

Dataset	Turbines	Location	Coordinates	Time span	Interval	Operational data	Additional information	Origin and License
Tjæreborg[46]	1	Denmark	55.448233, 8.593803[47]	20.01.1988–18.01.1993	10 min	Several information regarding availability and operating hours and technical details can be found in the reports[47]	The turbine in Tjæreborg is a production of Elsam Projekt A/S[48] has an installed capacity of 2MW. Additional data of a close-by met mast is also available.[49]	DTU Database,[42] Attribution 4.0 International (CC BY 4.0)
Eolos Wind Research Station[5051]	1	United States	44.73, −93.05[52]	01.01.2017–31.12.2017	10 min	The authors report "significant turbine curtailment" and other limitations of the dataset.[50] Curtailment data is not provided.	The Clipper Liberty turbine with an associated 130 m met mast has an installed capacity of 2.5 MW. The dataset consists of incomplete SCADA data (74 different SCADA and other variables). See description on the same page for further details.[50]	University of Minnesota (Brian Davison),[51] CC0 1.0 Universal (CC0 1.0) Public Domain Dedication
EDP[53]	4	Portugal	Unknown	01.01.2016–31.12.2017	10 min	Historical failure log book and wind turbine logs are available.[53]	Data of 4 out of 16 turbines is available, the wind park has a total installed capacity of 32MW.[54] Registration is necessary for data access. Description of the variables and additional data of a close-by met mast can be found on the same page.	EDP Inovação,[55] Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)
Kaggle 1[56]	1	Unknown	Unknown	31.12.2017–31.03.2020	10 min	Not available		Unknown, CC0 1.0 Universal (CC0 1.0) Public Domain Dedication
Kaggle 2[57]	1	Unknown	Unknown	01.05.2014–09.04.2015	10 min	Includes fault and status data.		Unknown, Unknown

Note: The datasets contain SCADA data and the separation line divides the datasets with known location from the datasets without known location. A description of the variables can be found in the beginning of this section. Additionally, the number of turbines that the dataset covers is provided in **Turbines**.

TABLE 3 Overview of SCADA-subset turbine-level wind power datasets

Dataset	Turbines	Location	Coordinates	Time span	Interval	Operational data	Additional information	Origin and License
Maelstrom[58]	45	Germany	Accessible after download	6 months	1h	Error number of the turbines, feed-in management inferred from turbine logs and from the power provider, status of the turbine[59]	Average production, minimal production, maximal production, mean wind speed, minimal wind speed, maximal wind speed, mean rotor speed, minimal rotor speed, maximal rotor speed, [59]	Notus Energy,[58] Apache License Version 2.0
Kaggle Turkey[60]	1	Turkey	40.585469158, 28.990284697[60]	01.01.2018–13.12.2018	10 min	Theoretical power curve	Active power, wind speed and wind direction	Unknown, unknown
NM92[61]	1	Denmark	57.039, 10.075722	19.11.2005-20.01.2006	10 min, raw time series	Availability of wind turbine, operational mode wind turbine, generator status (on/off)	The NM92 turbine has an installed capacity of 2.75 mW and a hub height of 70 m. Mast, turbine power and load measurements (25 Hz) together with 10-min statistics of these measurements are provided. Documentation is available.[62]	DTU Database,[42] Attribution 4.0 International (CC BY 4.0)
Nordtank[63]	1	Denmark	55.684436, 12.096689	21.10.2004–20.04.2006	10 min, raw measurements	Wind turbine operational mode, Thies inductive rain detector, tip activated, disc brake activated, generator grid connected	The Nordtank turbine has an installed capacity of 0.5mW and a hub height of 36 m. Mast, turbine power and load measurements (appr. 35 Hz) including 10-min statistics of these measurements are provided. The turbine turbine measurement laboratory has been	DTU Database,[42] Attribution 4.0 International (CC BY 4.0)

(Continues)

TABLE 3 (Continued)

Dataset	Turbines	Location	Coordinates	Time span	Interval	Operational data	Additional information	Origin and License
							developed by the DTU. Documentation is available.[64]	
Denmark Data[65]	>5000	Denmark	See "Additional information"	2002–2020	1 month	Not available	The dataset is called *Monthly data 2002-2020* and contains yearly production data for all wind turbines > 6kW. In *Data on operating and decommissioned wind turbines* location information can be found.[65]	Danish Energy Agency[65]
Wind Spatio-Temporal Dataset2[66]	200	Unknown	Relative position of the turbines	01.09.2010–31.08.2011	1 h	Not available	Hourly wind speeds, wind speed and wind direction at three met masts on the same wind farm[67]	Texas University,[68] Attribution 4.0 International (CC BY 4.0)
Wind Spatio-Temporal Dataset1[69]	120	Unknown	Relative position of the turbines	01.01.2009–31.12.2010	1 h	Not available	Average wind speed, hourly standard deviation of wind speed[67]	Texas University,[68] Attribution 4.0 International (CC BY 4.0)
Offshore Wind Farm[70]	10	Offshore, unknown	Relative position of the turbines	01.07.2007–31.08.2007	1 h	Not available	Wind speed, wind direction, air density, humidity, turbulence intensity, above-hub height wind shear, below-hub height wind shear[67]	Texas University,[68] Attribution 4.0 International (CC BY 4.0)
Inland Wind Farm Dataset1[71]	4 + 2 met masts	Offshore, unknown	Relative position of the turbines	different time horizons (1year)	10 min	Not available	Wind speed, wind direction, air density, below-hub height wind shear, turbulence intensity[67]	Texas University,[68] Attribution 4.0 International (CC BY 4.0)

TABLE 3 (Continued)

Dataset	Turbines	Location	Coordinates	Time span	Interval	Operational data	Additional information	Origin and License
Inland Wind Farm Dataset2[72]	4	Unknown	Relative position of the turbines	2008–2011	10 min	Not available	Wind speed, wind direction, air density, below-hub height wind shear, turbulence intensity[67]	Texas University,[68] Attribution 4.0 International (CC BY 4.0)
Offshore Wind Farm Dataset1[71]	2 + 1 met mast	Unknown	Relative position of the turbines	01.01.2009–31.12.2009	10 min	Not available	Air density, wind shear, turbulence intensity, humidity	Texas University,[68] Attribution 4.0 International (CC BY 4.0)
Offshore Wind Farm Dataset2[72]	2	Unknown	Relative position of the turbines	2007–2010	10 min	Not available	Air density, wind shear, turbulence intensity, humidity	Texas University,[68] Attribution 4.0 International (CC BY 4.0)
Wind Time Series Dataset[73]	1	Unknown	Unknown	07.10.2014–06.10.2015	10 min and 1 h	Not available	Wind speed	Texas University,[68] Attribution 4.0 International (CC BY 4.0)

Note: The datasets contain wind power, wind speed and other turbine-level measurements but no full SCADA data. These datasets stem originally from a SCADA system but only a subset of the usually measured SCADA data are publicly available. The separation line divides the datasets with known location from the datasets without known location. A description of the variables can be found in the beginning of this Section. Additionally, the number of turbines is provided in **Turbines**.

TABLE 4 Overview of the aggregated wind power datasets

Dataset	Location	Aggregation Level	Time span	Interval	Operational data	Additional information	Origin
ENTSO-E[21]	European Member States	Farm level and other	Since 2011	30 min to 1 h	For a detailed description of accessible data, we refer to the official documentation.[74] Among others, it includes information regarding congestion management such as redispatching.	Wind power data are stored under "Actual Generation Output per Generation Unit". For python the package entsoe-py can be used.[75] Data are updated regularly and a data subset covering the time period from 21.12.2014 to 11.04.2021 is provided by De Felice et al.[76] A second data subset containing hourly capacity factors for wind onshore from 1982 to 2019 at national and subnational (>140 zones) level is also provided by De Felice et al.[77] Location information is accessible, for example, by the Joint Research Centre.[78]	European Union,[79] mostly Attribution 4.0 International License (CC-BY 4.0). For further details we refer to the "List of data available for free re-use".[80]
Elexon[29]	Great Britain	Farm level and other	Since 2011	30 min	Among others, Performance Assurance Reporting and Monitoring System (PARMS) data, imbalance information and the Transmission Loss Factor (TLF)[81]	B1610 provides wind power generation data. Several software can help to work with the data, for example, for python the ElexonDataPortal library[82] can be used. Data are updated regularly with a lag of approximately 5 days. Energy identification codes (EICs) allow for location identification.	Elexon,[83] see license provided on the website[84]
New Zealand Data[85]	New Zealand	Farm level	Since 1997	30 min	Not available	Generation by plant data is estimated by mapping metered injections into the grid to the generating plant at those injection points. Location information is in the network supply points table,[86] POC_Code[85] and POC code[86] can then be mapped. The coordinates can be transformed to longitude/latitude (current is 'epsg:3857') e.g. with pyproj.[87]	Electricity Market Information New Zealand,[85] Attribution (CC-BY) 4.0 international license
Brazilian Data[88]	Brazil	Federal state	Since 2007	1 h to 1 year	Not available	*Tipo de Usina* has to be set to *Eólica*.	Operador Nacional Do Sistema Eletrico (Brazil),[89] license-free[90]
South African Data[91]	South Africa	Province	Since 2015	1 h	Not available	A manual on how to download the data is also provided.[92]	Department of Energy South Africa,[93] raw data is available for own analytical use

TABLE 4 (Continued)

Dataset	Location	Aggregation Level	Time span	Interval	Operational data	Additional information	Origin
Gefcom2014[94,95]	Australia	Farm level	01.01.2012–30.09.2012	1 h	Not available	In the dataset wind speeds as u10, u100, v10, v100 are included. The dataset covers 10 wind farms in Australia, their locations are disclosed.	Tao Hong (University of North Carolina),[95] cite paper to acknowledge the source
AEMO[96]	South-east Australia	Farm level	01.01.2012–31.12.2013	5 min	Not available	The dataset contains wind power data from 22 wind farms with known location. Abbreviations of the wind farm's names can be found in the csv-file with which locations can be mapped. Note that this is just a data subset, more AEMO data can be downloaded.[97]	AEMO and University of Strathclyde,[98] Attribution 4.0 International (CC BY 4.0)
4 Seconds Time Series[99]	Australia	Farm level	Starting on 01.08.2019	4 s	Not available	The dataset consists of one wind power time series from a single farm containing 7,397,147 values.	AEMO and Monash University,[100] Attribution 4.0 International (CC BY 4.0)
Wind Farm Data with Missing Values[101]	Australia	Farm level	01.08.2019–31.07.2020	1 min	Not available	The dataset contains 339 power time series of Australian wind farms with missing values (for some series more than seven consecutive days). Additionally, a second dataset where missing values are replaced by zeros is provided.[102]	AEMO and Monash University,[100] Attribution 4.0 International (CC BY 4.0)
EEM2020[103]	Sweden	4 Swedish regions	01.01.2000–31.12.2001	1 h	Not available	The datasets consists of aggregated data of 4 regions.	University of Strathclyde,[104] Attribution 4.0 International (CC BY 4.0)

Note: The datasets contain spatially aggregated data. A description of the variables can be found in the beginning of this Section. Additionally, the aggregation level is provided in **Aggregation level**.

TABLE 5 Overview of wind datasets

Dataset / Website	Type	Location	Time span	Interval	Additional information	Origin, License
MERRA[105]	Reanalysis	Earth	1979–02.2016	1 h to 6 h	Several datasets can be accessed, an additional description is provided.[106] MERRA 2 provides data beginning in 1980 and replaces MERRA since 2016.[107]	NASA,[108] full and open sharing of all data with research and applications communities, private industry, academia, and the general public[109]
ERA5[110]	Reanalysis	Earth	1979–ongoing	1 h	ERA 5 covers the Earth on a 30-km grid from the surface up to a height of 80 km.[111]	ECMWF,[111] Attribution 4.0 International (CC BY 4.0)
Public ECMWF NWP data[112]	NWP	Earth	-	6 h and 12 h	Two models based on HRES (single high resolution) and ENS (ensemble of forecasts) are publicly available. Both datasets contain mean sea level pressure, geopotential height, temperature and either u and v components of the wind or wind speed on a 0.5° by 0.5°. Notice that higher resolution models are available (see next row) for ECMWF member states. Other open-source datasets from ECMWF include data from two projects named TIGGE and S2S. TIGGE[113] is an ensemble model starting from October 2006 and the goal of S2S[114] is to improve forecast skill and understanding on the subseasonal to seasonal timescale.	ECMWF,[112] Attribution 4.0 International (CC BY 4.0)
ECMWF NWP data for member states[112]	NWP	Earth	-	hourly to 6-hourly	For ECMWF member states[115] the HRES model[116] is available with a resolution of 0.1° x 0.1° lat/long grid and the ensemble model[117] with a resolution of 0.2° x 0.2° lat/long grid.	ECMWF,[112] Attribution 4.0 International (CC BY 4.0)
NOAA[118]	NWP	Earth	-	hourly to 3-hourly (GFS) 6-hourly (GEFS)	An ensemble model GEFS[119] and a deterministic model GFS[120] are provided. The ensemble has a resolution of 1.0° and the horizontal resolution of the deterministic model is 13 km.	National Oceanic and Atmospheric Administration,[118] Attribution 4.0 International (CC BY 4.0)
Orsted[121]	Offshore lidar wind measurements	Baltic sea, Denmark	2012–2017	10 min	The three datasets contain data from met masts close to offshore wind farms. They cover different time spans between 2012 and 2017, the largest covers around 2.5 years. Documentation of two of the datasets can be downloaded after registration. Further information on Fino 2 and the other Fino research platforms in the Baltic Sea is also accessible.[122]	Orsted,[121] see terms and conditions on the website
Energydata.info[123]	Met mast measurements	Various	-	-	On the page many different wind datasets from different countries all around the globe are provided. Some of them are updated regularly.	World Bank Group,[124] dataset specific but mostly Attribution 4.0 International (CC BY 4.0)
Tall Tower Set[125]	High met mast measurements	Various	-	-	Tower data of tall towers around the globe.	Supercomputing Center Barcelona,[126] Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)

Note: The separation line divides the global weather datasets from wind measurement datasets and databases. The mentioned separation line is not visible in this preview?. The Orsted dataset is well known in the wind power research community because the signals are measured close to large offshore wind farms. The last two entries reference databases containing met mast data.

TABLE 6 Overview of synthetic datasets

Dataset	Location	Time span	Interval	Information	Origin, License
WIND Toolkit[127]	United States (over 126,000 locations)	2007–2013	5 min	The WIND toolkit is one of the largest synthetic grid integration and wind power datasets publicly available. Several data subsets can be downloaded. The WIND Toolkit is widely used in wind power research (e.g., literature[128,129]).	National Renewable Energy Laboratory,[130] Attribution 3.0 United States License (CC BY 3.0 US)[131]
Renewables Ninja[132]	EU-28, Norway, Switzerland	1980–2016	1 h	Additionally, hourly power output from wind and solar power plants can also be simulated.[133]	Iain Staffell and Stefan Pfenninger,[134] Attribution 4.0 International (CC BY 4.0)
EMHIRES[135]	EU-28, Norway, Switzerland, non EU countries from the Western Balkans	01.01.1986–31.12.2015	1 h	Data are aggregated by country (onshore and offshore), power market bidding zone and by the European Nomenclature of territorial units for statistics (NUTS) as defined by EUROSTAT.[136] A comparison with Renewables Ninja was performed by Moraes et al[137] and Gonzalez Aparicio et al.[138]	Gonzalez Aparicio et al.,[138] European Commission Reuse[139]

Note: The datasets contain meteorologically derived wind power data. A description of the variables can be found in the beginning of this Section. Renewables Ninja can additionally be used to simulate wind farms and create further synthetic data.

set for different wind power forecasting tasks. As the different data groups come with their group specific properties, we first address the tasks for which the datasets are often used and then discuss the shortcomings and advantages of each group. The discussion focuses on the role of specific variables as well as spatial and temporal granularity and coverage of the datasets. We do not evaluate the properties of the different wind measurements in this section but refer to Section 4 for this. The goal of wind power forecasting is mainly threefold. It consists of wind farm site selection (1), efficiently harnessing the wind (2)—mainly by controlling the turbine—and efficiently integrating that generated wind power into the grid (3). Additionally, wind power forecasting is performed hours to days ahead for power system management and energy trading.[33] Most other goals, such as maintenance planning, can be derived from results that tackle these three primary issues. To control the turbine and harness the wind efficiently (2), it is usually assumed that finely-resolved data on a second or even millisecond scale are needed.[2] Therefore, even though eleven of the datasets contain control parameters, their temporal resolution of 10 min is not sufficient for efficient turbine control and we will not discuss turbine control in this review. Consequently, the tasks we cover mainly aim to improve site selection (1) and wind power grid integration (3). The goal of the former is to optimize turbine or wind farm site selection under certain constraints. These constraints can be diverse; among them are the average wind speed, geological and geographical properties of the surrounding area, and political guidelines that, for example, restrict the closeness of turbines to cities.[34] Furthermore, while site selection has to rely on pure wind and weather data only, wind power grid integration models usually also consider wind power data.

3.1 | Wind power data

We will now first elaborate on the direct use of wind power data to perform wind power forecasting. The associated goal of wind power forecasting is usually efficient wind power grid integration. To achieve this goal, different sub-tasks can be defined. Among them are wind power time series forecasting on different temporal and spatial scales, ramp forecasting and variability forecasting. The main difference between these forecasting tasks is not always their underlying model or data but, in many cases, the evaluation metric used to assess the forecast quality. Therefore, the tasks mentioned here are neither disjoint nor do they necessarily require different models. An example of this can be found in Bianco et al,[35] where redefining the error metric allows using time series forecasts to investigate variability of the wind or ramp events.

The underlying main task of wind power forecasting is usually wind power time series forecasting, a regression task that aims to predict wind power generation at future time points given historical data. A common division of wind power time series forecasting is given by the World Meteorological Organization, ranging from very short term (in the range of a few minutes to hours) to long-term (in the range of 1 month to years). Additionally, wind power forecasting models can also be classified by their prediction methodology (usually physical, statistical or hybrid).[3] While very short-term forecast regression is usually performed with statistical models using historical wind (power) data, more advanced physical models do in general rely on exogenous data from NWP models. The physical properties that are preserved by these NWP models are in general needed for high quality long-term forecasts. Wind power forecasts on a time horizon of hours to days ahead can also be used for energy trading and power system operations. It has been shown that leveraging the forecast on wind farm level rather than turbine level can increase forecast skills.[140] Another subtask of wind power forecasting is wind power ramp forecasting which aims to predict large and sharp variations in the wind and its associated wind power. While there is no unique definition of a wind ramp, such a ramp can be characterized by the direction and magnitude of the power variation and its corresponding duration.[141] Additionally, the rate at which such large variations occur and the time point at which they occur can play a role when evaluating wind ramps. Key characteristics of wind ramps are that they can lead to a power output that is far below or above the usual one which comes either with decreased power generation or potential damage of the turbine.[142] Both of these outcomes can have a negative impact on the energy supply. The third task we mention here is wind variability forecasting. In contrast to wind power ramps, wind power variability refers to large amplitude, periodic changes in wind speed.[2] However, there is no clear definition of wind variability either. Davy et al[143] introduce a variability index that "is defined as the standard deviation of a band-limited signal in a moving window." In general, wind fluctuations that are part of wind variability can, for example, exhibit climatic patterns with regard to the season or atmospheric processes such as cloud coverage. Having introduced these tasks, namely, wind power time series forecasting, ramp forecasting, and variability forecasting, we now elaborate the usefulness of the three groups of wind power data for these tasks. However, there is no guarantee for completeness of the tasks discussed here. In the subsequent part, we discuss the spatial and temporal resolution of the datasets, the informative value of location information, and the transferability of the results.

The spatial and temporal resolution of the individual datasets plays an important role in the usefulness and expressiveness of the data. SCADA data are the best with respect to spatial and temporal granularity; they consist of measurements of individual wind turbines and usually preserve a temporal resolution of 10 min. The SCADA-subset turbine-level datasets have the same spatial resolution but often a coarser temporal resolution. While the spatial resolution within a limited region is high for the turbine level datasets, their spatial coverage is currently low due to data regulations. Therefore, if one aims to predict or analyze the wind power of more than one wind farm, other datasets have to be considered. Among these are the aggregated datasets presented in Table 4, which cover large spatial areas. Additionally to spatial and temporal coverage, the time horizon that a dataset covers can also be crucial. For example, seasonal patterns can usually only be detected in datasets that cover every season at least twice, that is, have a length of at least 2 years.

For most countries, wind turbine locations are accessible (see, e.g., the global power plant database[144]) and the aggregated datasets can be combined with these wind farm locations, which can give a broader insight into the wind power network. For three of the datasets (see ENTSO-E, Elexon, and New Zealand Data in Table 4), location mapping is straightforward, and we include resources on how to do so. However, this mapping can usually not be performed for most turbine-level SCADA-subset datasets found as they lack location information. Furthermore, incorporating control specific parameters is generally not possible afterwards. Nevertheless, some forecasting models explicitly take variables such as nacelle orientation, yaw error, or blade pitch angle into account (e.g., Lin et al.[22]) and can therefore only be explored, validated, and tested using turbine-level data that include these variables. In general, the more finely resolved datasets contain more variables, and Figure 4 shows that, on average, the SCADA datasets contain the most variables while weather data on turbine-level (except for wind data) are never included in the aggregated datasets. However, given location information, NWP data could be included.

Another crucial point of wind power forecasting models is their ability to be generalized and transferred to other data. In general, it is hard to elaborate how expressive an individual forecast is. Turbine–turbine interactions and environmental conditions give each turbine and wind farm a unique set of complex physical properties that make evaluating the expressiveness of one turbine's forecast difficult. However, SCADA data for many turbines in larger areas are currently not publicly available, to the best of our knowledge. Therefore, the informative value of SCADA data mostly remains limited to a specific range of conditions. In contrast, aggregated datasets cover larger regions, and when forecasting wind power generation of multiple turbines and parks together, local wind effects can become less relevant. Additionally, Holttinen et al[145] show that forecast errors drop when predicting aggregated wind power production instead of individual sites. Nevertheless, training models on one single wind power dataset always limits transferability.

To summarize, the application area of the different wind power datasets is mainly limited by the spatial and temporal aggregation level and coverage. Additionally, non-disclosed locations of the turbines play a crucial role as they limit the possibility of including additional weather or terrain data that are often used in more advanced models. However, other models that do not take this information into account are not affected by missing location data. Nevertheless, some variables such as control variables can not be included afterwards and make some datasets unsuitable for testing and validating specific models—independent of the data group to which the datasets belong. Furthermore, the possibility of transferring wind power forecasting models validated on one dataset to another is generally restricted, especially in the case of single turbines and wind farms where local turbulences play a significant role and highly influence the results.[8] However, taking several sites into account can reduce this effect.[145]

3.2 | Wind-based data

Especially for larger regions, fine-grained wind power data are not accessible. Whenever this is the case, we can exploit the underlying physical properties of wind power generation. For example, as generated wind power is proportional to wind speed cubed,[30] wind speed mainly determines the performance of an operating wind turbine. However, the main advantage of using wind data instead of wind power data is their high spatial coverage. Therefore, some research focuses on wind speed forecasting in order to perform wind power forecasting—sometimes, this is called indirect wind power forecasting.[31] One example of this is presented by Demolli et al[146] who aim to predict the wind speed and map it with a turbine-specific power curve to its corresponding wind power. Using such a turbine-specific power curve assumes that this wind speed to wind power mapping is deterministic, that is, given a wind speed and a power curve the corresponding wind power generation can be computed.

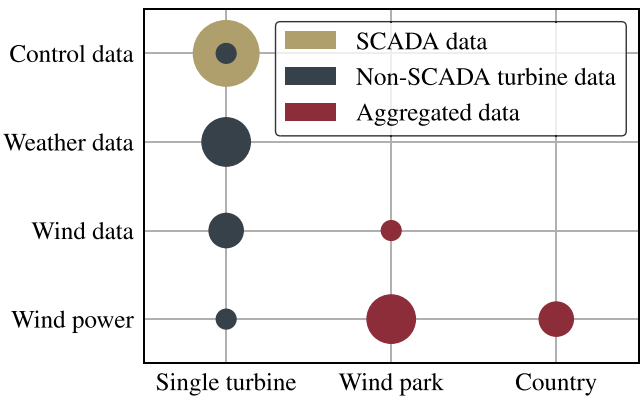

FIGURE 4 Overview of the spatial aggregation and information contained in the different datasets. The diameter of each circle scales linearly with the number of datasets associated with the respective category. The y-axis is ordered with increasing information content, for example, datasets that contain control data always contain wind power, wind and weather data as well. While none of the aggregated datasets contains weather data, six of them contain location information that allows to incorporate weather variables for example from NWPs

However, this assumption usually does not hold (see Figure 5). Therefore, other approaches model the wind power curve empirically based on wind data and its corresponding known wind power output to account for this stochasticity. Still, the choice of a power curve model remains difficult; Wang et al[147] study several power curve models and show that none of the models can outperform all of the other models.

While the expressiveness of wind speed forecasting without additional power data is usually limited, we can perform wind assessment for turbine site selection and efficient planing of wind turbine and wind farm constructions. This is usually done to assess whether the wind speed at a possible future site is high enough to generate a rewarding amount of energy, optimize the turbine locations given the area and location of a planned wind farm, or assess the maximum wind power capacity of a region. For completeness, it has to be mentioned that the maximum and average predicted capacity are not the only relevant considerations for site selection. Additionally to meteorological components, environmental, economic, and societal factors can further restrict site selection.[34]

Instead of using the wind data directly, we can also simulate wind power data based on wind measurements. This simulation allows incorporating the impact of orography into synthetic datasets and can reduce biases.[134] Overall, wind data thus have the advantage of covering the whole world, allowing us to use it for indirect wind power forecasting, site selection and synthetic data generation.

4 | SYSTEMATIC ERRORS IN WIND DATA

There are mainly two problems with the different types of wind data, namely, data availability and bias in the data. The first, data availability, mainly affects wind power data (see Figure 1) as pure wind data are open-source in many cases. The reason for this is that datasets that include wind power and wind measurements on turbine-level usually belong to wind power companies that do not want to disclose sensitive information about their operations. Without a shift in industry mindset, this lack of available data makes the collection of these datasets in Section 2.3 even more important. In contrast, the second problem, bias in the data, affects all types of wind data and can be, at least in parts, addressed in the wind power forecasting models themselves. The different wind data measurements include a range of known systematic measurement and modeling errors. As a result, each form of wind data comes with its specific advantages and shortcomings. In the following, we will address these known systematic errors for each form of wind data, namely, wind measurements from met masts and turbines as well as computed wind data from NWPs. In order to evaluate the different data types, we first compare wind measurements on turbine-level to wind measurements of met masts and proceed with a discussion of NWP models and reanalysis data.

4.1 | Wind measurements

Wind measurements for wind power forecasting are usually performed close to the turbine's nacelle or by close-by independent met masts. The main advantage of nacelle anemometer data is that the wind is measured close to where wind power is generated. However, the moving turbine blades influence these measurements and these measurements are therefore considered inaccurate.[148] Still, turbine-level data is often used to forecast wind power and more recent studies by Cutler et al[149] show an improvement in power curve modeling when using turbine-level wind speed as compared to met mast measurements. Instead of dealing with these known perturbations induced by the turbine, forecasting with other types of wind data, for example, in the form of close-by met mast LiDAR measurements, can be performed. Some renewable energy companies also support and motivate this approach.[121] Nevertheless, these wind speeds measured by close-by met masts may not be representative either, especially if the terrain conditions are complex or the met masts are too far away from the turbine of interest.[149] Another shortcoming of many of these met mast measurements is the installation height of the average met mast. With a standard height of 10 m above the earth surface, they do

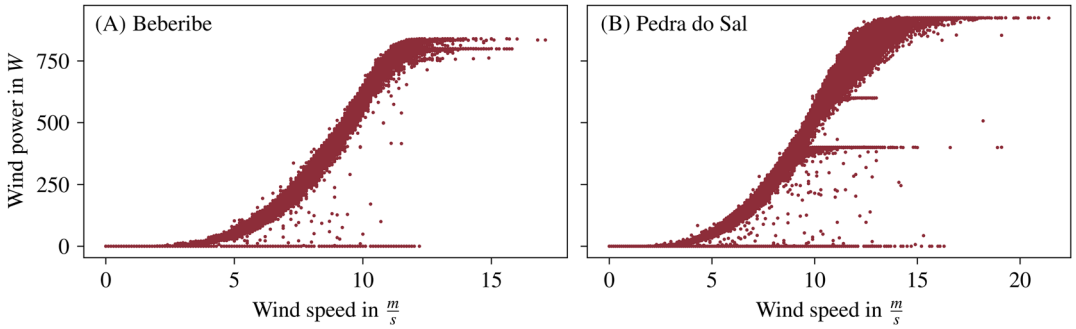

FIGURE 5 Scatterplot that shows wind power on the y-axis and wind speed on the axis. The plots show data from the first turbine of the dataset in Beberibe and Pedra do Sal[36] respectively. It can be seen that the mapping from wind speed to wind power is not deterministic

not cover the average turbine height of 50–150 m.[126] Therefore, Draxl et al[150] argue that validation against 10-m wind speeds is not sufficient. To address this height-issue, Ramon et al[126] present existing non-standard meteorological observations from 222 taller met masts. Nevertheless, the closeness to turbines and associated wind power generation are usually unknown, and their usefulness therefore remains limited.

To summarize, wind measurements can either be taken close to the turbine, where the turbine influences the measurements, or further away from the turbine, where they might not preserve the local wind structure. However, knowing about the challenges these different types of wind data pose allows us to consider them during the development and validation of a forecasting model.

4.2 | Weather models

As the number of met masts and monitoring stations is finite, there are no met mast wind measurements for every possible site. Therefore, globally, the density of measurements differs, and weather models covering the whole earth interpolate and model wind and weather data to fill gaps in the measurements. This section will discuss the uncertainties of these weather models, their expressiveness, and some of their known biases. In wind power forecasting numerical weather prediction (NWP) models are commonly used.[151–153] They describe atmospheric processes with input from weather observations and atmospheric and oceanic simulations. However, due to their nonlinearity, complexity and the gaps in measurements, they are solved with numerical approximations.[154] These numerical approximations and the interpolation of data points introduce uncertainty into the NWPs wind predictions. In fact, the NWPs accuracy is generally considered to be the most influential factor on the accuracy and uncertainty of the wind power forecast.[155] Nonetheless, Holttinen et al[145] show that the errors of the meteorological forecast model drop when data of many sites are being taken into account. Weather models are, in contrast to wind power measurements, available in large amounts covering all countries (see Section 2.4). However, the underlying meteorologic measurements are of different quality for different regions and cannot account for turbine-specific turbulences. Furthermore, they usually contain interpolated data and measurements below hub-height. Therefore, reanalysis or NWP datasets induce biases when they are used to simulate wind power data.[134] Reducing these biases is the goal of many forecasting algorithms that rely on weather models.[156] Additionally, planetary boundary layer parameterization aims to reduce these errors.[150]

Given the two different wind data types, we can decide which data to prefer by comparing the wind power curves that are associated with either wind measurement or modeled wind speed. The more scattered these curves map wind speed to wind power, the harder it gets to forecast wind power given wind speed. The wind power curve of the wind speed that is empirically determined using NWP data is more scattered than the power curve that describes the relationship between wind power output and wind speed measured at a reference wind mast.[157] This difference in variance indicates that measured wind speed is more accurate compared with wind speed data from weather models. Nevertheless, data availability crucially limits the use of wind measurements, and therefore, none of the wind data types can and should be preferred for each and every application and research question. When working with wind data or simulated wind power data, it should be kept in mind that the mapping from wind speed to wind power is nonlinear and stochastic. Wind speed forecasting is therefore just a subproblem of wind power forecasting, not an equivalent.

5 | DATA QUALITY

Data quality is crucial whenever data are involved in research or industry. However, the definition of data quality and its demands depend on the research question, making generalization difficult. Nevertheless, we want to address some high-level points of interest when using wind (power) datasets. Among these points are data outliers, the relevance of the datasets listed to current wind power forecasting research, the data's origin, and its associated credibility. The latter is closely related to the documentation and metadata that comes with the dataset of interest; therefore, we will elaborate on its importance.

It is generally assumed that fitting models to "bad quality or abnormal data"[158] results in biases and inaccuracies. However, incomplete and noisy data usually represent the data measured in the real world better than consistent and complete datasets, although we usually attribute a higher quality to the latter. The datasets messiness evokes a separate research subfield, data pre-processing, and various pre-processing techniques are used, for example, in order to eliminate outliers[159] or classify different weather types.[160] None of the datasets that are presented in Table 2 is described as being pre-processed. We, therefore, assume that the datasets contain non-processed data. However, several datasets do not provide raw measurements but averaged data over the measurement interval. An example is the Eolos wind research station data; the measurements are averaged over 10 min rather than recorded every 10 min. For other datasets, for example, the Wind Spatio-Temporal Dataset1 in Table 3, it is known that missing data are imputed. Pre-processing and outlier detection are, in general, choices that the researcher makes. However, this choice is often restricted by limited data availability, with data needing to be used as provided due to the policy of the data owners.

Having addressed pre-processing and outlier detection, we will now elaborate on the shortcomings and relevance of the wind power datasets described. While wind data from met masts and weather models are updated frequently, the wind power datasets are usually historical records.

Furthermore, most of the datasets in Table 2 are closed datasets; that is, it is not planned to update them. Because of this lack of actuality and because some datasets are already several years old—for example, Tjæreborg provides data that is already over 30 years old—their usefulness is limited. One reason for this limitation are the developmentsof wind measuring instruments[161] and the developments in NWP models.[162] Both the former and the latter can be traced back to technological progress. However, while the technological progress in wind measurements stems from the transition from cup anemometers to sonic anemometers and remote sensing by Doppler techniques,[161] the advances in NWP techniques can be mainly traced back to higher computational power and the "steady accumulation of scientific knowledge."[162] This progress makes the models and data better but, on the flip side, introduces inconsistencies within data that are measured or modeled over many years. Another point of importance that makes the older datasets less transferable to current wind power research is the trend of increasing capacities, hub heights, and turbine sizes.[163,164] However, increasing hub height is a minor issue in the datasets presented here, that is, the turbine in Tjæreborg has a hub height of 60 m and the turbine in Penmanshiel of 59 m. With an approximate hub height of 130 m,[165] the Eolos Wind Research Station is the highest of the turbines presented here. Nevertheless, the trend of increasing capacity is partially present in the datasets presented; for example, one shortcoming of the older datasets is that they only cover small farms or single turbines. Therefore, this collection does not cover large wind farms—the usefulness of forecast models validated on the datasets of this collection is therefore not known for large (off-shore) wind farms such as Hornsea 2[166] with an installed capacity of over 1.3 GW and 165 turbines. This lack of knowledge is critical as it is generally assumed that off-shore wind power will play an even more critical role in the future as land area is limited and wind speeds are on average higher over the sea.[167]

Additionally to the issues with measured data, one main shortcoming of many datasets is their lack of comprehensive documentation, metadata, and additional operational data. While several of the datasets from ECMWF and SCADA datasets that have already been used in wind power forecasting research come with helpful documentation and metadata, the datasets provided by wind energy companies, such as the La Haute Borne dataset, are less comprehensive. Furthermore, for most of the datasets, no information on turbine downs or wind power curtailment is provided. This lack of information makes the mapping of wind speed to wind power noisy because unknown confounding factors can influence this mapping. Current research often deals with these types of errors by eliminating values that are *too far away* from the theoretic deterministic mapping of wind speed and wind power; this is usually called wind power curve cleaning.[168] However, these methods cannot be validated as the underlying truth is unknown and measurement noise and curtailment are not always distinguishable. Another crucial point for the credibility of the data is the data's origin and the licenses associated with it. Therefore, these should also be considered part of data quality and reliability. It has to be mentioned that the Kaggle datasets[56,57,60] except for the dataset of GEFCom2012,[94] lack a good documentation of the variables and the datasets' origins are unknown. For the data provided by the Australian Energy Market Operator,[97] the data provided by the South African Energy Department[91] and the data provided by the Danish Energy Agency,[65] no information regarding data licenses was available online.

6 | DISCUSSION

Having presented the open-source datasets and the shortcomings and advantages of the different measurements, we now discuss their use and explain how we selected the different datasets. Open-source data and code play an important role in research. While many researchers would like to publish the data they use—we conclude this after various discussions with colleagues—most of the currently used wind power data is subject to confidentiality agreements and therefore non-accessible for the public. Our main motivation behind this work is therefore to compile a large collection of open-source wind power datasets that helps researchers in the field to use a suitable, non-confidential dataset. Figure 2 shows that the important characteristics such as known location, time horizons of at least 1 year and fine-grained 10-min data are covered by several datasets. Nevertheless, the collection lacks a fine-grained (≤10 min) dataset with large spatial coverage (more than wind farm size). However, additionally to the fact that we can only provide limited spatio-temporal resolution and coverage, wind power forecasting performance is dependent on terrain and site conditions.[4,8] The datasets presented cover neither very large nor off-shore wind farms. However, especially in large (off-shore) wind farms, the influences of surrounding turbines can significantly decrease power generation; similar effects can be observed at neighboring wind farms.[5] Therefore, on the one hand, the datasets presented are valuable for wind power forecasting research because they provide a broad range of open-source wind and wind power data. However, on the other hand, the collection is not complete. Furthermore, past technological limits reduce the quality of some of the older datasets. Both are not only a drawback of this collection of datasets but a consequence of (power) data regulations. We hope that political decision-makers' desire for a sustainable energy supply will bring about corresponding changes and that wind power data will become available publicly more easily and quickly in the future. However, the datasets we present are only a well-prepared snapshot of datasets. If the availability of open-source renewable energy data grows in the upcoming years, research in the field of wind power forecasting will also become more FAIR (see Section 1). Other initiatives are working on maintaining an up-to-date resource of open wind power data, such as the OPSD project,[169] the Wind Resource Assessment Group,[19] or the C3S Energy operational service.[170] These resources are of high value for experienced energy researchers, and we hope to provide a clear and easy implementable starting point for researchers from all different fields.

We compiled the datasets listed in this paper in several different ways: Searching online for datasets, getting in contact with wind power forecasting researchers from every continent to ask for available open-source data and energy data regulations in their region, and searching for papers that work with disclosed data. There is no guarantee for the completeness of the data. While we did not find other SCADA datasets, more countries probably provide or will provide some form of aggregated wind power data, and we know that many more resources of wind data exist. While we try to present all datasets that are hard to find and of high interest, such as the turbine-level datasets in Section 2.3, the aggregated and wind-based datasets should be seen as representatives of their corresponding groups. In these categories, we tried to cover different spatial and temporal resolutions.

7 | CONCLUSION

Wind power forecasting is an essential tool that can help to integrate wind power efficiently into the power grid. In addition, wind power forecasting can be used to make turbine-specific adjustments, and the mapping of wind speed to wind power allows for wind power resource assessment. All of this makes high-quality wind power forecasting and other forms of renewable energy forecasting important in developing a more sustainable power grid. This paper provides a categorization and overview of open-source wind power datasets that can be used for various wind power forecasting tasks. These tasks include wind power forecasting on different time and aggregation scales, wind ramp forecasting, wind variability forecasting and wind turbine condition monitoring. Wind power forecasting is a complex task. Its accuracy and quality depend not only on the aim of the forecast and the quality of the data but also on the real-world properties that the data aims to represent. Among these are, for example, the climatic characteristics of the country or region that is studied, terrain complexity and atmospheric (in-)stability and site-specific turbulences. This limits the comparability of models tested and validated on different datasets—the transferability of results is therefore hard to assess. However, a first step towards facing this issue can be taken by using open-source data and disclosing the research process. In order to make this search for datasets easier, we present the, to our best knowledge, largest variety of different datasets that can be used for different wind power forecasting tasks. Our categorization into five different data groups, such as fine-grained turbine-level power data, aggregated datasets, or even synthetic data, allows researchers to find a suitable dataset for various tasks. With more than 40 open-source datasets, we claim that suitable datasets do not have to be confidential for many tasks in wind power forecasting. Future work will benchmark different forecasting models on the open-source wind power datasets in this overview.

ACKNOWLEDGEMENTS

This study was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645 and the Athene Grant of the University of Tübingen. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Nina Effenberger. We acknowledge support by Open Access Publishing Fund of University of Tübingen. Open Access funding enabled and organized by Projekt DEAL. WOA Institution: Eberhard Karls Universitat Tubingen Consortia Name : Projekt DEAL.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/we.2766.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created. All referenced datasets are open-source.

ORCID

Nina Effenberger https://orcid.org/0000-0002-0713-1164
Nicole Ludwig https://orcid.org/0000-0003-3230-8918

REFERENCES

1. Renewables EU. Accessed Febuary 08, 2022. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Renewable_energy_statistics
2. Giebel G, Kariniotakis G. Wind power forecasting—a review of the state of the art. *Renewable Energy Forecasting: From Models to Applications*. Woodhead Publishing; 2017:59-109.
3. Hanifi S, Liu X, Lin Z, Lotfian S. A critical review of wind power forecasting methods—past, present and future. *Energies*. 2020;13(15):3764.
4. Kariniotakis G, Martí I, Casas D, et al. What performance can be expected by short-term wind power prediction models depending on site characteristics. In: CD-Rom Proceedings of the European Wind Energy Conference EWEC; 2004:1-9.
5. Nygaard NG. Wakes in very large wind farms and the effect of neighbouring wind farms. *J Phys Conf Ser*. 2014;524:12162.
6. Barthelmie RJ, Pryor SC, Frandsen ST, et al. Quantifying the impact of wind turbine wakes on power output at offshore wind farms. *J Atmos Oceanic Technol*. 2010;27(8):1302-1317.

7. McKay P, Carriveau R, Ting DS-K. Wake impacts on downstream wind turbine performance and yaw alignment. *Wind Energy*. 2013;16(2):221-234.
8. Marti I, Kariniotakis G, Pinson P, et al. Evaluation of advanced wind power forecasting models—results of the anemos project. In: Proc. of the European Wind Energy Conference; 2006:hal-00526668.
9. Menezes D, Mendes M, Almeida JA, Farinha T. Wind farm and resource datasets: a comprehensive survey and overview. *Energies*. 2020;13(18):4702.
10. Clifton A, Hodge B-M, Draxl C, Badger J, Habte A. Wind and solar resource data sets. *Wiley Interdisciplin Rev Energy Environ*. 2018;7(2):e276.
11. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The fair guiding principles for scientific data management and stewardship. *Scientif data*. 2016;3(1):1-9.
12. El-Gebali S, Mistry J, Bateman A, et al. The PFAM protein families database in 2019. *Nucleic Acids Res*. 2019;47(D1):D427-D432.
13. Sinaci AA, Núñez-Benjumea FJ, Gencturk M, et al. From raw data to fair data: the fairification workflow for health research. *Methods Inform Med*. 2020;59(S 01):e21-e32.
14. Vuong Q-H, Bui Q-K, La V-P, et al. Cultural additivity: behavioural insights from the interaction of confucianism, buddhism and taoism in folktales. *Palgrave Commun*. 2018;4(1):1-15.
15. Frank CW, Kaspar F, Keller JD, et al. Fair: A project to realize a user-friendly exchange of open weather data. *Adv Sci Res*. 2020;17:183-190.
16. Presse- und Informationsamt der Bundesregierung. Accessed January 28, 2022. https://www.bundesregierung.de/breg-de/suche/open-data-strategie-1939808
17. Open data EU. Accessed January 28, 2022. https://data.europa.eu/en/news/open-data-maturity-report-2021-out
18. IEA wind task 36, benchmarks. Accessed January 28, 2022. https://www.ieawindforecasting.dk/work-packages/workpackage-2/task-2-3
19. Wind resource assessment group. Accessed April 08, 2022. https://groups.io/g/wrag/wiki/13236/162238
20. Vargas SA, Esteves GRT, Maçaira PM, Bastos BQ, Oliveira FLC, Souza RC. Wind power generation: a review and a research agenda. *J Cleaner Prod*. 2019;218:850-870.
21. Public datasets by ENTSO-E. Accessed January 28, 2022. https://transparency.entsoe.eu
22. Lin Z, Liu X. Wind power forecasting of an offshore wind turbine based on high-frequency SCADA data and deep learning neural network. *Energy*. 2020;201:117693.
23. Bilal B, Ndongo M, Adjallah KH, et al. Wind turbine power output prediction model design based on artificial neural networks and climatic spatiotemporal data. In: 2018 IEEE International Conference on Industrial Technology (ICIT) Proceedings of the IEEE International Conference on Industrial Technology; 2018:1085-1092.
24. Gao L, Hu H. Wind turbine icing characteristics and icing-induced power losses to utility-scale wind turbines. *Proc Nat Acad Sci*. 2021;118(42):e2111461118.
25. Battisti L. *Wind Turbines in Cold Climates: Icing Impacts and Mitigation Systems*: Springer; 2015.
26. Corten GP, Veldkamp HF. Insects can halve wind-turbine power. *Nature*. 2001;412(6842):41-42.
27. Bartolomé L, Teuwen J. Prospective challenges in the experimentation of the rain erosion on the leading edge of wind turbine blades. *Wind Energy*. 2019;22(1):140-151.
28. Astolfi D, Castellani F, Lombardi A, Terzi L. Multivariate scada data analysis methods for real-world wind turbine power curve monitoring. *Energies*. 2021;14(4):1105.
29. Elexon datasets. Accessed January 28, 2022. https://www.elexon.co.uk/documents/training-guidance/bsc-guidance-notes/bmrs-api-and-data-push-user-guide-2/
30. Grogg K. Harvesting the wind: the physics of wind turbines. *Phys Astron Comps Papers*. 2005;7:1-41.
31. Zjavka L, Mišák S. Direct wind power forecasting using a polynomial decomposition of the general differential equation. *IEEE Trans Sustain Energy*. 2018;9(4):1529-1539.
32. Rienecker MM, Suarez MJ, Gelaro R, et al. MERRA: NASA's modern-era retrospective analysis for research and applications. *J Climate*. 2011;24(14):3624-3648.
33. Colak I, Sagiroglu S, Yesilbudak M. Data mining and wind power prediction: a literature review. *Renew Energy*. 2012;46:241-247.
34. Rediske G, Burin HP, Rigo PD, Rosa CB, Michels L, Siluk JCM. Wind power plant site selection: a systematic review. *Renew Sustain Energy Rev*. 2021;148:111293.
35. Bianco L, Djalalova IV, Wilczak JM, et al. A wind energy ramp tool and metric for measuring the skill of numerical weather prediction models. *Weather Forecast*. 2016;31(4):1137-1156.
36. Passos J, Sakagami Y, Santos P, Haas R, Taves F. Costal operating wind farms: Two datasets with concurrent SCADA, LiDAR and turbulent fluxes; 2017. https://doi.org/10.5281/zenodo.1475197
37. Sakagami Y. Influência da turbulência e do perfil de velocidade do vento no desempenho de aerogeradores em dois parques eólicos na costa no nordeste brasileiro. *Ph.D. Thesis*: Universidade Federal de Santa Catarina; 2017.
38. Plumley C. Penmanshiel wind farm data; 2022. https://doi.org/10.5281/zenodo.5946808
39. Cubico sustainable investments ltd. Accessed January 28, 2022. https://www.cubicoinvest.com
40. Delabole wind farm data. Accessed January 28, 2022. https://data.dtu.dk/articles/dataset/Scada_data_from_Delabole_wind_farm/14077004
41. Delabole documentation. Accessed January 28, 2022. https://gitlab.windenergy.dtu.dk/fair-data/winddata-revamp/winddata-documentation/-/blob/master/delabole.md
42. Dtu database. Accessed January 28, 2022. https://gitlab.windenergy.dtu.dk/fair-data/winddata-revamp/winddata-documentation
43. Plumley C. Kelmarsh wind farm data; 2022. https://doi.org/10.5281/zenodo.5841834
44. La haute borne wind farm data and documentation. Accessed January 28, 2022. https://opendata-renewables.engie.com/explore/index
45. La haute borne wind farm open data policy. Accessed January 28, 2022. https://opendata-renewables.engie.com
46. Tjareborg wind farm data. Accessed January 28, 2022. https://data.dtu.dk/articles/dataset/Wind_resource_SCADA_data_and_time_series_of_wind_and_turbine_loads_from_Tjareborg_DK/16701961
47. Tjareborg documentation. Accessed January 28, 2022. https://gitlab.windenergy.dtu.dk/fair-data/winddata-revamp/winddata-documentation/-/blob/master/tjare.md
48. Manufacturer tjareborg. Accessed January 28, 2022. https://en.wind-turbine-models.com/turbines/1012-elsam-tj-reborg

49. Tjareborg metmast data. Accessed November 16, 2021. https://gitlab.windenergy.dtu.dk/fair-data/winddata-revamp/winddata-documentation/-/blob/master/tjare_2.md
50. Eolos wind research station data and documentation. Accessed January 28, 2022. https://conservancy.umn.edu/handle/11299/205162
51. Davison B. Rich data for wind turbine power performance analysis. Retrieved from the data repository for the university of minnesota; 2019. https://doi.org/10.13020/1etn-1q17
52. Davison B. The evaluation of data filtering criteria in wind turbine power performance assessment. *Ph.D. Thesis*: Edinburgh Napier University; 2019.
53. Edp wind turbine dataset. Accessed January 28, 2022. https://opendata.edp.com/explore/?refine.keyword=visible&refine.keyword=Wind+Farm+1&sort=modified
54. Installed capacity EDP. Accessed April 05, 2022. https://opendata.edp.com/pages/Windfarms/
55. Edp open data policy. Accessed January 28, 2022. https://opendata.edp.com/pages/aboutus/
56. First kaggle dataset. Accessed January 28, 2022. https://www.kaggle.com/theforcecoder/wind-power-forecasting
57. Second kaggle dataset. Accessed January 28, 2022. https://www.kaggle.com/wasuratme96/iiot-data-of-wind-turbine
58. Wind farm data maelstrom. Accessed January 28, 2022. https://github.com/4castRenewables/climetlab-plugin-a6
59. Maelstrom documentation. Accessed January 28, 2022. https://www.maelstrom-eurohpc.eu/content/docs/uploads/doc6.pdf
60. Yalova wind turbine dataset. Accessed January 28, 2022. https://www.kaggle.com/berkerisen/wind-turbine-scada-dataset
61. Nm92 wind farm dataset. Accessed January 28, 2022. https://data.dtu.dk/articles/dataset/_/16743739
62. Nm92 documentation. Accessed January 28, 2022. https://gitlab.windenergy.dtu.dk/fair-data/winddata-revamp/winddata-documentation/-/blob/master/nm92.md
63. Nordtank wind farm dataset. Accessed January 28, 2022. https://data.dtu.dk/articles/dataset/Turbulence_and_turbine_measurements_from_the_Nordtank_turbine/16782598
64. Nordtank documentation. Accessed January 28, 2022. https://gitlab.windenergy.dtu.dk/fair-data/winddata-revamp/winddata-documentation/-/blob/master/nordtank.md
65. Monthly wind power denmark. Accessed January 28, 2022. https://ens.dk/en/our-services/statistics-data-key-figures-and-energy-maps/overview-energy-sector
66. Ding Y. Wind spatio-temporal dataset2; 2021. https://doi.org/10.5281/zenodo.5516550
67. Data science for wind energy, texas a&m university (engineering). Accessed January 28, 2022. https://aml.engr.tamu.edu/book-dswe/dswe-datasets/
68. Ding Y. *Data Science for Wind Energy*: CRC Press; 2019.
69. Ding Y. Wind spatio-temporal dataset1; 2021. https://doi.org/10.5281/zenodo.5516543
70. Ding Y. Wind spatial dataset; 2021. https://doi.org/10.5281/zenodo.5516541
71. Ding Y. Inland-offshore wind farm dataset1; 2021. https://doi.org/10.5281/zenodo.5516552
72. Ding Y. Inland-offshore wind farm dataset2; 2021. https://doi.org/10.5281/zenodo.5516554
73. Ding Y. Wind time series dataset; 2021. https://doi.org/10.5281/zenodo.5516539
74. Entso-e detailed data description. Accessed April 04, 2022. https://eepublicdownloads.entsoe.eu/clean-documents/pre2015/resources/Transparency/MoP_Ref_02_-_Detailed_Data_Descriptions_v1r2.pdf
75. Entso-e python package. Accessed January 28, 2022. https://github.com/EnergieID/entsoe-py
76. De Felice M. ENTSO-E Actual Generation of Wind units: data from 21-12-2014 to 11-04-2021; 2021. https://doi.org/10.5281/zenodo.4682697
77. De Felice M. ENTSO-E Pan-European Climatic Database (PECD 2021.3) in Parquet format; 2021. https://doi.org/10.5281/zenodo.5780185
78. Locations ENTSO-E. Accessed January 28, 2022. https://data.jrc.ec.europa.eu/dataset/9810feeb-f062-49cd-8e76-8d8cfd488a05
79. Entso-e. Accessed January 28, 2022. https://www.entsoe.eu
80. Data available entso-e. Accessed April 04, 2022. https://transparency.entsoe.eu/content/static_content/download?path=/Static%20content/terms%20and%20conditions/220218_List_of_Data_available_for_reuse.pdf
81. Elexon operational data. Accessed April 13, 2022. https://www.elexonportal.co.uk/category/view/179
82. Elexon python package. Accessed January 28, 2022. https://github.com/OSUKED/ElexonDataPortal
83. Elexon portal. Accessed January 28, 2022. https://www.elexonportal.co.uk
84. License elexon. Accessed April 07, 2022. https://www.elexon.co.uk/operations-settlement/bsc-central-services/balancing-mechanism-reporting-agent/copyright-licence-bmrs-data/
85. New zealand wind power generation. Accessed January 28, 2022. https://www.emi.ea.govt.nz/Wholesale/Datasets/Generation/Generation_MD/
86. New Zealand network supply points. Accessed January 28, 2022. https://www.emi.ea.govt.nz/Wholesale/Reports/R_NSPL_DR?_si=v|3
87. Pyproj python package. Accessed January 28, 2022. https://pyproj4.github.io/pyproj/stable/index.html#
88. Brazilian wind power generation. Accessed January 28, 2022. http://www.ons.org.br/Paginas/resultados-da-operacao/historico-da-operacao/geracao_energia.aspx
89. Brazilian data transparency of ons. Accessed January 28, 2022. http://www.ons.org.br/paginas/resultados-da-operacao/historico-da-operacao/dados-gerais
90. License ONS. Accessed April 06, 2022. https://dados.ons.org.br/about
91. South african wind power generation. Accessed January 28, 2022. http://redis.energy.gov.za/electricity-production-details/
92. Download south african wind power generation. Accessed January 28, 2022. http://redis.energy.gov.za/how-to-download-data/
93. South african data source and referencing. Accessed January 28, 2022. http://redis.energy.gov.za/data-source-and-referencing/
94. Gefcom2014 dataset. Accessed January 28, 2022. https://www.dropbox.com/s/pqenrr2mcvl0hk9/GEFCom2014.zip?dl=0
95. Hong T, Pinson P, Fan S, Zareipour H, Troccoli A, Hyndman RJ. Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond: Elsevier; 2016.
96. Australian dataset. Accessed January 28, 2022.
97. Aemo website. Accessed April 05, 2022. https://aemo.com.au
98. Dowell J, Pinson P. Very-short-term probabilistic wind power forecasts by sparse vector autoregression. *IEEE Trans Smart Grid*. 2015;7(2):763-770.

99. Godahewa R, Bergmeir C, Webb G, Abolghasemi M, Hyndman R, Montero-Manso P. Wind power dataset (4 seconds observations); 2020. https://doi.org/10.5281/zenodo.4656032
100. Godahewa R, Bergmeir C, Webb GI, Hyndman RJ, Montero-Manso P. Monash time series forecasting archive. arXiv preprint:210506643; 2021.
101. Godahewa R, Bergmeir C, Webb G, Abolghasemi M, Hyndman R, Montero-Manso P. Wind farms dataset (with missing values); 2020. https://doi.org/10.5281/zenodo.4654909
102. Godahewa R, Bergmeir C, Webb G, Abolghasemi M, Hyndman R, Montero-Manso P. Wind farms dataset (without missing values); 2020. https://doi.org/10.5281/zenodo.4654858
103. Eem 2020 dataset. Accessed January 28, 2022. https://pureportal.strath.ac.uk/en/datasets/data-and-code-for-the-eem2020-wind-power-forecasting-competition
104. European energy market conference, eem 2020. Accessed January 28, 2022. https://eem20.eu/forecasting-competition/
105. MERRA data. Accessed January 28, 2022. https://disc.gsfc.nasa.gov/datasets?project=MERRA
106. MERRA documentation. Accessed January 28, 2022. https://goldsmr2.gesdisc.eosdis.nasa.gov/data/AtrainMERRA/MAT1NXSLV_CPR.5.2.0/doc/MERRA.README.pdf
107. Merra 2. Accessed January 28, 2022. https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/
108. MERRA NASA. Accessed January 28, 2022. https://gmao.gsfc.nasa.gov/reanalysis/MERRA/
109. Nasa data policies. Accessed April 13, 2022. https://earthdata.nasa.gov/collaborate/open-data-services-and-software
110. ERA 5 data. Accessed January 28, 2022. https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=form
111. ECMWF reanalysis data. Accessed January 28, 2022. https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5
112. Public datasets by ECMWF. Accessed January 28, 2022. https://www.ecmwf.int/en/forecasts/datasets/wmo-essential
113. Tigge. Accessed April 11, 2022. https://confluence.ecmwf.int/display/TIGGE
114. S2s. Accessed April 11, 2022. https://www.ecmwf.int/en/research/projects/s2s
115. Ecmwf member states. Accessed April 11, 2022. https://www.ecmwf.int/en/about/who-we-are/member-states
116. Ecmwf hres. Accessed April 11, 2022. https://www.ecmwf.int/en/forecasts/datasets/set-i
117. Ecmwf ensemble model. Accessed April 11, 2022. https://www.ecmwf.int/en/forecasts/datasets/set-iii
118. Noaa. Accessed April 11, 2022. https://www.ncei.noaa.gov/products/weather-climate-models/numerical-weather-prediction
119. Noaa gefs. Accessed April 11, 2022. https://www.ncei.noaa.gov/products/weather-climate-models/global-ensemble-forecast
120. Noaa gfs. Accessed April 11, 2022. https://www.emc.ncep.noaa.gov/emc/pages/numerical_forecast_systems/gfs/documentation.php
121. Orsted offshore data. https://orsted.com/en/our-business/offshore-wind/wind-data
122. Fino 2 documentation. Accessed January 28, 2022. https://www.fino2.de/en/fino2/location.html
123. Energydata. Accessed January 28, 2022. https://energydata.info/dataset
124. Energydata responsibles. Accessed January 28, 2022. https://energydata.info/about_us
125. Tall tower dataset. Accessed January 28, 2022. https://talltowers.bsc.es
126. Ramon J, Lledó L, Pérez-Zanón N, Soret A, Doblas-Reyes FJ. The tall tower dataset: a unique initiative to boost wind energy research. *Earth Syst Sci Data*. 2020;12(1):429-439.
127. WIND toolkit data. Accessed January 28, 2022. https://www.nrel.gov/grid/wind-toolkit.html
128. Howland MF, Lele SK, Dabiri JO. Wind farm power optimization through wake steering. *Proc Nat Acad Sci*. 2019;116(29):14495-14500.
129. Chen Y, Wang Y, Kirschen D, Zhang B. Model-free renewable scenario generation using generative adversarial networks. *IEEE Trans Power Syst*. 2018;33(3):3265-3275.
130. Draxl C, Clifton A, Hodge B-M, McCaa J. The wind integration national dataset (wind) toolkit. *Appl Energy*. 2015;151:355-366.
131. License wind toolkit. Accessed April 06, 2022. https://registry.opendata.aws/nrel-pds-wtk/
132. Renewables ninja dataset. Accessed January 28, 2022. https://www.renewables.ninja/downloads#details-wind
133. Renewables ninja. Accessed January 28, 2022. https://www.renewables.ninja
134. Staffell I, Pfenninger S. Using bias-corrected reanalysis to simulate current and future wind power output. *Energy*. 2016;114:1224-1239.
135. Emhires dataset. Accessed January 28, 2022. https://data.jrc.ec.europa.eu/dataset/jrc-emhires-wind-generation-time-series#dataaccess
136. Eurostats NUTS. Accessed January 28, 2022. https://ec.europa.eu/eurostat/web/nuts/history
137. Moraes Jr L, Bussar C, Stoecker P, Jacqué K, Chang M, Sauer DU. Comparison of long-term wind and photovoltaic power capacity factor datasets with open-license. *Appl Energy*. 2018;225:209-220.
138. Gonzalez Aparicio I, Zucker A, Careri F, Monforti F, Huld T, Badger J. Emhires dataset part I: wind power generation. 2016.
139. European commission reuse and copyright notice. Accessed April 06, 2022. https://data.jrc.ec.europa.eu/licence/com_reuse
140. Gilbert C, Browell J, McMillan D. Leveraging turbine-level data for improved probabilistic wind power forecasting. *IEEE Trans Sustain Energy*. 2019;11(3):1152-1160.
141. Ferreira C, Gama J, Matias L, Botterud A, Wang J. A Survey on Wind Power Ramp Forecasting. tech. rep., Argonne, IL (United States), Argonne National Lab.(ANL); 2011.
142. Gallego-Castillo C, Cuerva-Tejero A, Lopez-Garcia O. A review on the recent history of wind power ramp forecasting. *Renew Sustain Energy Rev*. 2015;52:1148-1157.
143. Davy RJ, Woods MJ, Russell CJ, Coppin PA. Statistical downscaling of wind variability from meteorological fields. *Bound-Layer Meteorol*. 2010;135(1):161-175.
144. Global power plant database. Accessed January 28, 2022. https://datasets.wri.org/dataset/globalpowerplantdatabase
145. Holttinen H, Saarikivi P, Repo S, Ikäheimo J, Koreneff G. Prediction errors and balancing costs for wind power production in Finland. In: Proceedings of 6th workshop on Offshore and Large Scale Integration of Wind Power; 2006: 1-11.
146. Demolli H, Dokuz AS, Ecemis A, Gokcek M. Wind power forecasting based on daily wind speed data using machine learning algorithms. *Energy Conv Manag*. 2019;198:111823.
147. Wang Y, Hu Q, Li L, Foley AM, Srinivasan D. Approaches to wind power curve modeling: a review and discussion. *Renew Sustain Energy Rev*. 2019;116:109422.

148. Antoniou I, Friis Pedersen T. Nacelle anemometry on a 1 MW wind turbine. Comparing the power performance results by use of the nacelle or mast anemometer; 1997.
149. Cutler NJ, Outhred HR, MacGill IF. Using nacelle-based wind speed observations to improve power curve modeling for wind power forecasting. *Wind Energy*. 2012;15(2):245-258.
150. Draxl C, Hahmann AN, Peña A, Giebel G. Evaluating winds and vertical wind shear from weather research and forecasting model forecasts using seven planetary boundary layer schemes. *Wind Energy*. 2014;17(1):39-55.
151. Foley AM, Leahy PG, Marvuglia A, McKeogh EJ. Current methods and advances in forecasting of wind power generation. *Renew Energy*. 2012;37(1):1-8.
152. Bossavy A, Girard R, Kariniotakis G. Forecasting ramps of wind power production with numerical weather prediction ensembles. *Wind Energy*. 2013;16(1):51-63.
153. Chen N, Qian Z, Nabney IT, Meng X. Wind power forecasts using gaussian processes and numerical weather prediction. *IEEE Trans Power Syst*. 2013;29(2):656-665.
154. Al-Yahyai S, Charabi Y, Gastli A. Review of the use of numerical weather prediction (NWP) models for wind energy assessment. *Renew Sustain Energy Rev*. 2010;14(9):3192-3198. http://www.sciencedirect.com/science/article/pii/S1364032110001814
155. Sanchez I. Short-term prediction of wind energy production. *Int J Forecast*. 2006;22(1):43-56.
156. Costoya X, Rocha A, Carvalho D. Using bias-correction to improve future projections of offshore wind energy resource: a case study on the iberian peninsula. *Appl Energy*. 2020;262:114562.
157. Buhan S, Özkazanç Y, Çadırcı I. Wind pattern recognition and reference wind mast data correlations with NWP for improved wind-electric power forecasts. *IEEE Trans Indust Inform*. 2016;12(3):991-1004.
158. Zou M, Djokic SZ. A review of approaches for the detection and treatment of outliers in processing wind turbine and wind farm measurements. *Energies*. 2020;13(16):4228.
159. Morrison R, Liu X, Lin Z. Anomaly detection in wind turbine SCADA data for power curve cleaning. *Renew Energy*. 2022;184:473-486.
160. He B, Ye L, Pei M, et al. A combined model for short-term wind power forecasting based on the analysis of numerical weather prediction data. *Energy Rep*. 2022;8:929-939.
161. Probst O, Cárdenas D. State of the art and trends in wind resource assessment. *Energies*. 2010;3(6):1087-1141.
162. Bauer P, Thorpe A, Brunet G. The quiet revolution of numerical weather prediction. *Nature*. 2015;525(7567):47-55.
163. Lantz EJ, Roberts JO, Nunemaker J, DeMeo E, Dykes KL, Scott GN. Increasing wind turbine tower heights: Opportunities and challenges; 2019.
164. Increasing capacity. Accessed April 12, 2022. https://www.energy.gov/eere/articles/wind-turbines-bigger-better
165. Height eolos wind turbine. Accessed April 12, 2022. https://www.cleanenergyresourceteams.org/clean-energy-blows-umore-eolos-wind-research-station
166. Hornsea 2. Accessed April 12, 2022. https://hornseaprojects.co.uk/hornsea-project-two
167. Global wind atlas. Accessed April 12, 2022. http://energybc.ca/wind.html
168. Morrison R, Liu X, Lin Z. Anomaly detection in wind turbine SCADA data for power curve cleaning. *Renew Energy*. 2022;184:473-486.
169. Opsd project. Accessed April 07, 2022. https://open-power-system-data.org
170. C3s energy operational service. Accessed April 07, 2022. https://climate.copernicus.eu/operational-service-energy-sector

How to cite this article: Effenberger N, Ludwig N. A collection and categorization of open-source wind and wind power datasets. *Wind Energy*. 2022;1-25. doi:10.1002/we.2766