

Manuscript Number:

Title: Open Power System Data - Frictionless data for electricity system modelling

Article Type: VSI:Data on Energy

Keywords: electricity system modelling; frictionless data; data transparency; data package; data platform; open data

Corresponding Author: Dr. Ingmar Schlecht,

Corresponding Author's Institution: University of Basel

First Author: Frauke Wiese

Order of Authors: Frauke Wiese; Ingmar Schlecht; Wolf-Dieter Bunke; Clemens Gerbaulet; Lion Hirth; Martin Jahn; Friedrich Kunz; Casimir Lorenz; Jonathan Mühlenpfordt; Juliane Reimann; Wolf-Peter Schill

Abstract: The quality of electricity system modelling heavily depends on the input data used. Although a lot of data is publicly available, it is often dispersed, tedious to process and partly contains errors. We argue that a central provision of input data for modelling has the character of a public good: it reduces overall societal costs for quantitative energy research as redundant work is avoided, and it improves transparency and reproducibility in electricity system modelling. This paper describes the Open Power System Data platform that aims at realising the efficiency and quality gains of centralised data provision by collecting, checking, processing, aggregating, documenting and publishing data required by most modellers. We conclude that the platform can provide substantial benefits to the modelling community, and that it can also contribute to improving the quality of original data published by respective providers.

Highlights

- Quality of electricity system modelling heavily relies on input data.
- Centralised and open provision of input data increases both transparency and efficiency.
- A lot of data is publicly available, but often dispersed and tedious to process.
- The Open Power System Data platform is an approach to tackle these challenges.

Open Power System Data - Frictionless data for electricity system modelling

Frauke Wiese^a, Ingmar Schlecht^{b,c,*}, Wolf-Dieter Bunke^d, Clemens Gerbaulet^e,
Lion Hirth^{c,f,g}, Martin Jahn^d, Friedrich Kunz^{h,i}, Casimir Lorenz^{h,e,j}, Jonathan
Mühlenpfordt^{c,k}, Juliane Reimann^c, Wolf-Peter Schill^h

^aTechnical University of Denmark, DTU Management Engineering, Produktionstorvet,
Building 426, 2800 Kongens Lyngby, Denmark
^bUniversity of Basel, Faculty of Business and Economics, Peter Merian-Weg 6, 4052 Basel,
Switzerland
^cNeon Neue Energieökonomik GmbH (Neon), Karl-Marx-Platz 12, 12043 Berlin, Germany
^dEuropa-Universität Flensburg, Energy and Environmental Management, Auf dem Campus
1, 24943 Flensburg, Germany
^eTechnische Universität Berlin, Straße des 17. Juni 135, 10236 Berlin, Germany
^fHertie School of Governance, Friedrichstraße 180, 10117 Berlin, Germany
^gMercator Research Institute on Global Commons and Climate Change, Torgauer Straße
12, 10829 Berlin
^hDIW Berlin, Mohrenstraße 58, 10117 Berlin, Germany
ⁱTenneT TSO GmbH, Bernecker Straße 70, 95448 Bayreuth
^jSachverständigen Rat für Umweltfragen, Luisenstraße 46, 10117 Berlin, Germany
^kHumboldt-Universität zu Berlin, School of Business and Economics, Spandauer Straße 1,
10178 Berlin, Germany

Abstract

The quality of electricity system modelling heavily depends on the input data used. Although a lot of data is publicly available, it is often dispersed, tedious to process and partly contains errors. We argue that a central provision of input data for modelling has the character of a public good: it reduces overall societal costs for quantitative energy research as redundant work is avoided, and it improves transparency and reproducibility in electricity system modelling. This paper describes the Open Power System Data platform that aims at realising the efficiency and quality gains of centralised data provision by collecting, checking, processing, aggregating, documenting and publishing data required by most modellers. We conclude that the platform can provide substantial benefits to the modelling community, and that it can also contribute to improving the quality of original data published by respective providers.

Keywords: electricity system modelling, frictionless data, data transparency, data package, data platform, open data

*Corresponding author
Email address: `ingmar.schlecht@unibas.ch` (Ingmar Schlecht)

1. Introduction

Applied energy research as well as energy and climate policy advice are often based on quantitative computer models [1, 2]. Activities in the field of electricity system modelling [3] have grown substantially over the last years [4, 5]. This is, among other reasons, driven by increasing shares of wind and solar power in many countries, and the desire to answer various questions of integrating such variable renewable energy sources into power systems and markets. For example, authors of this article have used European electricity models for analyses of wholesale prices [6], the integration of renewables [7], the changing dispatch of thermal power plants [8], the need for storage requirements [9], changing requirements for transmission grids [10–14], system effects of electric vehicles [15], or optimal configurations of 100% renewable electricity sectors and their respective transition pathways [16].

Since many model-based analyses consider the conditions prevailing in today’s power sectors as a starting point, current data is generally needed. While many model applications differ with respect to data requirements — caused by different model features and different geographical and temporal scopes — input data on generation capacities as well as time series of load and renewable power generation are relevant input parameters for the majority of electricity system models. Several models also require hourly wholesale prices as input parameters and for some weather data itself is an essential input. The quality of such input data has a major influence on the scientific quality of model-based studies and on their usefulness with respect to policy conclusions.

Yet not only the quality of input data, but also its availability and transparent application is essential for high-quality energy research. In recent years, energy system modelling has been increasingly criticised for its black box character [17, 18] while openness of code and data are identified as key requirements for energy system models to comply with scientific standards like improved reproducibility and greater scrutiny [19]. By allowing reuse and collaborative development, open models and data can increase productivity and – through greater transparency [20] – also increase credibility in the policy discourse [16, 21].

In general, the public availability of European electricity system data has increased significantly during the last years due to rising transparency regulation. An example of this has been the introduction of the ENTSO-E Transparency Platform [22] (see [23] in this issue for a review). Yet, data is provided by many different institutions and thus often dispersed over various repositories and formatted in inconsistent ways. Further, data is often poorly documented and sometimes includes obvious errors or missing values. In particular, there are challenges related to:

- the identification of appropriate sources
- combinations of different data sources
- manual downloads of multiple files

- merging and harmonising different file formats and inconsistent formatting within files
- a lack of a standardised nomenclature and classification of energy sources and technologies
- poorly documented original data, in particular a lack of metadata
- data quality (e.g inconsistencies, obvious errors, gaps in data)

Overall, the collection, processing, maintenance and verification of input data imply substantial work loads for electricity system modellers. So far, such activities are performed by many modelling teams in parallel, and repeated regularly, whenever new data becomes available, i.e. each year or even more frequently. It would clearly be much more efficient if these tasks would be conducted in a collaborative and/or centralised manner. Likewise, the quality of the processed data is likely to improve by knowledge sharing among electricity system data users.

Taking into consideration the relevance of input data for electricity system modelling, the need for its transparency and the challenges arising from the current situation, we present an approach to tackle these challenges in this paper: we have initiated the project *Open Power System Data (OPSD)* which aims at realising the efficiency and quality gains for the modelling community by collecting, checking, processing, aggregating, documenting and publishing data required for European electricity system modelling on an online platform. OPSD was set up in the context of a research project (first phase between 2015-2017) and is in its second phase at the time of writing (2018-ongoing).

The concept and methods of OPSD are described in section 2, while section 3 explains the content and value added of the five OPSD data packages. We conclude with a reflection on the current usage of the data platform as well as possible extensions and remaining challenges (Section 4).

2. The OPSD concept

To tackle the challenges outlined above, the authors created a central data platform, OPSD [24], which provides the key variables used for power system modelling at a central location in an aggregated and ready-to-download form. OPSD is a free-of-charge data platform dedicated to electricity system researchers. The data provided on the platform is collected from individual sources, checked, processed, documented and published on the platform.

2.1. Definition of target group and scope

The idea behind OPSD is to avoid redundant work when collecting, preparing and aggregating data for energy modelling. Thereby, the project is a service provider to the modelling community. The target group of OPSD are energy professionals such as modellers and analysts in academia, consulting firms and

industry. This choice of target group is instrumental in guiding the conceptual decisions for the data platform. We focus on long historical time series rather than high frequency updates of real-time information; and on providing bare CSV files rather than visualisations of data.

The selection of data we provide is based on the data inputs common to most electricity models as outlined in Section 1, namely time series data (solar, wind, load and prices), individual as well as aggregated power plant data (conventional and renewable) and weather data. We generally use data from official sources such as from statistical authorities or transmission system operators (TSOs) and only make careful and well-documented changes to the original data.

2.2. Transparency improves quality and reliability

To allow for both scientific and policy-relevant use of the provided data, it is crucial to ensure high quality, trustworthiness and reliability. The OPSD concept was designed to meet these criteria. Besides choosing to only rely on data from official sources, the main principle in this regard are open source scripts, transparency of used input data and documentation of both the script as well as the final output.

Along with each of our data packages, we provide the scripts that are used to generate them. These scripts implement all steps including downloading of data from the original sources, processing and reformatting and writing the final output data for download on the OPSD platform. This not only enables data users to read the script to see how the data was treated (such as finding out which method to interpolate missing values was used), but also to take the script and change parts of it in order to implement alternative data processing steps at any point in the pipeline. Users can still benefit from the remainder of the script to generate clean output data.

We put a special emphasis on documentation and chose a two-level approach towards information about data on the platform. On a first level, for users mainly interested in knowing what is to be found in the data files for download, we provide a first short description of data contents along with column-by-column documentation of individual data files directly on the download page. For those who want to dive deeper and know exactly how the data was treated, we use Jupyter notebooks, a format that combines script code and documentation in a single file. The notebooks contain the most detailed documentation of the provided data files and enable users to find out how the data was sourced, treated and packaged up to the final product.

We refrain from making large-scale adjustments to the data in order to remain close to the original sources. This enables data users to choose their individual strategies to treat the data for their use case and trust in the provided data to be close to original sources. However, we do make careful modifications. These include the correction of obvious errors in the original data such as GW-scale rooftop solar plants, alphabetical characters in numerical data fields, missing commas, or aligning of different spellings for the same categorical labels. To remain transparent to the data user about modifications that were applied, a separate marker column marks each instance where data was changed by OPSD.

A frequent problem with data in the power sector is that data that was once available is removed from the original sources, or the location changes. We tackle this problem on the one hand by making a copy of the original input data available on our server next to the data package using it. On the other hand, we try to do better on our side and keep historic versions of data packages online under a stable and clear URL. This will be complemented in the near future by stable Digital Object Identifiers (DOIs) that can be used to reference individual versions of data packages.

While the majority of data packages on our platform are provided by the core team behind OPSD, we also invite data contributions by third parties. Contributions have to be relevant to our target audience, conform to the technical standards (see Section 2.3) and be maintained for a prolonged time period to be admitted to the platform.

2.3. Lightweight and scalable technical architecture

On all technical aspects of our platform, we aim at lightweight decentralised solutions rather than complex centralised structures. We build on established industry standards for open data (Data Packages), existing open-source software solutions (Jupyter Notebooks and Python) and collaborative software development tools (Git and GitHub).

The guiding principle by which we organise the data on our website comes from the Data Package standard [25], a standard invented by Open Knowledge International to foster frictionless data. The standard is very simple and relies on existing formats. It defines a tabular data package to consist of at least one comma-separated values (CSV) file and a text file containing structured metadata in the JSON format.

The CSV file format allows for easy and fast download, as no database is required on the server side. Furthermore, since CSV files are human-readable and can be used by almost any software, they are a well-established format. However, some variations of the standards exist, making usage not always intuitive. We tackle this problem by following the tabular data package standard very strictly for our CSV files and by providing Excel and SQLite files in addition.

Metadata is supplied in JSON files, which are both human-readable as well as machine-readable and contain information about properties such as title of the package, sources, contributors and the structure of the CSV files contained in the data package. The machine readability of the metadata files allows users of the data to automatically feed the data into their systems.

The scripts that we use to automatically download, process, and aggregate data are written in Python. We use Jupyter Notebooks as a single file for coding and documentation. A Jupyter Notebook is a documents that contains live code, equations, visualisations and explanatory text. The notebooks created in this project are available under the open-source MIT license on GitHub [26].

2.4. Central vs. decentral data preparation

Section 1 has highlighted a number of key issues with data for power system modelling today, often being scattered across multiple sources, non-open

and time-consuming to process. Therefore, central data platforms, providing a solution to these problems, are undoubtedly beneficial in regards to cutting the time and cost involved in data preparation. However, centralisation of data preparation also entails certain risks, which we outline and discuss below, as they are crucial to the key choices made in creating the platform.

First, a major risk in central data provision concerns error propagation. A single error in data preparation on the central data platform could propagate into many studies based on the centrally provided data. We tackle this risk by providing our scripts open source, so that others are invited to use and scrutinise the scripts (which, according to the detailed feedback we get, actually takes place). This makes error-detection easier than on closed-source platforms, where users do not get to see the programming code. Furthermore, we mark all instances where we intentionally modify and correct data, which points users at the most critical data points to scrutinise.

Second, there might be a bias in data selection. While for some data sets many competing sources are available (such as installed power plant capacity figures being available from statistical offices, energy ministries and TSOs), the data source selected by a central platform such as OPSD might be over-represented in final data usage by scientists and analysts. We tackle this issue by providing, where available, competing data sources so that users can compare them against each other. This is especially the case for the national generation capacity data package (see Section 3.4), but also concerns the time series data package (see Section 3.5) regarding load values from two different ENTSO-E sources.

Third, a central data platform might lead to a reduction of the plurality of data preparation methods in use before. By centralising model data provision, all users of the centrally provided data will base their studies on the selected approaches instead of implementing their own methodologies in regard to interpolation of missing values or detection of improbable data outliers, which might lead to a bias if there are issues with the chosen treatment. We address this issue by applying data modifications carefully only and in cases where a correction is not possible and mark improbable data items, leaving it to the user to do final modifications.

In summary, we argue that while there are risks from centralising data preparation, these are to a large degree addressed by the design choices we take.

3. Five data packages

3.1. Overview

For the illustration of our work on and methods for tackling the challenges of input data for electricity system modelling, we describe scope, coverage and value added for each of the five currently available OPSD data packages. These cover a large share of input data required by electricity system models:

- Conventional power plants

- Renewable power plants
- (Aggregated) National generation capacity
- Time series
- Weather data

Due to a varying availability of original data, the geographical as well as the temporal coverage of the resulting data packages differ. Therefore, an overview of coverage is provided in the data availability matrix (Figure 1). Countries covered in the different data packages are indicated by blue coloured cells. The cell content further specifies the earliest year of availability (time series) or the type of power plants covered (renewable and conventional).

	Time series				List of power plants		National generation capacity	Weather data
	Prices	Load	Wind	Solar	Renewable	Conventional		
AT	2015+	2006+	2015+	2015+				
BE	2015+	2006+	2015+	2015+				
CH	2015+	2007+	2015+	2015+	KEV plants	Hydro, nuclear		
CZ	2015+	2006+	2012+	2012+		> 100 MW		
DE	2005+	2006+	2012+	2012+	EEG plants	> 10 MW		
DK	2006+	2010+	2014+	2014+	Wind, solar			
ES	2015+	2006+	2015+	2015+				
FI	2015+	2010+	2015+	2015+				
FR	2015+	2006+	2015+	2015+				
IT	2015+	2006+	2015+	2015+				
NL	2015+	2006+	2015+	2015+				
NO	2006+	2006+	2015+	2015+		> 100 MW		
PL	2015+	2008+	2013+	2015+		Centrally dispatched units		
SE	2006+	2005+	2005+	2011+		> 100 MW		
UK	2015+	2011+	2015+	2015+				
20+ more								

Figure 1: Overview of the current data availability (as of May 2018).

A time-consuming barrier for input data handling from different sources are divergent naming conventions and classifications from different original data sources. Thus, an essential contribution of OPSD for smoother input data handling and better comparability of original data is a consistent classification of energy sources and technologies (see Figure 2) which we consistently apply to the data packages where this is relevant, i.e. the first three in the list above. On the first level, we generally distinguish between renewable energy sources, fossil fuels, nuclear and other or unspecified sources. The second level contains different renewable sources such as wind and solar power as well as different

fossil fuels. The third level is necessary to distinguish between different forms of bio energy.

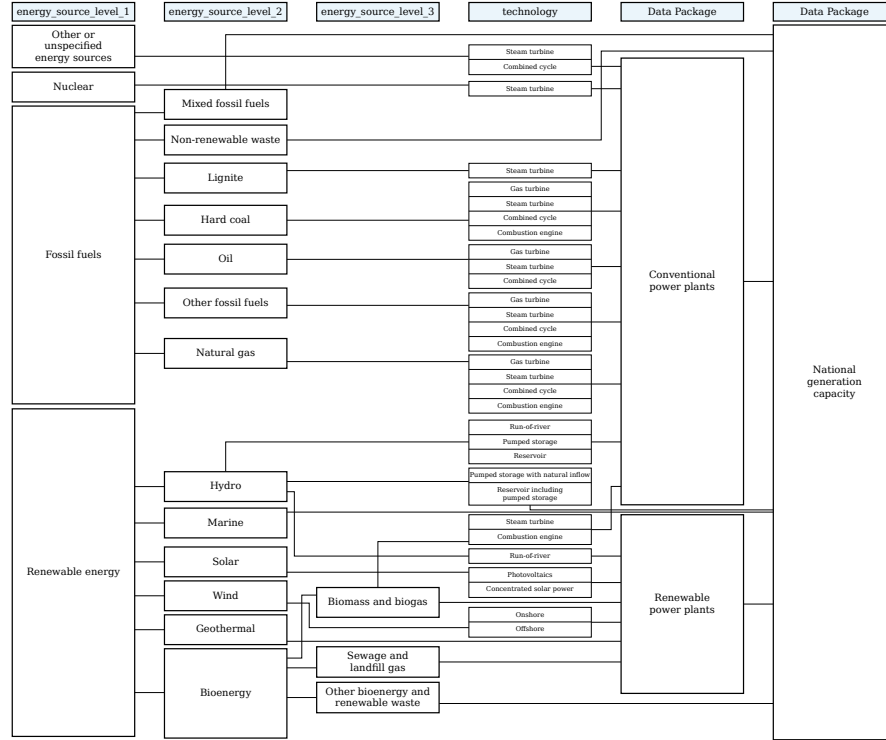


Figure 2: Classification of energy sources and technologies

In the following, we describe the content, scope and particular features of each data package. Please note that only the most relevant of the large amount of original data sources are mentioned while a full list can be found on the OPSD platform itself [24].

3.2. Conventional power plants

Content and Scope. This package contains data on conventional power plants for Germany and selected other European countries. It covers nuclear reactors and thermal plants fuelled by lignite, hard coal, natural gas or oil, as well as hydro power and pumped hydro storage. The package focuses on individual power plants, or power plant blocks, and their technical characteristics. These include installed capacity, main energy source and type of technology. Where available, we also include capabilities with respect to combined heat and power (CHP) generation, commissioning and closure years as well as geographical information and Energy Identification Codes (EIC). We also provide information on thermal

1
2
3
4
5
6
7
8
9 efficiency of individual plants, either by means of desk research or based on an
10 estimation method described in [27].

11 The main data source for Germany is a detailed power plant list provided
12 by the federal network regulator (BNetzA [28]). It is complemented with a
13 list provided by the German Environment Agency (UBA [29]) which contains
14 additional information on CHP, installed gross capacities, and turbine types.
15 Sources for other countries include detailed power plant lists provided by TSOs,
16 ministries, or associations and market participants as well as manually compiled
17 data for countries without system-wide power plant lists.

18 While we implemented a single data structure for all countries covered, the
19 package consists of two separate Jupyter Notebooks and two separate output
20 files: one for Germany and one for the other countries. This is because the orig-
21 inal German data provides many more details compared to the other countries,
22 and also requires a specific merging activity of the lists provided by BNetzA
23 and UBA.

24
25 *Geographic coverage.* Geographically, the data package focuses on Germany and
26 its neighbouring countries. The data package currently covers Austria, Belgium,
27 the Czech Republic, Denmark, Germany, Finland, France, Italy, the Nether-
28 lands, Norway, Poland, Spain, Sweden, Switzerland, Slovakia, Slovenia, and the
29 United Kingdom. Due to varying availability and quality of original data, not
30 all information is available for each country.

31
32 *Value added.* This data package is designed to benefit electricity system mod-
33 ellers in several ways. First, we provide a single data structure for many coun-
34 tries, including a harmonisation of fuels and technologies, the latter sometimes
35 inferred from other information provided in the original data or based on own
36 assessments. Specifically for Germany, merging the two official lists provided
37 by BNetzA and UBA adds value with respect to CHP information, net / gross
38 capacities, as well as commissioning and shut-down years. Further, complemen-
39 tary information which is not included in the primary sources but researched
40 manually, i.e., geo coordinates, EIC codes, and thermal efficiency, should be of
41 use for many modellers.

42 43 44 3.3. Renewable power plants

45 *Content and Scope.* This data package contains a list of existing solar, wind,
46 bio, run-of-river, geothermal and other renewable power plants. Due to different
47 data and parameter availability in the countries, one list is provided per coun-
48 try. These are of different accuracy and partly include different parameters.
49 However, all lists state energy source and technology, electrical capacity and
50 data source. If available in the original data, the respective system operators
51 (transmission and distribution), support scheme ID or technology-specific pa-
52 rameters like hub height for wind turbines are included. While a high amount
53 of parameters and units (more than 1.8 million power plant entries) are avail-
54 able for Germany, these only cover plants eligible under the renewable support
55 scheme (EEG). Also for Switzerland just the supported plants are provided
56
57
58
59
60
61
62
63
64
65

(Swiss feed-in tariff KEV) since these are well documented and information is publicly provided. For Denmark, detailed data on solar and wind power plants including rotor diameter etc. is provided, but no other technologies. Aggregated capacity per energy source and municipality is stated in the lists for Poland and France.

Most original data sources either provide conventional or renewable power plants. However, especially with respect to hydro power, there is an overlap between sources, and harmonisation is challenging. With the objective to provide data packages that can be directly applied by the modellers, we eliminated the overlap to some extent by clarifying which energy sources and which technologies are covered by the conventional and renewable data package as illustrated in Figure 2.

Geographic coverage. Currently, lists of German, Danish, Swiss, Polish and French renewable power plants are provided. Since the location of renewable power plants is important for resulting feed-in of wind and solar, the geographic location is provided wherever information was available. This is the case for Germany, Denmark and France. These coordinates are of different accuracy depending on the original source being derived from coordinates in UTM-format (BNetzA), zip-codes (German TSOs) or districts.

Value added. The amount of renewable energy units by far exceed the conventional ones due to smaller unit size, which makes processing tedious and manual verification on a unit-by-unit base impossible. Furthermore, naming conventions and classification of renewable energy technologies and sources differ significantly between data providers. The added value of the script is thus bringing it to the same energy source structure as well as to the same formatting. Although some parameters differ between the country scripts, the country lists can be consistently combined for a subset of essential parameters. Furthermore, inconsistent data points (like e.g. rooftop solar installations of several GW) are marked and can thus easily be filtered out.

An additional output of this data package and value added are historic daily time series of the installed renewable capacities per energy source type for Germany. These are applied in the time series script: Since solar and wind feed-in time series are derived from real feed-in data, feed-in levels tend to increase over the year due to increasing installed capacity. For deriving relative, normalised hourly profiles, the corresponding installed capacity of the respective day is required.

3.4. National generation capacity

Content and Scope. This data package complements the *Conventional power plants* and *Renewable power plants* packages by providing aggregated numbers on installed capacities for all types of generators. To create this package, we compiled a broad range of sources and structured them according to the harmonised classification of energy sources and technologies provided in Figure 2.

This proved to be challenging because of very heterogeneous reporting conventions across national statistics. Overall, we provide more than 1500 annual country statistics from four international and 25 national sources, i.e. around four statistics for each country covering several years each. As for the international sources, we use data provided by ENTSO-E and EUROSTAT, the latter going back to 1990. National data sources include statistical offices, ministries, TSOs and industry associations. While the international sources are automatically processed, the national sources generally require manual compilation.

Geographic coverage. The package provides data on 39 countries and thus virtually comprises all European countries. 25 of these countries are not only covered by international sources, but also complemented by individual country statistics. For example, 2015 data for France comprises four distinct data sources.

Value added. The main output of this package is a single data file that includes installed generation capacities for all European countries, with a harmonised coverage of energy sources and generation technologies. This is valuable for all electricity system models that do not require much technological or geographic detail, for example linear models that consider whole countries as nodes. Additional value is created by contrasting different sources for a given country and year, located in adjacent columns of the data file. This enables the user, inter alia, to draw conclusions on the quality and coverage of specific sources or to infer missing values. For example, overall generation capacity installed in France in 2015 varies between 105GW and 129GW across the different sources, with particularly large variation among renewable sources.

To name another benefit, the temporal and geographical coverage of this package is much better compared to the other two data packages covering conventional and renewable generators. Not least, we also provide a strict distinction between values that are actually zero and such that are not available or not specified in national / international statistics.

3.5. Time series

Content and Scope. The *Time series* data package contains highly granular time series data for a number of variables. Included are power consumption (load), day-ahead power prices, as well as three types of time series related to generation by variable renewables (wind and solar). The wind and solar time series include historical forecast generation, actual generation and installed production capacity. The latter is calculated from the *Renewable power plants* package and is required to calculate renewable generation profiles as share of installed production capacity.

Depending on the national market set-up, the time series have a resolution of either 15 minutes (i.e. Austria, Germany and some neighbours), 30 minutes (i.e. the British Isles) or 60 minutes (all other countries).

While from 2015 onward, many of the data are available from the ENTSO-E Transparency Platform, earlier records have to be obtained through individual national sources. Data are extracted mostly from national TSO's websites,

ENTSO-E's "Monthly statistics data collection" as well as ENTSO-E Transparency.

Many original data sources report data by referring to the local time, in most cases Central European (Summer) Time (CET/CEST). Transitions between summer daylight saving time and winter time are a source of confusion, since they imply jumping an hour in March and reporting an hour twice in October. We thus consistently convert all time stamps to Coordinated Universal Time (UTC), which avoids daylight saving time.

Since power system modelling usually requires time series data to be complete for at least an entire year, data gaps of up to two hours are filled by applying linear interpolation. In these cases, tables are annotated with a marker, allowing users to trace back original data manipulation, while longer periods of missing data are left as-is.

Geographic coverage. Overall, 35 countries are covered while varying between variables and improving in more recent years. While load data is published for all European countries, availability of renewable generation data varies between countries. Historic records of day-ahead prices are freely available since 2015, when the ENTSO-E Transparency Platform was introduced.

Value added. This package relieves researchers of the burden to collect time series data from many different sources, each using different formats and ways to organise data: Some TSOs allow downloading one file containing all available data (Amprion), while others require downloading and combining monthly (TenneT) or even daily files (PSE). The time series data package combines all data with the same temporal resolution in one file each, facilitating its use for modellers. Additionally, 15 and 30 minutes data are aggregated to hourly time series since this is the resolution required by most power system models and enables one file with maximal geographic coverage while keeping the file size manageable.

3.6. Weather data

Content and Scope. The *Weather data* package provides a script to automatically access, subset, download and process wind, solar, temperature and air data based on the MERRA-2 dataset [30, 31] made available by NASA. This data package differs significantly from the other data packages as the size of the data to be downloaded can reach GB and TB if a large amount of variables, a long time span and a huge geographical area are chosen. Thus, instead of offering the download of ready-made data packages, the strategy here is to provide a documented methodological Python script that can be run on the user's computer. This script serves as a documented tutorial and enables the user to specify the geographical area and the time span for which the data shall be downloaded and processed. Only a small sample data set (Germany, 2016) is available for direct download.

Besides connection to the database, download and conversion of the three-dimensional NetCDF4-format to CSV and SQL-output, the script includes the

calculation of the output parameters relevant for electricity system modellers. For example, the wind speed is calculated from the north and east pointing wind vectors provided in the original MERRA-2 data set. Resulting parameters covered by the script are:

- wind velocity in 2, 10 and 50 meters height as well as roughness length
- solar radiation
- temperature
- air density and pressure

Geographic coverage. The MERRA-2 data set offers reanalysis data with a worldwide coverage in a resolution of $0.625^\circ \times 0.5^\circ$ which correlates to approximately 50 x 50 km in Central European latitudes. By specifying the northeastern and southwestern corner coordinates in the script, a rectangular area can be chosen.

Value added. Due to the rising influence of wind and solar feed-in for the power system, their realistic representation and thus weather parameters are crucial for electricity system models. The OPSD weather data script fills a gap between original meteorological data and ready-made feed-in time series. On the one end of the spectrum, the direct usage of original data sets like MERRA-2 with hundreds of different weather parameters to choose from as well as a tedious manual download and processing requires specific knowledge to retrieve the relevant data for electricity system modelling. Additionally, the NetCDF-file format preferred by meteorologist due to its multidimensional structure is hard to handle and not directly compatible with data formats preferred by energy system modellers. On the other end, ready-made feed-in time series as provided in the data package described in Section 3.5 are less flexible for the automatic adaption to own modelling code and are restricted by fixed assumptions made for the calculation of the resulting electricity feed-in, e.g. the hub height or type of wind turbines. Thus, for electricity system modellers who require decades of historical weather data, and who carry out own feed-in calculations, the script lowers the barrier of utilising the wealth of available reanalysis weather data.

4. Discussion and Outlook

The OPSD platform has been designed to provide a service to the electricity system modelling community. It provides a wide range of data at one place, is easily accessible, clean and ready-to-use, permanently available and version-controlled. The large number of users — around 100,000 unique visitors during the last year (2017) — and, more importantly, the amount of research that makes use of OPSD — 18 published papers [8, 32–48] since the go-live in late 2016 — suggest that the platform fulfils a need and provides value to electricity system modellers.

Clearly, there is room for improvement. We are currently working on expanding the geographic scope of the platform, adding additional data items, and improving usability. We also work on implementing Digital Object Identifiers (DOIs), which will make it easier to reference and retrieve specific versions of the individual data packages.

An ongoing challenge for projects like OPSD is the academic incentive structure. Collecting, cleaning, aggregating and providing data is essential, everyone agrees. However, it is challenging to attract funding for these kinds of projects and even more difficult to publish journal articles on data processing and aggregation. As a consequence, open data, like many other public goods, tends to remain short in supply.

In the long term, we hope that OPSD provides additional benefits by serving as a proof of concept. It may help to convince the providers of data – system operators, statistical offices, generation companies, and public authorities – to provide data that is complete and can be easily used for research. With machine-readable data and metadata that is packaged and version-controlled, along with quality and consistency checks and detailed data documentation, the electricity sector would make a huge step towards frictionless data.

Acknowledgements

The OPSD project has been supported by the German Federal Ministry for Economic Affairs and Energy (BMWi) in its first phase 2015-2017 under grant number 03MAP324. Since 2018, it is maintained and amended as part of the project "Open Source Energiewende" (060/17) for BMWi. The authors also thank all colleagues contributing to the platform, the open energy modelling initiative and all users of OPSD providing feedback.

References

- [1] S. Pfenninger, A. Hawkes, J. Keirstead, Energy systems modeling for twenty-first century energy challenges, *Renewable and Sustainable Energy Reviews* 33 (2014) 74–86. doi:10.1016/j.rser.2014.02.003.
- [2] J. DeCarolís, H. Daly, P. Dodds, I. Keppo, F. Li, W. McDowall, S. Pye, N. Strachan, E. Trutnevyte, W. Usher, M. Winning, S. Yeh, M. Zeyringer, Formalizing best practice for energy system optimization modelling, *Applied Energy* 194 (2017) 184–198. doi:10.1016/j.apenergy.2017.03.001.
- [3] D. Connolly, H. Lund, B. Mathiesen, M. Leahy, A review of computer tools for analysing the integration of renewable energy into various energy systems, *Applied Energy* 87 (4) (2010) 1059–1082. doi:10.1016/j.apenergy.2009.09.026.

- [4] L. M. Hall, A. R. Buckley, A review of energy systems models in the UK: Prevalent usage and categorisation, *Applied Energy* 169 (2016) 607–628. doi:10.1016/j.apenergy.2016.02.044.
- [5] R. Fernandez Blanco Carramolino, F. Careri, K. Kavvadias, I. Hidalgo Gonzalez, A. Zucker, E. Peteves, Systematic mapping of power system models, Tech. rep. (2017). doi:10.2760/422399.
- [6] L. Hirth, What caused the drop in European electricity prices? A factor decomposition analysis, *The Energy Journal* 39 (1) (2018) 143–157. doi:10.5547/01956574.39.1.lhir.
- [7] F. Kunz, Improving Congestion Management: How to Facilitate the Integration of Renewable Generation in Germany, *The Energy Journal* 34 (4) (2013) 55–78. doi:10.5547/01956574.34.4.4.
- [8] W.-P. Schill, M. Pahle, C. Gambardella, Start-up costs of thermal power plants in markets with increasing shares of variable renewable generation, *Nature Energy* 2 (17050) (2017) . doi:10.1038/nenergy.2017.50.
- [9] W.-P. Schill, A. Zerrahn, Long-run power storage requirements for high shares of renewables: Results and sensitivities, *Renewable and Sustainable Energy Reviews* 83 (2018) 156–171. doi:10.1016/j.rser.2017.05.205.
- [10] J. Egerer, W.-P. Schill, Power system transformation toward renewables: Investment scenarios for Germany, *Economics of Energy & Environmental Policy* 3 (2) (2014) 29–44. doi:10.5547/2160-5890.3.2.jege.
- [11] I. Schlecht, H. Weigt, Linking Europe: The Role of the Swiss Electricity Transmission Grid until 2050, *Swiss Journal of Economics and Statistics* 151 (2) (2015) 125–165. doi:10.1007/BF03399415.
- [12] F. Kunz, Quo Vadis? (Un)scheduled electricity flows under market splitting and network extension in central Europe, *Energy Policy* 116 (2018) 198–209. doi:10.1016/j.enpol.2018.01.051.
- [13] J. Egerer, C. Gerbaulet, C. Lorenz, European electricity grid infrastructure expansion in a 2050 context, *The Energy Journal* 37 (1) (2016) 101–124. doi:10.5547/01956574.37.SI3.jege.
- [14] C. Gerbaulet, A. Weber, When regulators do not agree: Are merchant interconnectors an option? Insights from an analysis of options for network expansion in the Baltic Sea region, *Energy Policy* 117 (2018) 228–246. doi:10.1016/j.enpol.2018.02.016.
- [15] W.-P. Schill, C. Gerbaulet, Power system impacts of electric vehicles in Germany: Charging with coal or renewables?, *Applied Energy* 156 (2015) 185–196. doi:10.1016/j.apenergy.2015.07.012.

- [16] F. Wiese, G. Bökenkamp, C. Wingenbach, O. Hohmeyer, An open source energy system simulation model as an instrument for public participation in the development of strategies for a sustainable future, *WIREs Energy Environ* 3 (5) (2014) 490–504. doi:10.1002/wene.109.
- [17] S. Pfenninger, J. DeCarolus, L. Hirth, S. Quoilin, I. Staffell, The importance of open data and software: Is energy research lagging behind?, *Energy Policy* 101 (2017) 211–215. doi:10.1016/j.enpol.2016.11.046.
- [18] S. Pfenninger, L. Hirth, I. Schlecht, E. Schmid, F. Wiese, T. Brown, C. Davis, M. Gidden, H. Heinrichs, C. Heuberger, S. Hilpert, U. Krien, C. Matke, A. Nebel, R. Morrison, B. Müller, G. Pleßmann, M. Reeg, J. C. Richstein, A. Shivakumar, I. Staffell, T. Tröndle, C. Wingenbach, Opening the black box of energy modelling: Strategies and lessons learned, *Energy Strategy Reviews* 19 (2018) 63–71. doi:10.1016/j.esr.2017.12.002.
- [19] F. Wiese, S. Hilpert, C. Kaldemeyer, G. Pleßmann, A qualitative evaluation approach for energy system modelling frameworks, *Energy, Sustainability and Society* 8 (1) (2018) . doi:10.1186/s13705-018-0154-3.
- [20] K.-K. Cao, F. Cebulla, J. J. Gómez Vilchez, B. Mousavi, S. Prehofer, Raising awareness in model-based energy scenario studies – a transparency checklist, *Energy, Sustainability and Society* 6 (28) (2016) . doi:10.1186/s13705-016-0090-z.
- [21] C. Dieckhoff, A. Eberspächer, Consulting with energy scenarios - Requirements for scientific policy advice, Accessed 24 June 2016 (2016). URL http://www.akademienunion.de/fileadmin/redaktion/user_upload/Publikationen/Stellungnahmen/Stellungnahme_Energy_scenarios.pdf
- [22] ENTSO-E Transparency Platform. URL <https://transparency.entsoe.eu>
- [23] L. Hirth, J. Mühlenpfordt, M. Bulkeley, The ENTSO-E Transparency Platform - A review of Europes most ambitious electricity data platform, *Applied Energy* 225 (2018) 1054–1067. doi:10.1016/j.apenergy.2018.04.048.
- [24] Open Power System Data - A free and open data platform for power system modelling. URL <https://open-power-system-data.org/>
- [25] Open Knowledge International, Frictionless Data Specifications and Software: Data Package. URL <https://frictionlessdata.io/specs/data-package/>
- [26] Open Power System Data - Repositories on github. URL <https://github.com/Open-Power-System-Data>

- [27] J. Egerer, C. Gerbaulet, R. Ihlenburg, F. Kunz, B. Reinhard, C. von Hirschhausen, A. Weber, J. Weibezahn, Electricity sector data for policy-relevant modeling: Data documentation and applications to the German and European electricity markets, Tech. rep., Data Documentation, DIW (2014).
URL http://www.diw.de/documents/publikationen/73/diw_01.c.440963.de/diw_datadoc_2014-072.pdf
- [28] Bundesnetzagentur (German Federal Network Agency, Power Plant List (in German).
URL https://www.bundesnetzagentur.de/EN/Areas/Energy/Companies/SecurityOfSupply/GeneratingCapacity/PowerPlantList/PubliPowerPlantList_node.html
- [29] Umweltbundesamt (German Federal Environment Agency, Database "Power plants in Germany" (in German, Datenbank "Kraftwerke in Deutschland").
URL <https://www.umweltbundesamt.de/dokument/datenbank-kraftwerke-in-deutschland>
- [30] Global Modeling and Assimilation Office (GMAO) at NASA Goddard Space Flight Center, Modern-Era Retrospective Analysis for Research and Applications, version 2.
URL <https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/>
- [31] M. M. Rienecker, M. J. Suarez, R. Gelaro, R. Todling, J. Bacmeister, E. Liu, M. G. Bosilovich, S. D. Schubert, L. Takacs, G.-K. Kim, S. Bloom, J. Chen, D. Collins, A. Conaty, A. da Silva, W. Gu, J. Joiner, R. D. Koster, R. Lucchesi, A. Molod, T. Owens, S. Pawson, P. Pegion, C. R. Redder, R. Reichle, F. R. Robertson, A. G. Ruddick, M. Sienkiewicz, J. Woollen, MERRA: NASAs Modern-Era Retrospective Analysis for Research and Applications, *Journal of Climate* 24 (14) (2011) 3624–3648. doi:10.1175/JCLI-D-11-00015.1.
- [32] M. Kendzioriski, M. Setje-Eilers, F. Kunz, Generation expansion planning under uncertainty: An application of stochastic methods to the German electricity system, in: 2017 14th International Conference on the European Energy Market (EEM), 2017, pp. 1–7. doi:10.1109/EEM.2017.7981891.
- [33] M. Klein, M. Deissenroth, When do households invest in solar photovoltaics? An application of prospect theory, *Energy Policy* 109 (2017) 270–278. doi:10.1016/j.enpol.2017.06.067.
- [34] J. Amme, G. Pleßmann, J. Bühler, L. Hülk, E. Kötter, P. Schwaegerl, The ego grid model: An open-source and open-data based synthetic medium-voltage grid model for distribution power supply systems, *Journal of Physics: Conference Series* 977 (1) (2018) 012007. doi:10.1088/1742-6596/977/1/012007.

- [35] U. P. Müller, I. Cusmann, C. Wingenbach, J. Wendiggensen, AC Power Flow Simulations within an Open Data Model of a High Voltage Grid, in: V. Wohlgemuth, F. Fuchs-Kittowski, J. Wittmann (Eds.), *Advances and New Trends in Environmental Informatics*, Springer International Publishing, Cham, 2017, pp. 181–193.
- [36] U. P. Müller, L. Wienholt, D. Kleinhans, I. Cusmann, W.-D. Bunke, G. Pleßmann, J. Wendiggensen, The ego grid model: An open source approach towards a model of german high and extra-high voltage power grids, *Journal of Physics: Conference Series* 977 (1) (2018) 012003. doi:10.1088/1742-6596/977/1/012003.
- [37] F. Fusco, S. Tirupathi, R. Gormally, Power systems data fusion based on belief propagation, in: *2017 7th IEEE International Conference on Innovative Smart Grid Technologies (ISGT)*, 2017. URL <https://arxiv.org/pdf/1705.08815.pdf>
- [38] P. Tafarte, P. Buck, Integration of wind power - 2014; challenges and options for market integration and its impact on future cross-sectorial use, in: *2017 14th International Conference on the European Energy Market (EEM)*, 2017, pp. 1–5. doi:10.1109/EEM.2017.7981941.
- [39] N. H. El-Amari, A. Balbaa, R. A. Swief, T. S. Abdel-Salam, A Reconfigured Whale Optimization Technique (RWOT) for Renewable Electrical Energy Optimal Scheduling Impact on Sustainable Development Applied to Damietta Seaport, Egypt, *Energies* 11 (3) (2018) . doi:10.3390/en11030535.
- [40] J. P. Gonzalez, A. M. S. Roque, E. A. Prez, Forecasting functional time series with a new hilbertian armax model: Application to electricity price forecasting, *IEEE Transactions on Power Systems* 33 (1) (2018) 545–556. doi:10.1109/TPWRS.2017.2700287.
- [41] Y. Yaslan, B. Bican, Empirical mode decomposition based denoising method with support vector regression for time series prediction: A case study for electricity load forecasting, *Measurement* 103 (2017) 52–61. doi:10.1016/j.measurement.2017.02.007.
- [42] B. H. Zaidi, S. H. Hong, Combinatorial double auctions for multiple microgrid trading, *Electrical Engineering* 100 (2) (2018) 1069–1083. doi:10.1007/s00202-017-0570-y.
- [43] F. Gotzens, H. Heinrichs, J.-F. Hake, H.-J. Allelein, The influence of continued reductions in renewable energy cost on the European electricity system, *Energy Strategy Reviews* 21 (2018) 71–81. doi:10.1016/j.esr.2018.04.007.
- [44] W.-P. Schill, A. Zerrahn, F. Kunz, Prosumage of Solar Electricity. Pros, Cons, and the System Perspective, *Economics of Energy & Environmental Policy* 6 (1) (2017) 7–31. doi:10.5547/2160-5890.6.1.wsch.

- [45] A. Zerrahn, W.-P. Schill, C. Kemfert, On the economics of electrical storage for variable renewable energy sources, preprint (2018) .
 URL <https://arxiv.org/abs/1802.07885>
- [46] M. Robinius, J. Linßen, T. Grube, M. Reuß, P. Stenzel, K. Syranidis, P. Kuckertz, D. Stolten, Comparative Analysis of Infrastructures: Hydrogen Fueling and Electric Charging of Vehicles, Vol. 408 of Schriften des Forschungszentrums Jülich Reihe Energie & Umwelt / Energy & Environment, Forschungszentrum Jülich GmbH Zentralbibliothek, Jülich, 2018.
 URL <http://juser.fz-juelich.de/record/842477>
- [47] L. Nacken, T. Möbius, The effects of harmonized European climate policy targets in comparison to national targets utilizing a European electricity market model, in: 2017 14th International Conference on the European Energy Market (EEM), 2017, pp. 1–5. doi:10.1109/EEM.2017.7981987.
- [48] J. Olauson, ERA5: The new champion of wind power modelling?, Renewable Energy 126 (2018) 322–331. doi:10.1016/j.renene.2018.03.056.

Prof. Jinyue Yan
Editor-in-Chief
Applied Energy

June 1, 2018

Dear Prof. Jinyue Yan,

We are pleased to submit an article entitled “Open Power System Data - Frictionless data for electricity system modelling” for consideration for publication in the *Special Issue of Applied Energy: Data on Energy, Climate change and Sustainability*.

This article describes the idea, content and benefit of the Open Power System Data platform that aims at realising the efficiency and quality gains of centralised data provision by collecting, checking, processing, aggregating, documenting and publishing data required by electricity system modellers.

We think that this manuscript is appropriate for publication by the Special Issue of Applied Energy since it supports the proceedings and analysis on the data that is required for electricity system analysis and thus contributes to its application in quantitative energy system analysis and its sustainable development.

Applied energy research as well as energy and climate policy advice are often based on quantitative computer models. Thus, quality of electricity system modelling heavily relies on input data, while being underrepresented in the academic discussion. The approach presented in the paper has the potential to significantly increase the quality of input data and the efficiency of data processing for electricity system modelling, thus contributing to quality improvements and more time for the system analysis itself.

The novelty of our work is the approach of providing processed ready-to-use data for electricity system modelling in a central manner, while using a decentralised lightweight, scalable technical architecture and keeping it fully transparent and flexible by offering the possibility to adapt the processing scripts. The main principles in this regard - open source scripts, transparency of used input data and documentation of both the script as well as the final output – comply with scientific standards.

This work has not been submitted previously to Applied Energy (in part or in whole), has not been published previously and is not under consideration for publication elsewhere. All authors approved the article.

If you feel that the manuscript is appropriate for your journal, we suggest the following reviewers:

- Sylvain Quolin; Contact: sylvain.quotin@kuleuven.be
- Francesco Gardumi; Contact: francesco.gardumi@desa.kth.se
- Mark Howells; Contact: mark.howells@energy.kth.se
- Chris Davis; Contact: c.b.davis@rug.nl
- Manuel Wetzel; Contact: manuel.wetzel@dlr.de

Thank you for your consideration.

Sincerely,

Ingmar Schlecht (Corresponding Author) and Frauke Wiese (1st Author) for all the authors

LaTeX Source Files

[Click here to download LaTeX Source Files: LatexSourceFiles.zip](#)