WORLD RESOURCES INSTITUTE

TECHNICAL NOTE

A GLOBAL DATABASE OF POWER PLANTS

LOGAN BYERS, JOHANNES FRIEDRICH, ROMAN HENNIG, AARON KRESSIG, XINYUE LI, COLIN MCCORMICK, LAURA MALAGUZZI VALERI

ABSTRACT

This technical note explains how World Resources Institute (WRI) experts and their partners created the Global Power Plant Database from official government data and independent sources around the world, integrated them with crowdsourced data such as analysis of satellite images, and delivered the final database as an open data resource. The note explains how the experts addressed three challenges: matching plants across databases, keeping the database up to date, and delivering information on the accuracy of the data.

1. INTRODUCTION

An affordable, reliable, and environmentally sustainable power sector is central to modern society. Governments, utilities, and companies make decisions that both affect and depend on the power sector. For example, if governments apply a carbon price to electricity generation, it changes how plants run and which plants are built over time. On the other hand, each new plant affects the electricity generation mix, the reliability of the system, and system emissions. Plants also have significant impact on climate change, through carbon dioxide (CO_2) emissions; on water stress, through water withdrawal and consumption; and on air quality, through sulfur oxides (SO_x), nitrogen oxides (NO_x), and particulate matter (PM) emissions.

CONTENTS

Technical notes document the research or analytical methodology underpinning a publication, interactive application, or tool.
This is an update to the original technical note published in April 2018, to reflect the addition of new data that have improved the global power plant database's coverage.

Suggested Citation: L. Byers, J. Friedrich, R. Hennig, A. Kressig, Li X., C. McCormick, and L. Malaguzzi Valeri. 2019. "A Global Database of Power Plants." Washington, DC: World Resources Institute. Available online at www.wri.org/publication/global-database-power-plants.

Despite the importance of the power sector, there is no global, open-access database of power plants. Existing databases fail to be either truly comprehensive or fully open. Many countries do not report their power sector data at the plant level, and those that do vary wildly in what they report, how they report it, and how frequently they report. The lack of reporting standards makes data gathering time intensive, as the data are in different formats and must be harmonized. This creates a barrier for conducting global and national research and analysis of the power sector.

The Global Power Plant Database leverages existing data sources and methodologies to build a comprehensive and open-access power sector database.[1] The database collects the following characteristics and indicators:

- All types of fuel
- Technical characteristics (fuel, technology, ownership)
- Operational characteristics (generation)
- Plants' geolocation
- Plants over 1 megawatt (MW)
- Plants in operation only (in first iteration)

We refer to this database as open, as all data are traceable to sources that are publicly available on websites. Most of the publicly available sources we draw upon are collected from national governments and other official sources. In addition, the database is published under a Creative Commons—Attribution 4.0 International license (CC BY 4.0), allowing it to be used and republished in any fashion, with source attribution. By providing a common information source, the database will facilitate collaborative analysis of the power sector. It is important to note that although the database is the most comprehensive in terms of fuel types and capacity covered, because power sector information is not fully reported or instantly updated, the database will never be fully comprehensive and will show the power sector data with some time delay.

The Global Power Plant Database builds on experiences from existing, similar efforts (see Table 1). It creates a new database that is easier to update because it collects data on a country-by-country basis and uses as many nationally reported and automatically updating sources as possible, focusing on primary sources or entities with legal authority (see Section 3.1 on data reliability). It builds on the methods employed by the sources provided in Table 1 and uses these sources' latitude and longitude information for some plants (see Section 5.2). The specific methods used to create the database are fully detailed in this technical note.

The Global Power Plant Database uses more than 600 sources to create a database in which more than 95 percent of the installed capacity information is from 2015 or later. A detailed description of the sources and years of the plant-level information, by country, is documented online and available upon request.

The aim is to create as comprehensive a database as possible. For the initial release of the database, only generators above 1 MW nameplate capacity will be

Table 1 | **Existing Open Global Power Plant Datasets and Their Limitations**

DATABASE	LIMITATIONS	URL
Global Energy Observatory (GEO)[a]	Only 10,000 power plants representing 58% of global capacity; opportunistically updated from public records	http://globalenergyobservatory.org
Enipedia[b]	Not accurately geolocated or regularly updated	http://enipedia.tudelft.nl/wiki/Enipedia
Carbon Monitoring for Action (CARMA)[c]	Not accurately geolocated; not updated since 2012; only covers a small set of indicators	http://carma.org/plant

Notes:
[a] Global Energy Observatory. Los Alamos National Lab. http://globalenergyobservatory.org.
[b] Enipedia. University of Delft, The Netherlands. https://enipedia.org.
[c] Carbon Monitoring for Action (CARMA). Global Center for Development. http://carma.org.
Source: WRI, April 2018.

included, independent of their fuel source and whether they are connected to the electricity grid. We intend to integrate information on smaller generators as data becomes available over time.

The database is built preferentially on data reported from trusted sources, as defined in Section 3. Section 2 discusses data limitations, and Section 5 reports on how the database is built to minimize maintenance costs and facilitate updates. Section 4 describes the indicators included in the database for each power plant and their coverage within the dataset. Some characteristics are rarely available from public sources, such as geolocation and annual electricity generated. The database adds plant geolocation, as described in Section 5. Section 6 outlines how annual electricity generation is estimated for individual plants when it is not reported by official sources. Section 7 details how the database is updated and maintained over time, and Section 8 discusses the next steps in the database's development.

2. LIMITATIONS

Many challenges and limitations arise when collecting data from a variety of different sources and when there is a lack of officially reported data. Although data in different languages can slow the collection process, we would consider this a challenge and not a limitation. With an international staff and translation services available through the web, we have been able to collect data in many languages. The primary limitations are listed below:

1. Data availability (of small and renewable power plants)
2. Reporting delay
3. Data reliability
4. Lack of operational data reported (electricity generation, emissions, water use)

Data availability is the primary limitation. Because most countries do not publicly report their power sector data, it is not feasible to assemble 100 percent power plant coverage. It is particularly difficult to identify the smaller, distributed power plants, a category that includes smaller renewables and diesel generators.

Renewable power plants are not always reported in public documentation as they are relatively new and smaller, causing wind and solar plants to have the lowest global coverage of installed capacity in the database (49 percent and 21 percent, respectively). Conventional power plants, including thermal and large hydro plants, are more extensively covered, as reported in Table 2.

Table 2 | **Database Coverage by Fuel Type**

FUEL TYPE	DATA COVERAGE/GLOBAL INSTALLED CAPACITY (%)	MISSING CAPACITY IN DATABASE (MW)[a]	SOURCE OF TOTAL CAPACITY, YEAR[b]
Nuclear	100.00%	0	International Atomic Energy Agency (IAEA), 2017
Geothermal	97.19%	362	Platts, 2017
Coal	>100%	0	Platts, 2017[c]
Hydro	89.34%	125,286	Platts, 2017[c]
Natural Gas/Oil	81.55%	392,452	Platts, 2017[c]
Biomass	76.38%	10,651	Platts, 2017[c]
Wind	49.48%	245,211	Global Wind Energy Council (GWEC), 2016
Solar	21.03%	239,237	International Energy Agency (IEA), 2016

Notes: Coverage statistics as of May, 2019.
[a] Missing capacity refers to the difference between the sum of capacity of plants (by fuel) in the database and the total capacity reported in listed sources. Capacity of plants that are not geolocated is included in missing.
[b] Total capacity reported in this column is used exclusively to calculate database coverage and MW of missing plant by fuel (columns 2 and 3).
[c] Platts figures are unofficial estimates but are preferred over IEA official statistics, which are updated with a two- or three-year lag.
Sources: IAEA data accessed in 2017 at https://www.iaea.org/pris/; Platts 2017; GWEC 2016; IEA-PVPS 2016.

Data availability also affects the amount of installed capacity that is geolocated and the accuracy of that location. As described above, most of the plants in the database are geolocated by matching them to other databases, which themselves have varying levels of accuracy. The collaboration with KTH Royal Institute of Technology in Stockholm improved the accuracy of geolocation by systematically verifying each location for Latin American and African plants. We have yet to develop a platform that can help collect and verify crowdsourced data provided by any private citizen.

The second and third challenges (data reliability and data updatability) impact the quality of data that is provided through the database and the frequency of updates. Both of these concerns are addressed in the data collection strategy (see Sections 3.1 and 3.2, respectively). How frequently the database is updated depends on when the source data is updated. As we explain in Section 3.2, we use the most authoritative source of information, which means we do not have more reliable sources to validate it against.

The final challenge with the database is that operational data, such as electricity generation by power plant, is rarely reported and, therefore, must be estimated. Generation describes the actual electricity produced by a plant on a yearly basis. Generation can vary significantly from year to year. Section 6 explains how generation is estimated and the challenges of defining and accurately reporting estimations. Generation data is an important input in estimating other power plant variables, such as power plant water use and emissions (CO_2, PM, SO_x, etc.).

3. DATA COLLECTION CRITERIA AND SOURCES

The database is built entirely from open sources, which are publicly available on the Internet, including data from national government agencies, reports from companies that build power plants or provide their components, data from public utilities, and information from multinational organizations. One goal of the data collection process is to identify reliable sources that are regularly updated and sources of data that are unavailable elsewhere.

Data collection occurs on a country-by-country basis, with each country requiring different strategies to integrate data. Ideally, we use sources that are comprehensive and can be integrated automatically (through application programming interfaces, or APIs), but this is not usually possible. For many countries, data collection requires manually gathering data from various sources. This section describes the data collection criteria (reliability and ease of updating) and which data sources were used.

3.1 Data Reliability

The reliability of this database depends on the reliability of the data sources we identify and aggregate. Sources that are directly linked to power plant operations or have the legal authority to gather power plant statistics are considered the most reliable. The quality of data sources is broadly ranked as follows:

1. Primary sources or entities with legal authority

Definition

- Data provider is a primary source (i.e., not aggregated from other sources)
- Source has the authority to collect statistics directly from plant-operating entities

Criteria

- Government agency reporting on a power plant within its boundaries; company reporting on a power plant it owns or operates; construction or manufacturing company reporting on a power plant it has serviced

Examples

- National governments
- Utilities
- Transmission or system operators
- Power plant construction companies
- Power plant operators
- Intergovernmental organizations

2. Secondary sources without legal authority but with quality assurance processes

Definition

- Data provider is a secondary or indirect source (e.g., aggregated data source)

Criteria

- Data provider has a system or process for cross-checking/verifying data claims
- Data are traceable to a source, although connecting a specific data point to a specific source may be difficult

Examples

- Global Energy Observatory
- IndustryAbout.com
- CARMA

3. Crowdsourced data

Definition

- Information by data provider is crowdsourced

Criteria

- Crowdsourced data must be linked directly to a source unless it is geolocated
- Crowdsourced data must then be cross-verified by WRI staff

Examples

- Wikipedia
- KTH crowdsourcing (see Section 3.2 for more details)

For primary and secondary sources, data must meet the definitions and criteria listed above. The same rules apply to crowdsourced data, although the verification stage makes the data collection process for crowdsourced data more time-consuming. The database uses the most reliable data available for each power plant observation. Although official data is preferred because it is most authoritative, we cannot guarantee it is completely accurate. In general, we do not have access to alternative and equally credible sources of data and cannot verify the accuracy of official data. However, we are able to verify plants' geolocation through satellite imagery and have conducted random sampling, which lets us estimate how reliably our data is geolocated (see Section 5.2).

The data sources can vary for each plant characteristic or indicator (shown in Table 3). For installed capacity, the database uses type 1 sources (primary sources or entities with legal authorities) for 67 percent of the data and type 2 sources (secondary sources) for 33 percent of the data. We use crowdsourcing to expand the number of characteristics that are covered for each plant. At this point we have not developed a formal process for contributing data or submitting corrections through crowdsourcing, although we have accepted some corrections from researchers who have used the beta version; we have also granted KTH access to our data for some manual additions and edits. At a later stage we will revisit a more robust crowdsourcing feature.

3.2 Ease of Updating the Data

The second criterion used to evaluate a data source is how easy it is to update the data. The database favors sources that provide automatic or mandated updates at set intervals to ensure it reflects the most recent information. Static or non-updating sources are avoided if an alternative exists.

Preferred sources

- Update information in a regular, timely fashion
- Data format is easy to read and load into the database

Examples

- Easiest to update: an API with machine-readable data maintained by a national government that updates annually
- Easy to update: a transmission operator that produces an annual spreadsheet of data
- Difficult to update: a utility website with information about a power plant in paragraph form

3.3 Data Sources

Although much of the data comes from information sources that report on a large number of power plants (such as the U.S. Energy Information Agency, Arab Union of Electricity, and European Network of Transmission System Operators), the majority of sources contained in the database provide only a small number of observations. More than 600 unique sources are used for the database that range in coverage from a single power plant to several thousand. Different characteristics of a power plant can be linked to different sources. Due to space limitations, we are not listing all the original sources in this technical note. Each power plant entry contains a direct link to where the data was obtained so users can check the original data source. We have documented the details of all sources of power plant characteristics by country in separate source documentation, which is available upon request. It is easy to track the data source within the database because every power plant entry is directly connected to its source(s). We also include the year associated with the capacity data and the electricity generation information.

The only information that is not linked to a source is geolocation (latitude and longitude), which users can verify independently through satellite imagery.[2] In some cases, power plant location is collected from web map providers such as Google Maps, so it is not always attributable to a unique data source. Geolocation information in the database is determined by a few methods, including parsing datasets produced by standard sources, matching plants to other global datasets containing geolocation data (detailed in the next section), using manual verification via satellite or aerial imagery (see Section 5.2).

4. COVERAGE AND INDICATORS

As of February 2018, the database included the following:

- Approximately 28,700 geolocated plants
- 80.2 percent of global installed capacity, compared to the most up-to-date country estimates[3]
- 600+ sources of data
- 164 countries

The final database includes only geolocated power plants. We have collected information on an additional 4,815 power plants that represent 337 gigawatts (GWs), or 5.2 percent of total installed capacity, that we have not yet geolocated. Including these plants would increase the database coverage to the following:

- Approximately 34,700 plants
- 85.4 percent of global installed capacity, compared to the most up-to-date country estimates[4]

WRI has partnered with the KTH Royal Institute of Technology to increase its geolocation accuracy and verify additional power plant characteristics through crowdsourcing efforts. The KTH partnership targets power plants in Latin America and Africa. The data collection conducted by KTH includes both satellite verification of plant traits and desk research to find missing data on the web. The main indicators that will be updated are those that can be verified through satellite imagery, such as location or cooling technology for thermal plants. Plant ownership will also be targeted in this effort. Improving coverage of power plant ownership over time is a workstream being developed in partnership with the Asset Data Initiative (ADI) led by Oxford University.

Table 3 lists the indicators available in the initial nonpublic beta release and their coverage.

Table 3 | **Indicator Coverage for Geolocated Plants**

INDICATOR	DESCRIPTION	PERCENT OF PLANTS WITH INDICATOR
Name	**Power plant name**	**100%**
Fuel type	fuel category	100%
Capacity	installed electrical capacity (MW)	100%
Location	latitude and longitude (xx.xx, xx.xx)	100% (by definition)
Year of capacity	year of reported capacity	100%
Year of generation	year of reported generation	100%
Data source	source of data	100%
URL	URL linking directly to data source	100%
Annual generation	annual generation (calendar year) in gigawatt hours (GWhs), gross	100%: 24% reported, 76% estimated (see Section 6)
Operational status	commissioned/retired/planned	100%
Generator technology	technology used to generate electricity	64%
Owner	primary owner of the power plant	60%
Commissioning year	first year plant generated electricity	45%

Note: Annual generation represents the only operational indicator in the first version of the database. More operational indicators will be included in future versions.
Source: WRI, 2018.

5. MATCHING AND DATABASE DESIGN

5.1 Building the Database

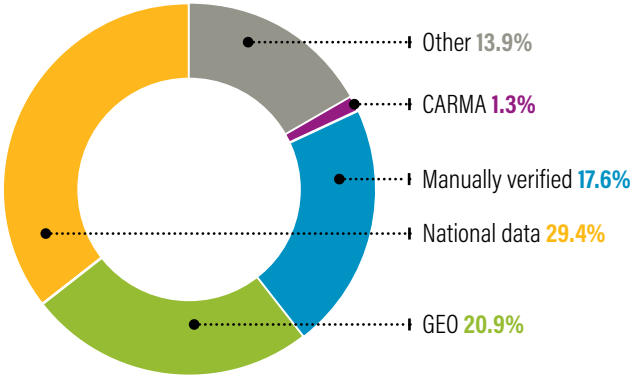
The database is built in several steps. First, data are collected from a wide variety of sources, most of which are country specific; cleaned; and converted into a standard format.

All numerical data collected from the sources are inserted in the database as is. String elements of the database, such as name of power plants or owner names, are often not reported in American English. Therefore, text fields are processed into Unicode and then certain characters are removed or replaced: control characters such as "new-line" and "non-breaking-space" are removed, unconverted characters indicating improper encoding are removed, commas are replaced with a space, and whitespace at the end of the text string is removed.

Data that are collected manually are stored in a set of publicly accessible Google Fusion Tables. When power plant data are available in computer-readable formats—such as Excel spreadsheets, comma-separated value (CSV) files, Keyhole Markup Language (KML) files, or scrapable Hypertext Markup Language (HTML) tables—these sources are automatically read and converted to the database's format using Python scripts. All coding associated with the project is available at the Global Power Plant Database.[5]

When the country-specific Fusion Tables are finalized for a specific database version, a master Python script aggregates the country-level data into a master database. If the country-level sources do not include plant geolocation, the script incorporates geolocation information from external databases, primarily GEO and CARMA. For a small number of countries with a small total capacity and/or number of power plants, the script incorporates all plant-level data directly from GEO. Raw data sources for some countries contain data of all capacities, but we exclude plants below 1 MW of capacity from the first version of the final database.[6] This is largely for consistency and simplicity; although sub–1 MW power plants account for a very small amount of capacity (around 0.5 percent of global capacity), our internal sources have more than 100,000 "plants" of this size. Submegawatt power plants are difficult to geolocate/verify through publicly available satellite imagery (Google Earth) due to their small size, and if included, they would constitute the majority of "plants" in the database despite the fact that they play a small role in terms of capacity and generation. Smaller plants are likely to be included in future iterations of the database.

Figure 1 | **Percent of Geolocated Capacity by Source, as of May 2019**



Note: Figures in table are a percentage of located capacity as a percentage of the total database. Figures are in gigawatts; there was a total of 5,572 GW in the database as of May 2019.
Source: WRI, June 2018.

The database design relies on several resource files, including a concordance table, which matches plant names and ID numbers used in different databases; a taxonomy of fuel types with a set of synonyms for each fuel category (including across languages); and a list of synonyms for the names of countries (e.g., "Burma"/"Myanmar") based on different naming conventions used in the source databases. The database standardizes all country names to ISO 3166.[7]

The logic underpinning the build process is to make the steps for collecting and processing data for each country as independent of the others as possible. Information from all countries is combined into the master database only in the final build stage. This allows current and future researchers to work in parallel on improving data for different countries, with minimal mutual dependencies.

The design logic is also built on the hybrid approach of combining manually collected data hosted in Fusion Tables with automated scripts. Automating data collection is preferable for many reasons but often is not possible for specific countries, given the lack of authoritative and comprehensive central datasets or for other complicating factors. Using Fusion Tables allows researchers with limited or no Python experience to contribute to the project by manually collecting and cleaning data.

5.2 Matching Data Sources to Determine Plant Location

All power plants in the output database are geolocated. This requires a significant effort because most geolocation information is not directly available from the original data sources. As shown in Figure 1, only 29.4 percent of capacity in the database is geolocated using national data. The rest of the capacity is located from various secondary sources or manually identified via satellite imagery. Percentages in Figure 1 are the capacity from each source type/total capacity of the database.

To use the geolocation information in GEO and CARMA, plants collected for this project are matched against plants in those databases using a data service based on Elastic search provided by Enipedia (Davis et al. 2015). This tool uses matching parameters specific to each database to return a best match, with an associated match quality score. The matching parameters are as follows:

- GEO: name (weighted 2x), country (weighted 1x), fuel type (weighted 1x)
- CARMA: name (weighted 2x), country (weighted 1x)[8]

Geolocation information reported in GEO is highly accurate. We randomly tested 50 plants and found that GEO's geolocation was within 300 meters of the actual plant 92 percent of the time. GEO reports location accuracy for 42 percent of the power plants in the database, but we found that all categories ("Approximate," "Exact," and "Unspecified") are geolocated with high accuracy and, thus, accept geolocation from GEO regardless of the category in the location accuracy field. Because of the high confidence in its location accuracy, plants in the database are first analyzed for matches to GEO using the Enipedia API for elastic search. In total, 1,797 plants representing 20.9 percent of capacity in the database have been matched and added to the final database. Those matches deemed to have a high confidence score (as provided by the Enipedia data service) have been added; a manual inspection of a few of these plants found them to be consistently accurate so, therefore, these are considered valid matches. Plants matched with medium confidence are manually verified; approximately half are ultimately judged to be valid matches. Of the remaining plants, very few are successfully matched with the Enipedia API.[9]

The plants still without geolocation after matching with GEO are next analyzed using the Enipedia API for matches with plants in CARMA. CARMA geolocates the plants less accurately. We randomly tested 50 plants and found that CARMA's geolocations are within 300 meters of the actual plant only 10 percent of the time. Exactly 1,295 plants representing 1.3 percent of the installed capacity in the database have been matched and added to the final database through the Enipedia API. Like with GEO, these matches all have been manually verified.

2,870 power plants, representing 17.6% of the total database, have been manually located via satellite imagery.

One complicating factor for the matching process is the use of non-Romanized characters in plant names. Different databases have different protocols for handling these characters, ranging from fully incorporating them using Unicode to fully Romanizing them (stripping out all non-Latin characters and/or accent marks). This is partly addressed using Python libraries to Romanize text (particularly Unidecode)[10] but remains a challenge for further matching.

All matches deemed valid (i.e., ones that refer to the same physical plant in different databases) are added to a concordance table linking ID numbers across databases, which serves as a master record linkage to connect geolocation data to plant records during the database building process. The matching procedures described here are ultimately intended to create this ID concordance table. After that has been created, the matching algorithms do not have to be rerun.

Table 4 | **Accuracy of Geolocation, by Capacity Grouping, for 1% Random Sample**

CAPACITY (MW) SIZE	TOTAL CAPACITY IN SAMPLE (MW)	NUMBER OF PLANTS IN SAMPLE	PERCENT OF CAPACITY CONFIRMED
1 to 5	168	70	86%
5 to 25	793	68	72%
25 to 100	2,520	45	90%
100 to 500	10,242	45	91%
500 to 1,000	14,944	21	95%
1,000+	20,650	12	100%

Source: WRI, April 2018.

We studied a random sample of 1 percent of the database (n = 262) to determine overall geolocation accuracy. We confirm the geolocation of a plant when the coordinates fall within the power plant campus or within 250 meters, making the plant easy to locate. The geolocation is unconfirmed if the coordinates are inaccurate or the imagery quality is poor. The geolocation of 95.2 percent of the capacity in the sample was "confirmed," and 4.2 percent was "unconfirmed." In terms of number of plants, 83.6 percent were "confirmed," and 16.4 percent were "unconfirmed." This discrepancy is due to the fact that smaller plants tend to be more difficult to spot with poor image quality and are also generally harder to geolocate manually because they are harder to see. The results of accuracy by size can be seen in Table 4.

We have information on about 8,000 plants that we have not been able to geolocate, representing around 5 percent of global installed capacity. However, to keep the final database consistent, power plants without location are excluded. These plants may be geolocated and added at a later date.

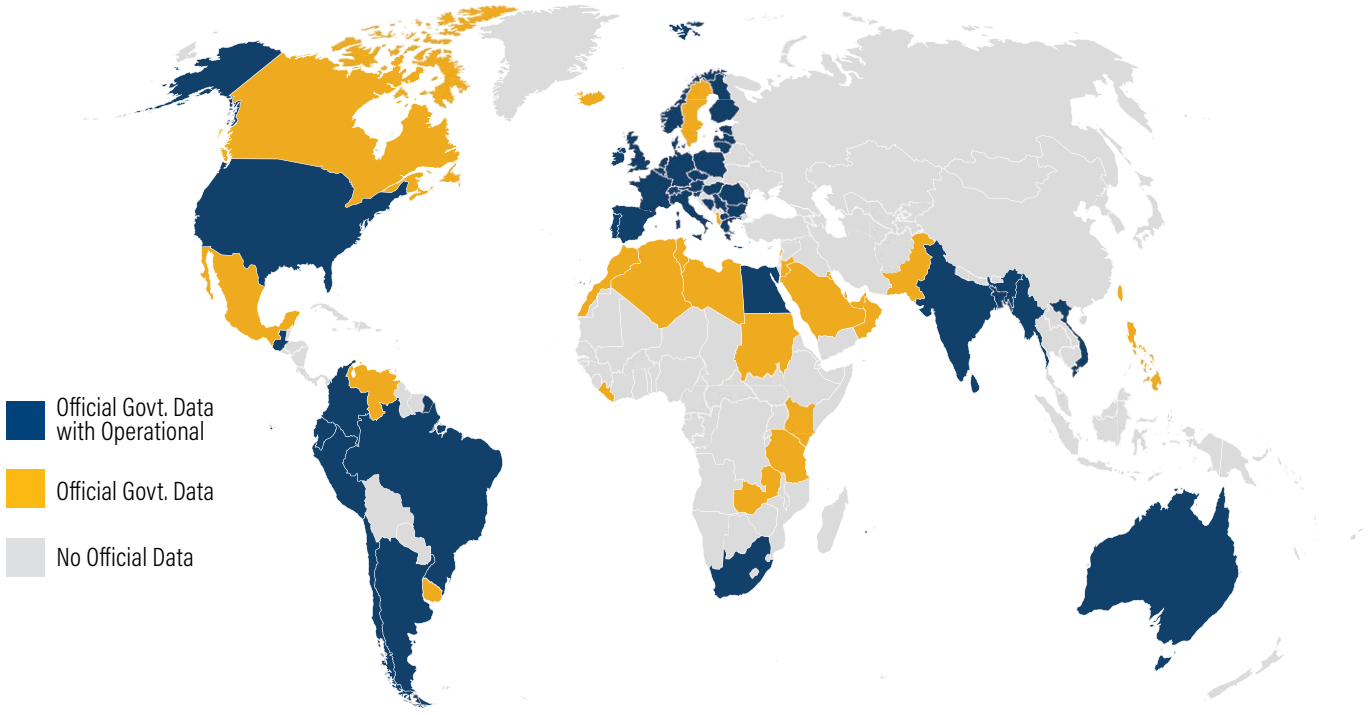
5.3 Matching Databases to Include Generation

Annual electricity generation by plant is often not reported. When it is, it is usually reported in a database that is separate from the plant characteristics data, with a different plant ID. Therefore, we have to match the records for generation with plant characteristics.

6. ESTIMATION OF YEARLY GENERATION BY PLANT

Although many sources report data on a power plant's physical characteristics, such as fuel type and capacity, far fewer report operational characteristics such as actual yearly electricity generated. (See Figure 2 for a map of countries that report data.) Data on generation is important on its own and also because it is linked to plants' fuel and water use, greenhouse gas emissions, and particulate matter emissions that are relevant for monitoring air pollution.

Figure 2 | **Countries Identified by WRI that Provide Official Government Power Plant-Level Data**



Note: Operational data includes either generation or emissions by plant. Even those countries that report data often only include thermal plants, excluding renewables. WRI has done its best effort to identify officially reported data by governments but cannot assure that some countries who report data have not been omitted.
Source: WRI, April 2018.

When (annual) generation is reported, we include it directly in the database. When generation information is not available for a specific plant, we estimate it, although estimating annual generation at the plant level is challenging. Electricity generated by each plant varies by time period and depends on factors such as the regulatory environment, level of demand, cost of fuels, and extent of planned and unplanned maintenance in addition to plant-level characteristics. We devised two potential ways to estimate annual plant-level electricity generation: scaling information on aggregate generation by plant size and a machine-learning approach.

To scale total generation to the plant level, we begin with data from the IEA on total annual electricity generation, disaggregated by country and fuel type (see the IEA online statistics database).[11] The categories used within the database for fuel type and subfuel type can be viewed in the project GitHub file "Fuel Type Thesaurus."[12] For each power plant, we determine its fraction of the total generation capacity of plants with the same fuel type in its country. We then multiply the total annual generation by country and fuel type by this fraction and use it as the generation estimate for the plant. This method essentially allocates the known total generation among all plants according to their relative capacity. In more formal terms, generation for plant h is as follows:

$$(1) \qquad G_{?hfcy} = \frac{K_{hfcy}}{K_{fcy}} \bullet G_{fcy}$$

G represents generation (in megawatt hours, or MWh), and *K* represents installed capacity (in MW). These measures are indexed by period *y* (year), country *c,* and fuel type *f*.

This method has the advantage of ensuring that the individual plant-level generation estimates add up to the correct national totals. However, it does not incorporate any plant-specific information about age, efficiency, or other operating characteristics (beyond capacity).

We have investigated the use of a machine-learning algorithm to refine these estimates, based on generation data reported for power plants in several countries. However, the performance of this method is poor and does not produce estimates that are demonstrably more accurate than the simple method of allocating generation by relative capacity. (See Appendix A for more information on the machine-learning model.) In the future, we intend to revisit the machine-learning method to improve it, and we will incorporate it as the basis for generation estimates if it is significantly more accurate than the simple allocation method.

An alternative would be to use unit dispatch models (such as PLEXOS[15] and GE MAPS[16]) for each jurisdiction in each year. Whereas this approach has the potential to be more accurate than the capacity-weighted allocation method we currently use, it requires building unit commitment models and populating them with plant-characteristic data, which we have not done. The hope is that more independent researchers will contribute to improving the generation data as more of them use the database and the plant-level information it reports. At this point there is no formal process for contributing data or submitting corrections, although we have accepted some corrections from researchers who have taken a look at the beta version; we have also granted KTH access to our data for some manual data additions and edits. At a later stage we will revisit a more robust crowdsourcing feature. Ultimately, these complementary methods will be compared for accuracy, and the generation estimation methodology may be amended accordingly.

We include a clear flag for the data that is estimated to easily distinguish it from the reported data.

7. MAINTAINING THE DATABASE

Ease of updating is a major concern. There are two basic paths to updating the database that are not mutually exclusive. When new or updated data are identified for a specific country, they are either incorporated manually into the relevant Fusion Table, or the relevant automated script is run to incorporate them. When machine-readable data are identified for a country for the first time and replace manually collected data, a new automated script is developed to incorporate them. This replaces the manual Fusion Table (which is archived for later reference). Much of the data, especially for low-income countries, has been obtained manually. This means that any updating is likely to also be done manually.

At this stage, additional data are likely to come from smaller sources that provide the data occasionally and in formats that are not machine readable. Collecting more data manually expands the database's coverage but makes it more expensive to maintain over time.

To address the update and maintenance challenges, WRI has written guidelines for data updates for each country where the data were collected manually.[16] Through a partnership with KTH, the data for Africa and Latin America will be manually updated periodically.

In collaboration with IBM, the project is developing an artificial intelligence (AI) program that aims to query different web sources, identify announcements of power plant commissionings and closures, and translate the information into the database format. If successful, this will reduce the need to manually collect data, which is time intensive. These approaches will help ensure that the database continues to be up to date.

Finally, the hope is that official statistics on plant-level data will become more common over time, as the costs of providing the information decrease and the value of having information for all plants becomes clearer to stakeholders.

8. FUTURE STEPS

Partners and power plant data experts provided feedback on the database during its development in 2016 and 2017. Moving forward, the focus will be on improving the core database and launching new research based on the core dataset. Future steps include

- improving generation estimates by plant;
- expanding data coverage using artificial intelligence, remote sensing, and crowdsourcing; and
- expanding the number of indicators included by plant, specifically greenhouse gas emissions, water use, and particulate matter emissions.

The estimation of generation by plant will take place in close consultation with partner organizations (Carbon Tracker, GE, IBM). We aim to both identify new data that can improve the accuracy of generation estimation through the existing machine-learning model (see Appendix A) and explore alternative ways to estimate plant-level generation. This may include using country-level unit commitment models.

We plan to use advanced data collection methods to expand coverage of power plants in the database. In partnership with IBM, we will test whether AI programs can query text-based web sources, identify power plant commissioning and closure dates, and translate them into a format that is easy to read into the database.

We are testing new data collection methods, including text analysis from web sources and remote sensing using satellite imagery. Remote sensing can be especially useful for identifying power plant characteristics or locating new power plants. Using remote sensing effectively requires using a large amount of data as training data for the machine-learning algorithm, as well as recent and high-resolution imagery. Using crowdsourcing to confirm plant geolocation has been tested with KTH master students, to some success. Allowing the broader public to have access to the database via an online platform may let the project expand geolocated coverage of plants.

We also plan to expand the number of indicators reported for each plant. These indicators are likely to depend on a mix of estimated and reported data. At the same time, we will move to a more sophisticated database infrastructure where each plant characteristic has its own year indicator. Indicators we hope to add in future expansions of the database include the following:

- Water consumption and withdrawal (per MWh and per annum)
- CO_2 emissions (per MWh and per annum)
- Other emissions such as NO_x, SO_2, PM 2.5 (per MWh and per annum)
- Cooling type
- Cooling source
- Annual capacity factor
- Unit level information (i.e., with data disaggregated at the unit rather than the plant level)
- Grid balancing area to which the plant is connected

Improving power plants' reporting processes would enhance the quality and quantity of power plant characteristics captured in the dataset.

APPENDIX A. ESTIMATING GENERATION USING MACHINE LEARNING

We developed a method to estimate annual generation using a machine-learning approach similar to that described in Ummel (2012). We first gathered data for plants whose annual generation is reported by official sources. Next, we identified the subset of those plants for which we had a complete set of data on six basic explanatory variables: fuel type, capacity, age, relative share of national generation capacity, relative share of national generation capacity of the specific fuel type, and average capacity factor of plants of the same fuel type in the same country. Overall, the plants for which we have generation data and all the explanatory variables represent 23 percent of global capacity (1,546 GW, out of a global total of 6619 GW). Data on the amount of generation captured in the training dataset is shown in Table A1.

We use this data and the Gradient Boosted Regression Trees (GBRT) method to train a model that estimates the plant capacity factor (total annual generation divided by the plant's maximum potential generation) based on three explanatory plant-level variables (fuel type, capacity, and age) and three explanatory system-level variables (plant share of total national generation, plant share of total national generation for the plant fuel type, and average capacity factor by fuel and country). Once trained, we then apply the model to plants for which we know these explanatory variables but not the capacity factor. To produce a final estimate of generation, we convert the capacity factor into a generation amount (using the plant capacity). The final step scales all values for a given country by the same value to ensure that total estimated generation matches the total reported generation at the national level.

This approach is similar to the one described in Ummel (2012). This algorithm iteratively builds a set of decision trees that collectively covers the parameter space of plant- and system-level characteristics and attempts to explain the variation in the dependent variable (capacity factor). Each tree is fairly weak as a descriptor on its own but assembling a large number in a systematic way improves predictive performance (Elith et al. 2008).

We perform an empirical search through the model's hyperparameters to optimize the training process. (See below for the optimum values we identified.) We evaluate the model's performance using a cross-validation technique (the Python sklearn library *cross_val_score* function, with 10 folds) and by calculating the R-square (a measure of fit) of the full training dataset against the model predictions. For the optimum hyperparameters, the model achieves an R-square of 0.472. The relative importance of each of the explanatory variables is shown in Figure A1.

To further characterize the model, we examine how it performs for specific fuel types. Figure A2 visualizes the actual and estimated capacity factor for each plant in the training set by fuel type. As is apparent, the model performs far better for certain fuel types than others. In general, the estimated capacity factor is clustered around the capacity factor by fuel for each country.

Improving Yearly Generation Estimates in the Future

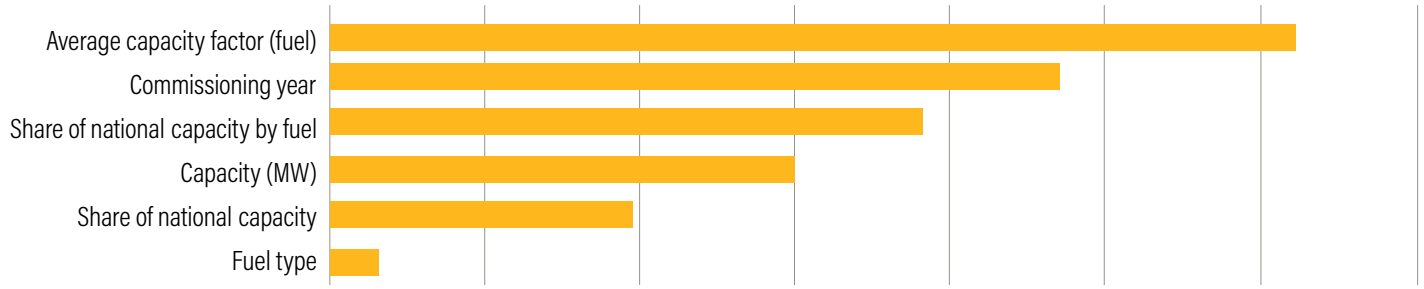
There are two fundamental ways to improve the accuracy of this estimation method: by increasing the amount of training data and including more explanatory variables. When including explanatory variables, we must ensure that they are available both for power plants included in the training sample and for power plants whose generation we will need to estimate. An additional complicating factor is that U.S. data currently provide most of the training data (over three-quarters of data, by total generation). U.S. data may not be representative of the power sector in other

Table A1 | **Training Data for Generation Model**

COUNTRY	DATABASE COVERAGE (GWH)	TOTAL GWH (IEA 2014)	PERCENT COVERAGE
United States	4,079,788	4,339,210	94.0
Argentina	118,330	141,586	83.6
Egypt	160,613	171,747	93.5
India	972,042	1,287,398	75.5
Total	5,330,773	5,939,941	89.7

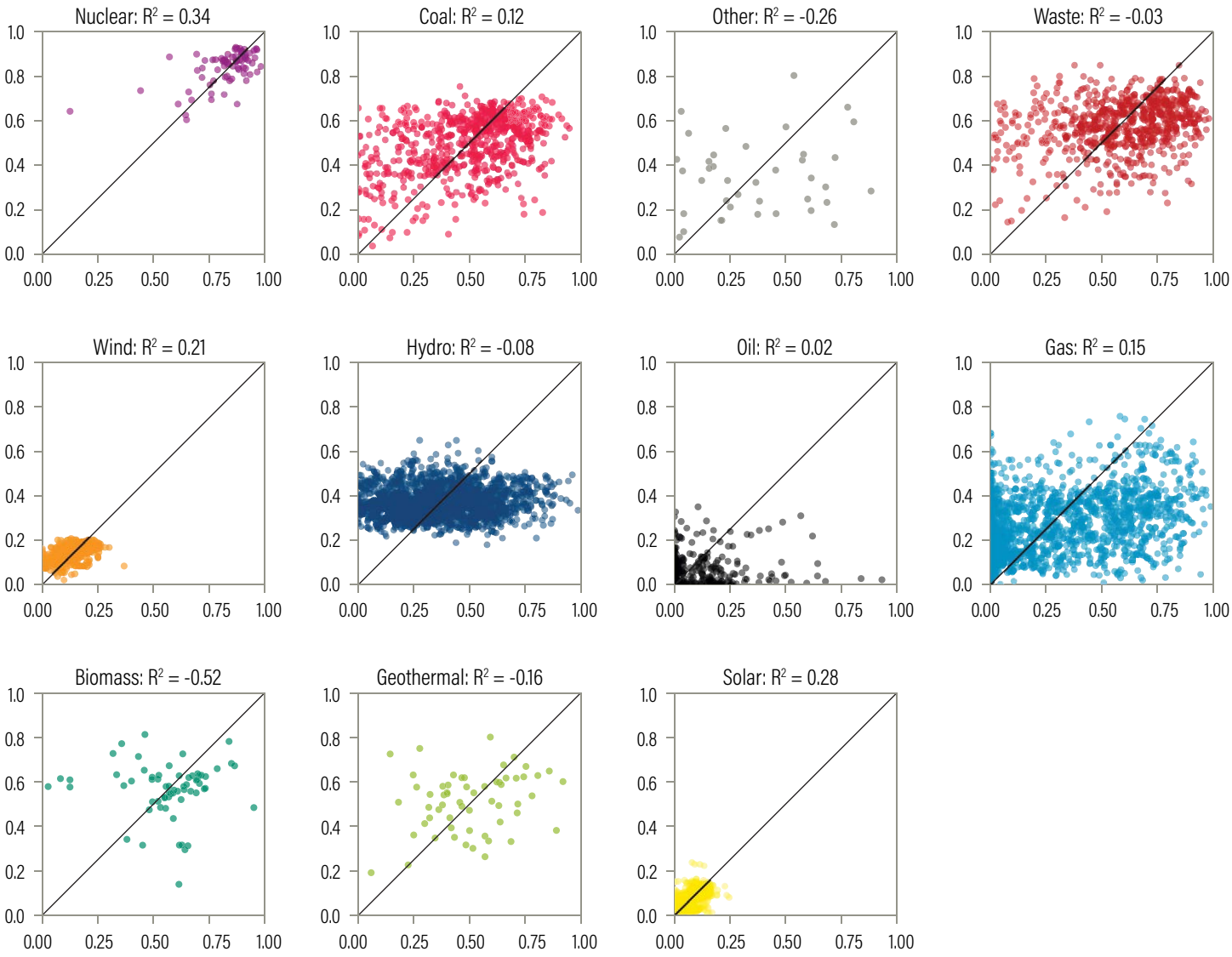
Source: WRI.

Figure A1 | Relative Influence of Each Independent Variable in Trained GBRT Model



Notes: See Table A3 for model details.
Source: WRI.

Figure A2 | Performance of Trained GBRT Model for Individual Fuel Types



Notes: Horizontal axis = ground-truth capacity factor; vertical axis = model-estimated capacity factor. R^2 = R-square.
Source: WRI.

countries, so we have attempted to balance this with data from Argentina, Egypt, and India. However, additional variables are more likely to be available for the United States than these (or other non-Organisation for Economic Co-operation and Development) countries, which would further skew the training data.

The possible additional explanatory variables to include are broadly grouped into technical characteristics of the plant; country-specific socioeconomic variables; weather-related variables; and characteristics of the relevant (regional or national) power system, including interactions between these groups:

(1) $CF(i) = f(Technical, Country, Weather, System)$

where $CF(i)$ is the capacity factor of the i-th plant. Table A2 describes the specific variables under consideration.

Explanatory Power and Accuracy

We measure the explanatory power and accuracy of the GBM projections by comparing how much of the in-sample variation is explained by the projections and by showing the size of the error, both in sample and out of sample. The out-of-sample measure refers to jurisdictions where we have the information on generation capacity and capacity factors by plant. This should be interpreted as a lower bound of the projection error for jurisdictions where we do not have access to generation information by plant. We cannot independently verify the validity of the projections in these jurisdictions. Table A3 summarizes the set of tuning parameters for the GBM process that we use in this application.

Number of iterations (n_estimators). This is the number of boosting stages (or, equivalently, the number of weak learners) that the model performs. This can generally be set fairly high, as gradient boosting techniques are not particularly vulnerable to overfitting. We determined empirically that the optimum R^2 value occurs for a value of 1,500.

Learning rate (learning_rate). This is the fraction by which the contribution to the loss function is reduced for each tree. A small value leads to a larger number of estimators being required to achieve a similar loss, although this also reduces the chance of the model getting stuck in a local rather than global optimum. We empirically find that 0.003 leads to good performance.

Depth of each tree (max_depth). This value specifies the number of splits at each node, or the degree of variable interactions. A value at 1 means there is no interaction between variables and an additive model will be applied. A value greater than 1 will add nonlinearities to the model and improve performance. We empirically find a value of 6 to be optimum.

Number of cross-validation folds (num_folds). To measure the performance of the model, we use a cross-validation technique; the training dataset is randomly divided into N equal-size subsets, and the model is repeatedly fit with a different set of N - 1 subsets and tested against the remaining one. The average accuracy is then reported. Following best practices, we choose a value of N = 10.

Subsampling fraction (subsample). To improve the model's robustness, we use a subsampling technique such that only a (random) fraction of the training data is used to train each weak learner. Following best practice, we set this fraction at 0.5.

Loss function (loss). The loss function is the quantity the model training process is attempting to optimize. We choose the Huber loss, which is a combination of least squares and least absolute deviation. This tends to be more robust against outliers in training data.

Table A3 | Value of Tuning Parameters in GBRT Regression

PARAMETER	NAME IN MODEL	VALUE
Number of iterations	n_estimators	1,500
Learning rate ("shrinkage")	learning_rate	0.003
Depth of each tree	max_depth	6
Number of cross-validation folds	num_folds	10
Subsampling fraction	subsample	0.5
Loss function	loss	Huber

Source: WRI.

APPENDIX B. COVERAGE BY COUNTRY AND SOURCE OF INSTALLED CAPACITY DATA

Documentation is available upon request and records the coverage of installed capacity by country within the Global Power Plant Database. This is done by tallying the sum of installed capacity by plant in the database and comparing it to the most recent information on aggregate capacity. Many countries have recently updated installed capacity figures (India, China, United States) because of regular reporting on total installed capacity of power fleets. If no easily available nationally reported data are available, the U.S. EIA International Energy Statistics (2014)[16] are used to measure aggregate installed capacity in a country. Information on the Global Power Plant Database's coverage by country and source of installed capacity data is available online.[17]

ENDNOTES

1. In this note, *comprehensive* means covering all fuel types and plant-level indicators, not covering all power capacity.
2. For more details on how characteristics are verified through satellite imagery, see WRI 2016.
3. WRI 2018.
4. WRI 2018.
5. Global Power Plant Database. World Resources Institute. https://github.com/wri/global-power-plant-database.
6. The total number of power plants below the 1 MW threshold found in the database is more than 100,000—mostly solar photovoltaic (PV) farms located in Spain and Germany.
7. Unidecode 1.0.22. Python. https://pypi.python.org/pypi/Unidecode.
8. Note that CARMA does not include fuel type.
9. Note that GEO only contains about 10,000 plants.
10. Unidecode 0.04.21. Python. https://pypi.python.org/pypi/Unidecode.
11. Online Data Statistics. International Energy Agency (IEA). https://www.iea.org/statistics/.
12. WRI GitHub. "Fuel Type Thesaurus." World Resources Institute. https://github.com/wri/powerwatch/tree/master/resources/fuel_type_thesaurus. (Currently not a public site.)
13. GE MAPS. GE Energy Consulting. http://www.geenergyconsulting.com/practice-area/software-products/maps.
14. PLEXOS Integrated Energy Model. Energy Exemplar. https://energyexemplar.com/software/plexos-desktop-edition/.
15. WRI 2018.
16. International Energy Statistics. U.S. Energy Information Administration (EIA). https://www.eia.gov/beta/international/data/browser/.
17. WRI 2018.

REFERENCES

Davis, C.B., A. Chmieliauskas, G.P.J. Dijkema, and I. Nikolic. 2015. "Energy and Industry Group, Faculty of Technology, Policy and Management, TU Delft." Delft, The Netherlands: Enipedia. http://enipedia.tudelft.nl.

Elith, J., J.R. Leathwick, and T. Hastie. 2008. "A Working Guide to Booster Regression Trees." *Journal of Animal Ecology* 77 (4): 802–13. https://www.ncbi.nlm.nih.gov/pubmed/18397250.

GWEC (Global Wind Energy Council). 2016. "Global Wind Report 2016—Annual Market Update." http://gwec.net/publications/global-wind-report-2/global-wind-report-2016/.

IEA-PVPS (International Energy Agency Photovoltaic Power Systems Programme). 2016. "Snapshot of Global Photovoltaic Markets." *Report IEA PVPS T1-29:2016*. http://www.iea-pvps.org/fileadmin/dam/public/report/statistics/IEA-PVPS_-___A_Snapshot_of_Global_PV_-_1992-2015_-_Final.pdf.

Platts. 2017. "The UDI World Electric Power Plants Database." S&P Global Platts. March. https://www.platts.com/products/world-electric-power-plants-database.

Ummel, K. 2012. "CARMA Revisited: An Updated Database of Carbon Dioxide Emissions from Power Plants Worldwide." Center for Global Development, Working Paper No. 304. https://ssrn.com/abstract=2226505.

WRI (World Resources Institute). 2016. "Identifying Power Plant Characteristics from Satellite Imagery" (unpublished manuscript). https://docs.google.com/document/d/1EYMpehabWpjn8gB7WsZnx2PLhym0DGGuimXCqZlzukU/edit?usp=sharing.

WRI. 2018. "Global Power Plant Database: Sources of Installed Capacity Information and Coverage by Country, January 2018" (unpublished manuscript). https://docs.google.com/document/d/1S97HqXHhi4CoI8IHIZ4nGE9lcRVCSGYsRvCDv7oD4nU/edit.

ACKNOWLEDGMENTS

The team acknowledges the help of Dani Barjum, Katie Lebling, and the following individuals and organizations who were instrumental in the creation of the database:

Saleem Van Groenou, Google

Rajan Gupta, Global Energy Observatory

Chris Davis, Enipedia

Shahid Hussain Siyal, Constantinos Taliotis, and the students in course MJ2413, KTH Royal Institute of Technology, in Stockholm

Ted Nace, Coal Swarm

Lion Hirth and Wolf-Peter Schill, Open Power Systems Data

Kevin Ummel, Carbon Monitoring for Action

This research is made possible with funding from Google.

When referring to the data, cite this technical note as well as the database directly. Here is the database citation:

Global Energy Observatory, Google, KTH Royal Institute of Technology in Stockholm, University of Groningen, World Resources Institute. 2018. Global Power Plant Database. Published on Resource Watch.

ABOUT THE AUTHORS

Logan Byers is a Research Assistant working in WRI's Climate Program.

Contact: logan.byers@wri.org

Johannes Friedrich is a Senior Associate with WRI's Climate Program and leads the Climate Tools projects.

Contact: jfriedrich@wri.org

Roman Hennig is a Research Analyst in WRI's Climate Program.

Contact: roman.hennig@wri.org

Aaron Kressig is a Research Analyst in WRI's Climate Program.

Contact: aaron.kressig@wri.org

Xinyue Li was a Research Assistant in WRI's Climate Program.

Contact: lixinyue.luna@gmail.com

Colin McCormick is a consultant working on the global power plant database.

Contact: cfmccormick@gmail.com

Laura Malaguzzi Valeri is the Deputy to Vice President of Science and Research at WRI.

Contact: lmalaguzzi@wri.org

ABOUT WRI

World Resources Institute is a global research organization that turns big ideas into action at the nexus of environment, economic opportunity, and human well-being.

Our Challenge
Natural resources are at the foundation of economic opportunity and human well-being. But today, we are depleting Earth's resources at rates that are not sustainable, endangering economies and people's lives. People depend on clean water, fertile land, healthy forests, and a stable climate. Livable cities and clean energy are essential for a sustainable planet. We must address these urgent, global challenges this decade.

Our Vision
We envision an equitable and prosperous planet driven by the wise management of natural resources. We aspire to create a world where the actions of government, business, and communities combine to eliminate poverty and sustain the natural environment for all people.

Our Approach
COUNT IT
We start with data. We conduct independent research and draw on the latest technology to develop new insights and recommendations. Our rigorous analysis identifies risks, unveils opportunities, and informs smart strategies. We focus our efforts on influential and emerging economies where the future of sustainability will be determined.

CHANGE IT
We use our research to influence government policies, business strategies, and civil society action. We test projects with communities, companies, and government agencies to build a strong evidence base. Then, we work with partners to deliver change on the ground that alleviates poverty and strengthens society. We hold ourselves accountable to ensure our outcomes will be bold and enduring.

SCALE IT
We don't think small. Once tested, we work with partners to adopt and expand our efforts regionally and globally. We engage with decision-makers to carry out our ideas and elevate our impact. We measure success through government and business actions that improve people's lives and sustain a healthy environment.

Maps are for illustrative purposes and do not imply the expression of any opinion on the part of WRI, concerning the legal status of any country or territory or concerning the delimitation of frontiers or boundaries.