WORLD
RESOURCES
INSTITUTE

TECHNICAL NOTE

ESTIMATING POWER PLANT GENERATION IN THE GLOBAL POWER PLANT DATABASE

LUOTIAN YIN, LOGAN BYERS, LAURA MALAGUZZI VALERI, AND JOHANNES FRIEDRICH

ABSTRACT

The benefits and costs of power plants, including their environmental impacts, depend on their technology and on how much electricity each plant actually generates. However, plant-level generation data are not reported in most countries, including developing countries where electricity generation is projected to rapidly expand. This technical note documents methods to estimate the annual electricity generation of power plants for the Global Power Plant Database. We use distinct estimation models for different fuel types, including wind, solar, hydropower (hydro), and gas power plants. The methodology combines statistical regression with machine learning techniques. Explanatory variables include plant-level characteristics such as plant size and fuel type, and country-level characteristics, such as country- and fuel-specific average generation per megawatt of installed capacity. We show that fuel-specific models can provide more accurate results for wind, solar, and hydro plants. Estimations for natural gas plants also improve, but the error remains high, especially for smaller plants.

1. INTRODUCTION

Electricity powers modern society. Despite its importance, information about actual electricity generation by power plants is often closely held by plant and system operators and difficult for others to access.

World Resources Institute (WRI) and its partners have created the Global Power Plant Database (GPPD) as an open-source, open-access dataset of the world's power

CONTENTS

Technical notes document the research or analytical methodology underpinning a publication, interactive application, or tool.

Suggested Citation: Yin, L., L. Byers, L. Malaguzzi Valeri, and J. Friedrich. 2020. "Estimating Power Plant Generation in the Global Power Plant Database." Technical Note. Washington, DC: World Resources Institute. Available online at: wri.org/upload/estimating-power-plant-generation-in-the-global-power-plant-database.

plants (Byers et al. 2018). It is assembled from hundreds of public sources and contains information on technical characteristics of power plants, including capacity (megawatts), location, and fuel type. The electricity generation (gigawatt-hours) of the plants is also included in the database when such information is publicly reported. As of June 2019, validated sources of reported plant generation are available for 33 countries (Appendix A).

Power plant capacity, usually measured in megawatts (MW), describes a facility's maximum electric power rate. If a 100 MW plant runs at its maximum capacity for one hour, it will generate 100 megawatt-hours of electricity. In other words, capacity measures the size of the plant and its potential generation rate, while generation describes the actual electricity output of the plant over a period of time.

Thermal plants use several inputs to produce electricity, including fuel to start and run their turbines and water to cool the plants. In addition to electricity, outputs typically include (warmer) water sent back to a water body and steam that evaporates, plus pollutants to air, water, and soil. Energy planners can use geo-located historical plant generation to both monitor emissions and evaluate how best to meet changes in electricity demand over time.

How often and how intensely a power plant runs varies across plant types. Annual power plant generation can be estimated using methodologies that are based on electricity grid optimization (also known as optimal dispatch) or statistical models.

Grid optimization models consider each power plant, its technical characteristics, and the cost of input fuels, and dispatch the simulated plants to minimize the total cost of generation subject to the following:

1. Meeting the quantity of electricity demanded in every period
2. Meeting any technical constraints, such as minimum down times or maximum ramp rates
3. Accounting for the transmission and distribution constraints (if modeled)

These models approximate plant-level generation by estimating "optimal generation," though they do not necessarily reflect actual historical generation, which may have been produced "nonoptimally." To produce high-quality results, optimal dispatch models can be computationally intensive and require detailed information on plant efficiency, which can vary based on operational load and plant age. Efficiency information is not currently available globally.

Statistical models use information on power plants with reported annual generation to estimate the correlation between annual generation and plant characteristics such as capacity, fuel type, and commissioning year. These estimated correlations are then applied to the characteristics of plants without reported generation to create their estimated yearly generation. In contrast to a dispatch model, a statistical model estimates generation based on a power plant's similarity to plants with reported generation, not based on a system optimization.

In our approach, statistical models with machine learning techniques are used to estimate annual plant generation as accurately as possible. Machine learning algorithms capture correlations between generation, technical characteristics, and system variables.

Ummel (2012) approached the problem using unit-level data. Since generation data are often reported at a plant level globally, here we estimate plant-level generation.

Section 2 summarizes the data. Section 3 introduces the methodology. Sections 4–7 describe the fuel-level estimation methodologies for natural gas, wind, solar, and hydropower (hydro). Section 8 explains the limitations of the methods and Section 9 concludes.

2. PLANT-LEVEL GENERATION DATA

Very few jurisdictions openly publish annual power-plant generation data. Even when published, the data are often not in a consistent format. Over the past years, we have aggregated trusted available information. Table 1 reports the number of plants and total capacity for which we have annual generation data for 2016, the year we use to perform the generation estimation analysis and modeling.

Generation data are reported for almost 50 percent of plants, but most of them are concentrated in the United States and other developed countries.

Appendix D lists the external data sources used in this analysis.

Table 1 | **Reported Generation by Country or Region in 2016**

	PLANTS IN GLOBAL POWER PLANT DATABASE	PLANTS WITH REPORTED GENERATION DATA	PLANTS WITH REPORTED GENERATION DATA (%)	CAPACITY WITH REPORTED GENERATION DATA (%)	SOURCE
UNITED STATES	8,644	7,944	91.9	97.5	EIA
INDIA	861	427	49.6	93.2	CEA
AUSTRALIA	429	248	57.8	69.1	NGER
EUROPEAN UNION	9,846	679	6.9	39.4	ENTSO-E & JRC-PPDB[b]
OTHER[a]	10,304	272	2.6	1.0	Multiple sources[c]
TOTAL	**30,084**	**9,570**	**31.9**	**--**	**--**

Notes: [a] See shaded countries in Appendix A for the full list. EIA stands for Energy Information Administration, CEA for Central Electricity Authority, and NGER for National Greenhouse and Energy Reporting. ENTSO-E stands for European Network of Transmission System Operators for Electricity, and JRC-PPDB for The Joint Research Centre Power Plant Database.

[b] Kanellopoulos et al. 2019.

[c] Egypt: Egyptian Electricity Holding Company, http://www.moee.gov.eg/english_new/report.aspx; Latvia: ENTSO-E; Montenegro: JRC-PPDB (Kanellopoulos et al. 2019); Morocco: Office National de l'Electricite, http://www.one.org.ma/FR/pdf/Rapport%20d'activit%C3%A9s%202016%20FR.pdf; Kenya: Kenya Electricity Generating Company; Vietnam: Open Development Vietnam, https://vietnam.opendevelopmentmekong.net/.

Sources: See Appendix D for links to the databases for EIA (Forms EIA-860 and EIA-923), CEA, NGER, and ENTSO-E.

3. ESTIMATING ANNUAL GENERATION

3.1 Plant Operation Depends on the Fuel the Plant Uses

Different types of plants have different generating patterns. Nuclear, coal, and some natural gas plants typically run continuously, or at baseload, since they have relatively low running costs once in operation, but require time and resources to turn on, shut off, or change operating level. Consequently, plant operators limit these ramp-up and ramp-down events. Other natural gas plants with quicker start-up and shut-down times are run when demand increases and are referred to as mid-merit or peaking plants. Plants that rely on intermittent renewable resources, such as the sun or wind, generate only when those resources are available.

Different daily generation patterns translate into different annual average generation by plant, which is commonly measured and expressed by the capacity factor. The capacity factor is a measure of the frequency and intensity of generation. The average capacity factor (*cf*) of plants using fuel *f* in country *c* and year *y* is the ratio of generation of all plants of fuel *f* within country *c* with respect to the maximum potential generation of those plants over the same time period:

Equation 1

$$\mathrm{cf}_{\mathrm{fcy}} = tg_{\mathrm{fcy}} \,/\, (tc_{\mathrm{fcy}} \bullet \#\ \text{hours in a year})$$

Here, tg_{fcy} represents total generation of plants of fuel *f* in country *c* for year *y*, in megawatt-hours; tc_{fcy} is the aggregate capacity of all plants with fuel *f* in country *c* for year *y*, in megawatts.

In general, the capacity factor of a power plant is always between 0 and 1. A capacity factor of 0 means the plant did not operate during the year, and a value of 1 occurs when a plant operates continuously at full power over the entire year. Annual capacity factors never reach 1 in practice, since plants require maintenance during the year and are therefore occasionally off-line. The highest realistic annual capacity factors are around 90 percent.

3.2 This Analysis Addresses Wind, Solar, Hydro, and Natural Gas

Since plant generation depends on the type of plant, we start by separating the power plants by fuel type to increase estimation accuracy. We focus on estimating generation for plants where such information is not readily available, but where we have sufficient information for the statistical estimates to be appropriate: plants powered by intermittent renewables (solar and wind), hydro, and natural gas.

We do not address nuclear power generation, as information for all nuclear power plants is reported by the International Atomic Energy Agency. In this analysis, we also exclude coal plants because coal-plant generation estimates are pursued by other dedicated projects; e.g., Gray et al. (2018) published by Carbon Tracker. For other types of plants, we have too few observations: oil plants represent only 4.7 percent of capacity in GPPD; biomass, waste, geothermal, wave, and tidal are also represented by a relatively small number of plants, though these fuel types may constitute a large portion of generation within some geographies. With such sparse data, statistical analysis would not provide accurate results, and it would be challenging to characterize and communicate errors for those fuels. We therefore impute generation for these types of plants by using average capacity factors by country for those fuel types.

In some cases, it is reasonable to assume that the generation of a power plant is not affected by the generation patterns of other power plant types in the system. For example, solar and wind generate only when the sun shines or the wind blows and tend to do so at very low marginal costs. These plants are part of the generation mix any time they are available, subject to transmission constraints. For thermal power plants this assumption is less appropriate, as the system operator will generally dispatch plants that are cheaper relative to all available plants. A plant's generation will therefore depend not only on its characteristics and costs, but also on the characteristics and costs of alternative plants and on total demand for electricity.

3.3 Baseline Model

For most countries, we have information on total annual generation by fuel type. The International Renewable Energy Agency (IRENA) provides both national-level capacity and national-level generation for renewable plants, so we can build a consistent measure of annual capacity factor. For fossil fuel plants in Organisation for Economic Co-operation and Development (OECD) countries, the International Energy Agency (IEA) reports both generation and capacity by country by fuel, which provides a measure of generation per megawatt installed. For fossil fuel plants in non-OECD countries, we have a more imperfect measure, as the reported generation by country by fuel comes from the IEA, but total capacity comes from national totals through the GPPD. In the first version of the Global Power Plant Database, Byers et al. (2018) used this information to define annual estimated generation of plant *i* using fuel type *f* in country *c* for a given year *y* as the average capacity factor for plants of fuel *f* in country *c* for year *y* multiplied by the capacity of plant *i*, where capacity factor is defined in Equation 1.

Equation 2

$$\hat{g}_{ifcy} = cf_{fcy} \cdot c_{ifcy} \bullet \ \# \text{ hours in a year}$$

Here, $\hat{g}_{\mathrm{ifcy}}$ is the estimated annual generation in megawatt-hours for plant *i* with fuel *f* in country *c* for year *y*; *cf* represents the capacity factor by fuel; c_i is the capacity of plant *i* in megawatts.

This approach, which we refer to as the baseline model in the rest of the document, has the desirable feature that the sum of the generation of all plants adds up exactly to the reported generation by fuel type. However, it assumes that all plants of the same fuel type generate at the same intensity (capacity factor) and ignores many other relevant factors. As a result, it is associated with relatively large estimation errors.

3.4 Including Other Predictors of Generation

Plant-level generation depends on plant-level and system-level characteristics. Wind, solar, and hydro generation depend in part on natural resources availability (wind speed, solar irradiation, and water runoff, respectively). These climatic measurements are available from global physical models like the Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2) by NASA (the U.S. National Aeronautics and Space Administration), and ERA5 by the ECMWF (European Centre for Medium-Range Weather Forecasts). In sections 4, 5, 6, and 7, we explain how these measurements are derived and used in our model.

The generation of thermal power plants, which mostly run on fossil fuels, depends on plant characteristics such as size and thermal efficiency, and the fuel used and its price, but also on system characteristics, such as the cost of generation for alternative plants and the level of electricity consumed, which jointly determine whether a plant will be dispatched or not. Not all of these variables are globally available, so we use country-level capacity factors as a proxy for relevant country-level information.

3.5 Machine Learning Model

Our main goal is to estimate annual historic plant-level generation accurately. Machine learning models are well suited for the challenge (Olden et al. 2008). We use supervised machine learning models, which define the relation between the dependent variable—in our case, plant-level annual generation—and independent variables, or predictors, chosen by the researchers. With finely tuned parameters, the model is optimized over many iterations to minimize the estimation error, given a set of data that includes both dependent and independent variables (called the labeled training data).

Machine learning models can use a large number of algorithms. We adopted the gradient boosting tree (GBT) regressor as the model algorithm. The GBT is an appropriate modeling algorithm in this case due to the following:

1. Regression trees can capture nonlinear relationships (e.g., wind generation and wind speed don't have a linear relationship. At some point a turbine won't generate any more electricity as the wind speed continues to increase).
2. Tree-based models allow us to easily see which predictors contribute more to the prediction.
3. Tree-based models predict by looking at similar samples in the training data, so the prediction won't go too far from the range of the target variable in the training set.

This algorithm iteratively builds a set of decision trees that cover the space of plant-level characteristics and aim to explain changes in the dependent variable. Using a large number of trees that are built systematically improves the predictive performance of the model, even if each tree is fairly weak on its own (Elith et al. 2008).

In our case, the dependent variable can be either plant-level annual generation or annual capacity factor. They are directly related, as outlined in Equation 1.

We choose capacity factor as the dependent variable for two reasons. First, using generation would put more weight on minimizing the error for larger plants, which tend to generate the most over the year. Second, it is easier to interpret results for capacity factor without further normalization. Once we obtain an estimate for the capacity factors, it is easy to calculate annual generation for a plant by inverting Equation 1.

One of the risks of machine learning models is that they "overfit" to the training data, leading to a small prediction error for the original training set, but a large one when applying the model to new or unseen data. This risk is enhanced if the training data are not representative of all potential observations, which occurs in our case as most of the labeled observations in the current iteration of the GPPD are for plants in the United States. We mitigate this risk by

- including the fuel-country capacity factor as an input variable in the plant-level generation estimation;
- dividing the United States into regions based on the North American Electric Reliability Corporation (NERC) classification and using separate region-level capacity factors, thereby increasing the variation in fuel-country capacity factors; and
- testing the models on test data that are not used during model training.

The U.S. Energy Information Administration (EIA) reports capacity and generation information for power plants of all fuel types in the United States. Each unit is labeled with the NERC region it belongs to. We derive NERC-region capacity factors by fuel by aggregating the unit-level information to regional levels.

3.6 Data Cleaning and Outlier Detection

The analysis is based on official information, but generation or capacity may be reported incorrectly, leading to an unrealistic capacity factor for a given plant. In other cases, plants may not operate for a whole year due to maintenance problems. We cannot predict prolonged maintenance periods within this type of analysis and focus on predicting generation for plants that are consistently available during the year. Outliers may unduly influence the model and lead to inaccurate predictions.

To reduce measurement error, we start by dropping cases where the capacity factor is either smaller than 0 or larger than 1. Capacity factors exceeding 1 likely represent mislabeled capacity or generation values. Capacity factors below 0 may be caused by power plants importing more power than they export, as is the case for pumped storage facilities.

We also eliminate outlier observations with capacity factors that are more than three standard deviations from the mean (in line with Iglewicz and Hoaglin 1993) calculated across all countries or regions for the particular fuel type.

Details on any additional fuel-specific data cleaning processes are provided in each section.

3.7 Model Validation and Testing

We split the labeled dataset into training, validation, and test sets. The model is fitted to the training set and tuned based on its performance on the validation set, in an iterative manner if necessary. It is then evaluated on the test or unseen data.

The test data consist of 20 percent of the original dataset, stratified by country to ensure that they match the regional distribution of the overall labeled dataset.

We use stratified K-fold cross validation to divide the sample between validation and training sets. With cross validation, the model is optimized over K equally sized subsamples, or folds, of the training data, as shown in Figure 1 for the case of 10 folds. The model is trained on (K-1) folds and tested, or validated, by applying the trained model on the subsample that was not used (Varian 2014). This is repeated by rotating the subsample that is kept out. By repeating this process K times, we are able to gain K validation scores. The final validation score is the arithmetic mean of the K scores.

K-fold cross validation mitigates the risk of a single training set not being representative of the larger data and skewing the model (Shulga 2018). It also allows for validation while using all of the data for training, which is particularly helpful when the data are limited, as in our case.

We tune the model to attain lower cross-validation scores. Eventually, we apply the fine-tuned model on the test set to assess the model's performance. Figure 2 illustrates the entire process of training, validation, and testing. The same process is conducted for each of the fuel-specific models.

3.8 Model Performance Evaluation

To evaluate model performance, we compare the estimated capacity factor with the reported capacity factor using two metrics, the mean absolute error (MAE) and the mean absolute percentage error (MAPE):

Equation 3

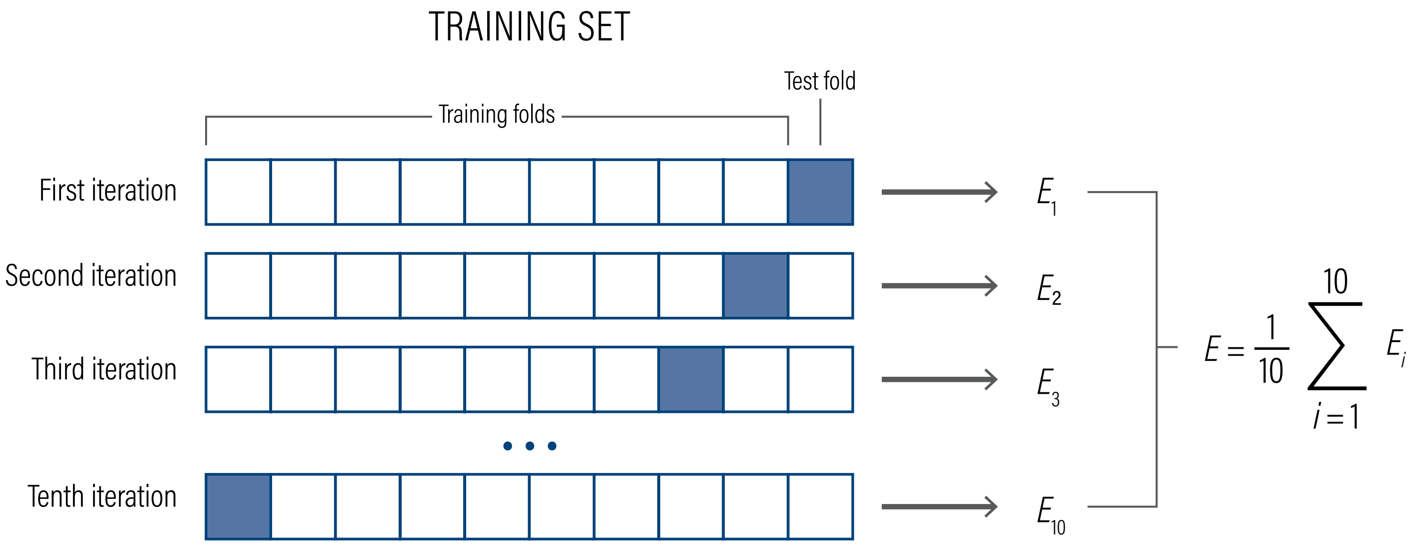
$$MAE = \frac{\sum_{i=1}^{n} \left|cf_i - \widehat{cf_i}\right|}{n}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \left|\frac{cf_i - \widehat{cf_i}}{cf_i}\right|$$

where n is the sample size, cf_i is the true capacity factor for observation i, and $\widehat{cf}_i$ is the estimated value of the predicted variable. The expression between vertical bars is the absolute value.

The MAE compares the true and estimated values directly and measures the average deviation across all observations, providing an error measure that is easy to interpret, but does not assess the relative size of the error. The MAPE, on the other hand, is unitless and is therefore easier to use when comparing accuracy across capacity factors with different magnitudes.

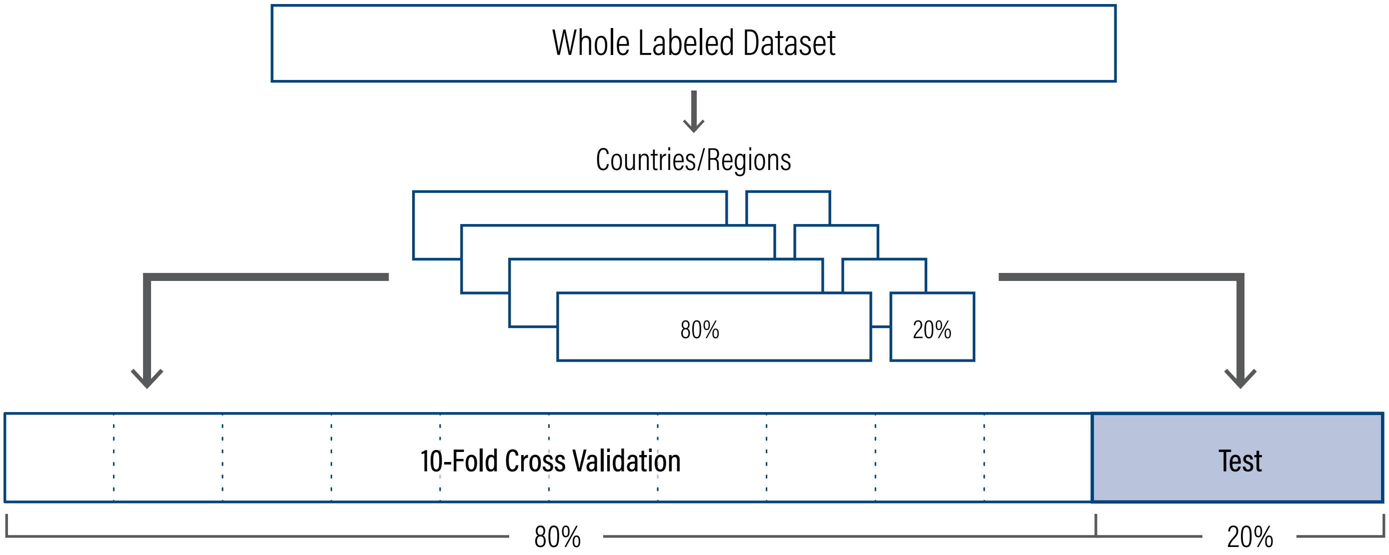
Figure 1 | Cross Validation for Case with K=10 Folds



Note: E_i = average of the predicted variable across training folds in iteration i.

Sources: Rosaen, K. 2016. "Learning Log / 2016-06-20." Blog. http://karlrosaen.com/ml/learning-log/2016-06-20/.

Figure 2 | Training, Validation, and Test Split



Source: Authors.

The MAE and MAPE both summarize the model performance by measuring its error. Throughout the rest of the paper, we compare the model outputs to the errors that derive from assuming that generation per megawatt is always the same for plants with the same fuel type within a country/region (the baseline model) and verify that the methods we use are an improvement over the baseline model.

4. GAS

4.1 Model Description

Gas-fired power plants generate electricity by burning natural gas and use different technologies, including combined cycle gas turbines (CCGTs), open cycle gas turbines (OCGTs), and steam turbines (STs), which are typically older. The technology can affect the plant's capacity factor. In general, CCGTs are more expensive to build, but are more efficient, which makes them more appropriate to use for continuous or baseload generation. CCGTs recycle heat from the combustion process and use it for further generation. OCGT power plants are smaller, have lower construction costs, and are less efficient, but they are faster to switch on and off, making them suitable for intermittent or occasional use such as peaking or mid-merit (IEA and NEA 2015). We include technology type as one of the input variables in the machine learning model. Whereas CCGT plants are often explicitly identified in the technology description, OCGTs can be associated with several technologies, although they are roughly equivalent to the combustion gas turbines referred to in Table 2. CCGT plants have generally higher capacity factors, as reported in Table 2 for the United States in 2016. Other technologies are not as common. Internal combustion engines have low efficiency, high emissions, and high maintenance costs. Fuel cells use an electrochemical process to convert the hydrogen obtained from natural gas to electricity. They tend to be expensive, but are used in combined heat and power applications (Darrow et al. 2017), which explains why they are used relatively intensely with a 0.66 capacity factor.

The average capacity factor of CCGTs was about 0.42 for U.S. natural gas plants, equivalent to their generating at maximum capacity for more than 40 percent of the year. Combustion gas turbines (GTs) and internal combustion engines (ICs) have much lower average capacity factors while STs tend to be older and have the lowest values.

The generating technology information is available from the EIA for U.S. gas plants and on the commercial World Electric Power Plants (WEPP) Database by S&P Global Platts for non-U.S. gas plants.

In addition to technology, age and capacity of the plant are also included as independent variables, together with the average capacity factor across all the natural gas plants for each region.

- Capacity
- Age
- Generating technology type (CCGT, GT, IC, ST, fuel cell [FC], single shaft combined cycle gas turbine [CS])
- Average natural gas capacity factor by region

Table 2 | **Natural Gas Average Capacity Factor by Generating Technology, United States (2016)**

GENERATING TECHNOLOGY	AVERAGE CAPACITY FACTOR	NUMBER OF PLANTS
Combined cycle gas turbine	0.42	417
Single shaft combined cycle gas turbine	0.45	17
Fuel cell	0.66	32
Combustion gas turbine	0.18	543
Internal combustion engine	0.23	88
Steam turbine	0.13	128

Source: Authors' elaboration of Global Power Plant Database data, originally from the U.S. Energy Information Administration. 2016. "Form EIA-860." https://www.eia.gov/electricity/data/eia860/.

Age is correlated with plant efficiency, with the newer plants being more efficient and therefore dispatched more frequently. Capacity is also correlated with more frequent dispatch or higher intensity of operation as larger plants tend to be used more intensely.

Plant capacity, age, and unit type are fed into the gradient-boosting tree model, together with the capacity factor by region, to determine the predicted capacity factor for each plant. The workflow is depicted in Figure 3.

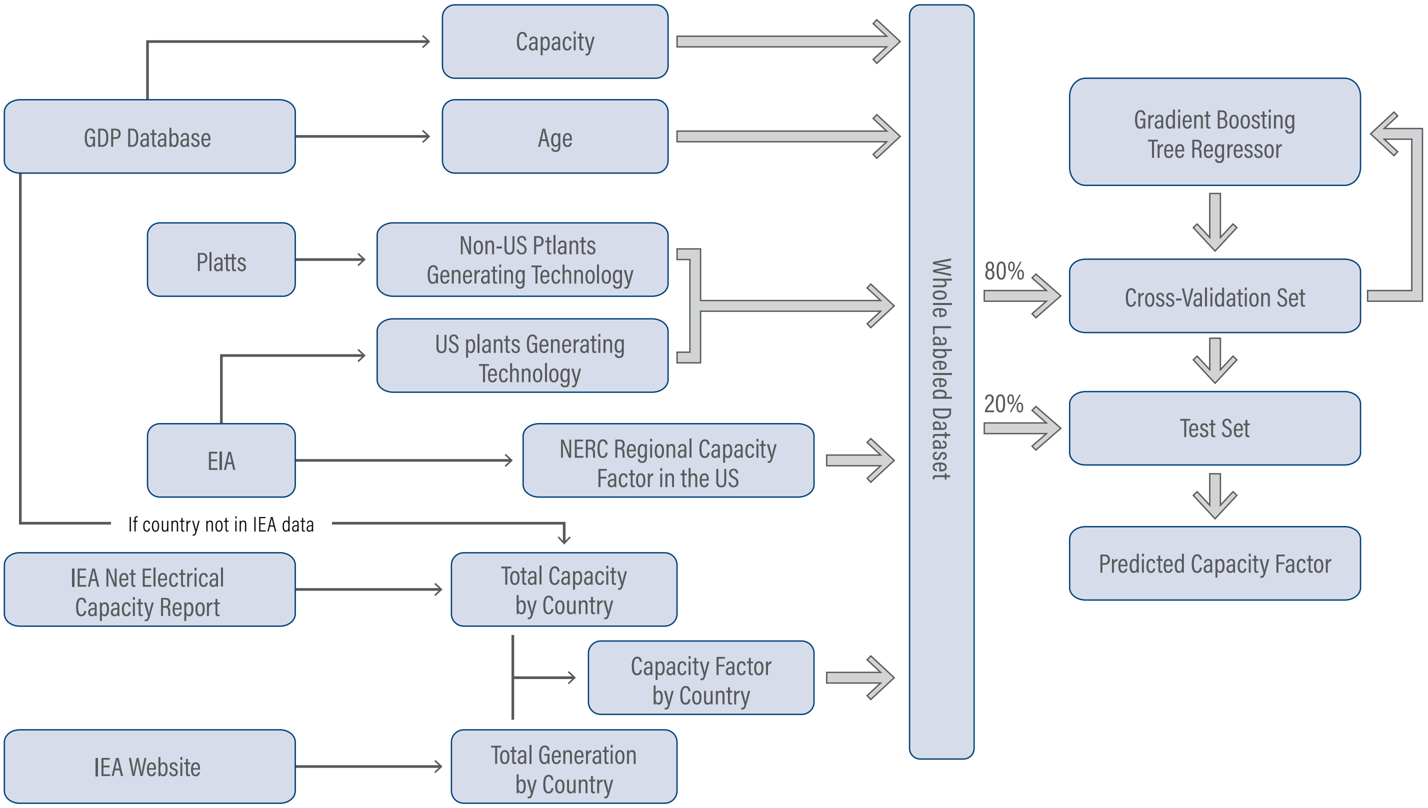
The model allows us to estimate generation for 1,482 plants whose generation amount in 2016 is currently not available in GPPD. For 487 plants where we do not have technology type, since they are not matched with any records in WEPP, we will continue to impute generation using the capacity factor by region or country.

4.2 Data

Global Power Plant Database. The GPPD contains 2,598 gas-fired power plants around the world commissioned before 2016; about 70 percent of the plants have reported generation. Table 3 shows that North America accounts for 66 percent of all gas plants, but 93 percent of plants with reported generation.

Many power plants use gas in addition to other fuels and some contain several turbines with different generating technologies. For the sake of consistency, we model only the gas plants whose proportion of gas-based capacity is above 95 percent (relative to all fuels) and whose proportion of any single generating technology (categorized as one of CCGT, CS, GT, ST, IC, or FC) is also above 95 percent of total plant capacity. This narrows the training set from 1,780 to 1,284 observations.

Figure 3 | Gas-Fired Power Plant Generation Estimation Workflow


Capacity
GDP Database
Age
Platts
Non-US Ptlants Generating Technology
US plants Generating Technology
EIA
NERC Regional Capacity Factor in the US
If country not in IEA data
IEA Net Electrical Capacity Report
Total Capacity by Country
Capacity Factor by Country
IEA Website
Total Generation by Country
Whole Labeled Dataset
80%
20%
Gradient Boosting Tree Regressor
Cross-Validation Set
Test Set
Predicted Capacity Factor

Notes: GPPD stands for Global Power Plant Database, EIA for Energy Information Administration, IEA for International Energy Agency, and NERC for North American Electric Reliability Corporation.

Source: Authors.

Table 3 | Gas-Fired Plants and Plants with Reported Generation by Region (2016) in the GPPD

	NUMBER OF GAS PLANTS BY REGION	GAS PLANTS BY REGION (% OF WORLD PLANTS)	NUMBER OF GAS PLANTS WITH REPORTED GENERATION DATA (PLANTS THAT CAN BE USED FOR TRAINING)	GAS PLANTS WITH REPORTED GENERATION DATA (% OF WORLD PLANTS WITH REPORTED GENERATION)
NORTH AMERICA	1,708	65.7%	1,651	92.8%
SOUTH AMERICA	128	4.9%	0	0.0%
EUROPE	267	10.3%	73	4.1%
AFRICA	68	2.6%	1	0.0%
ASIA	425	16.4%	55	3.1%
AUSTRALIA/OCEANIA	2	0.1%	0	0.0%
TOTAL	**2,598**	**100%**	**1,780**	**100%**

Source: Authors' elaboration of Global Power Plant Database data.

One hundred eighty-six gas plants, corresponding to 14 percent of remaining observations, have reported capacity factors lower than 1 percent (corresponding to approximately 100 hours of full-intensity generation per year). These plants are difficult to model with our estimation tools. We therefore removed the data points, leaving us with 1,098 labeled observations.

EIA-923

The EIA publishes plant-level generation and generating technology for each unit using information collected through "Form EIA-923." Matching this information with the GPPD, we label each plant with either a unique generating technology or set of generating technologies depending on the configuration of the plant.

World Electricity Power Plants by Platts

While the EIA collects generating technology only for gas plants in the United States, the WEPP database holds similar information for gas plants worldwide. We extracted relevant information from both sources for the analysis, but do not republish the WEPP information as it is proprietary.

The prediction errors on the test data of 220 observations for the natural gas plant generation information are reported in Table 4. The metrics on test data are more realistic reflections of the model estimation performance when using unseen datasets. Appendix B presents the validation scores for each fold in the cross validation.

The model predicts plant generation with an absolute error of 0.136 and an average percentage error of more than 170 percent. While high, this is a distinct improvement compared with the baseline, which uses uncorrected average capacity factor, resulting in an error of more than 350 percent, as shown in Table 4.

Figure 4 shows that the proposed model displays more variation across plants than the baseline approach, which leads to a more accurate prediction. There are a number of plants with very low generation during 2016, shown by the cluster of dots close to the vertical axis. Generally, the model overestimates the capacity factor for low capacity factor plants, as the points are above the 45-degree line, and underestimates the capacity factor of high capacity factor plants.

Figure 5 shows that the error varies by size of the plant and is typically smaller for larger plants, which tend to generate more continuously. The *x* axis is the capacity

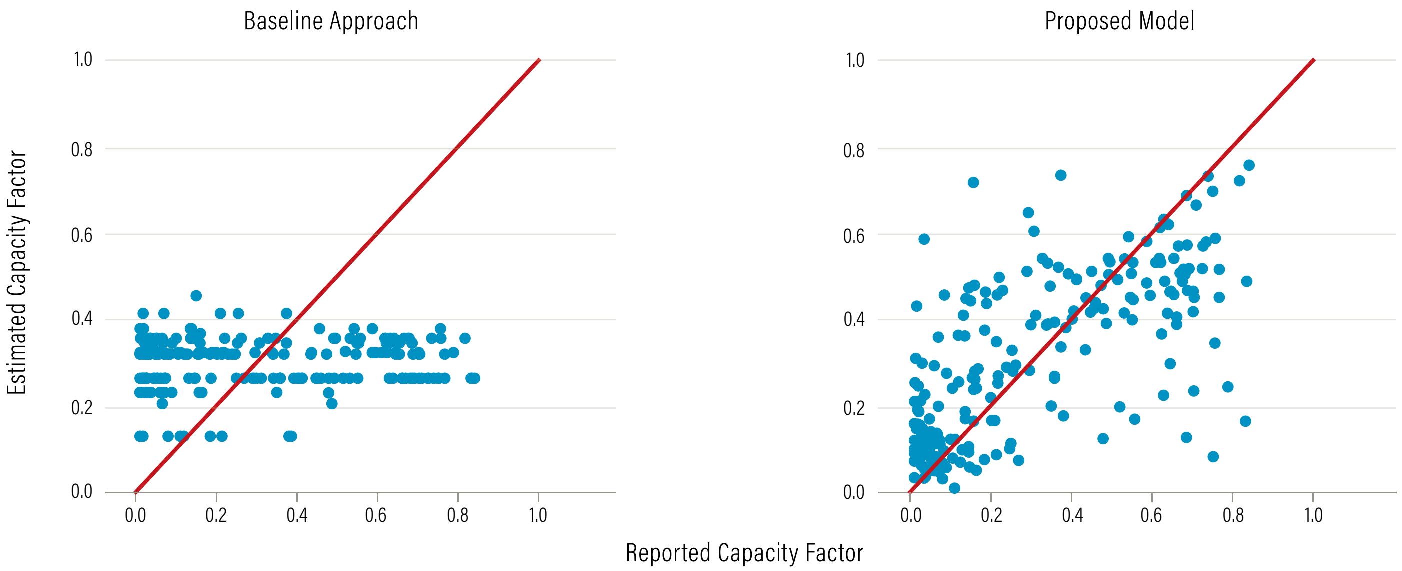
Table 4 | **Model Test Scores for Yearly Capacity Factor of Gas Plants (2016)**

	BASELINE MODEL	PROPOSED MODEL
MAE	0.231	0.136
MAPE	3.537	1.731

Notes: MAE stands for mean absolute error and MAPE for mean absolute percentage error.

Source: Authors.

Figure 4 | **Natural Gas Reported versus Estimated Capacity Factor for 2016 on Test Set, Baseline versus Proposed Model**



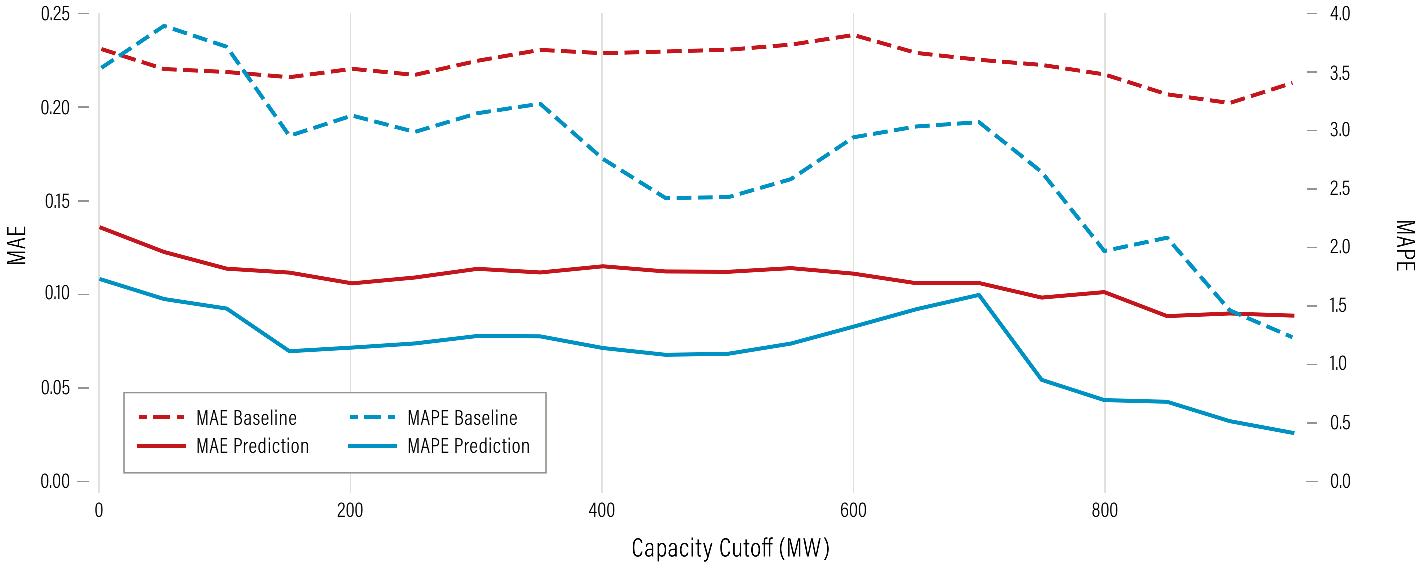
Source: Authors.

cutoff above which the accuracy metrics are recalculated, or, conversely, the capacity cutoff below which plants are not considered. The mean absolute percentage error never falls below 50 percent, suggesting that the predictors we have are not sufficient to accurately estimate plant-level generation for natural gas plants.

The training data selection and method implementation do not allow us to estimate which plants will be on extended maintenance or generate for fewer than 100 full-intensity hours during the year. In our sample, this corresponds to about 14 percent of plants with reported generation and technology type.

Tree-based models provide relative importance scores for each of the predictors used in the trained model. Figure 6 reports these "feature importance scores," which always sum to one across all predictors. In this case, a plant's capacity and age are the relatively more important predictors of its capacity factor.

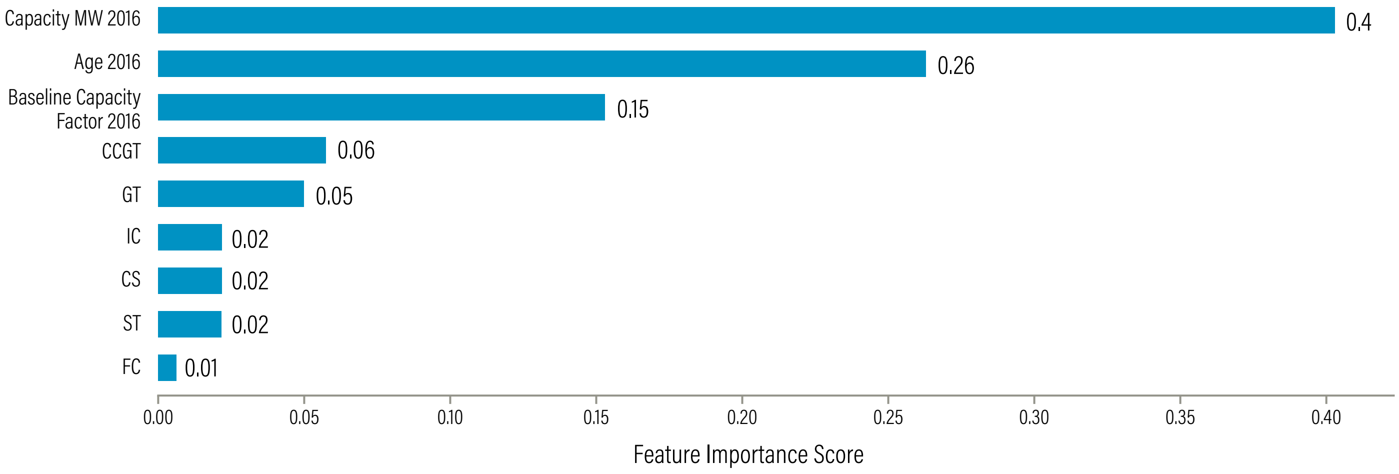
Figure 5 | Gas Model Accuracy by Plant Size



Notes: MAE stands for mean absolute error and MAPE for mean absolute percentage error. MW stands for megawatt.

Source: Authors.

Figure 6 | Relative Importance of Predictors in the Gas Model



Notes: Relative importance scores do not sum to exactly 1 due to rounding. MW stands for megawatt, CCGT for combined cycle gas turbine, GT for combustion gas turbine, IC for internal combustion engine, CS for single shaft combined cycle gas turbine, ST for steam turbine, and FC for fuel cell.

Source: Authors.

5. WIND

5.1 Model Description

The generation of a wind turbine depends on how much wind blows at its location, and factors such as transmission availability. IRENA reported in 2016 that the capacity factors of wind farms vary significantly by country, with 25 percent of country capacity factors below 18 percent and 25 percent larger than 31 percent. Even in the same location, year-on-year capacity factors can vary widely. For example, Irish national wind capacity factors have historically been between 24 and 33 percent (EirGrid 2016, Figure 5.1).

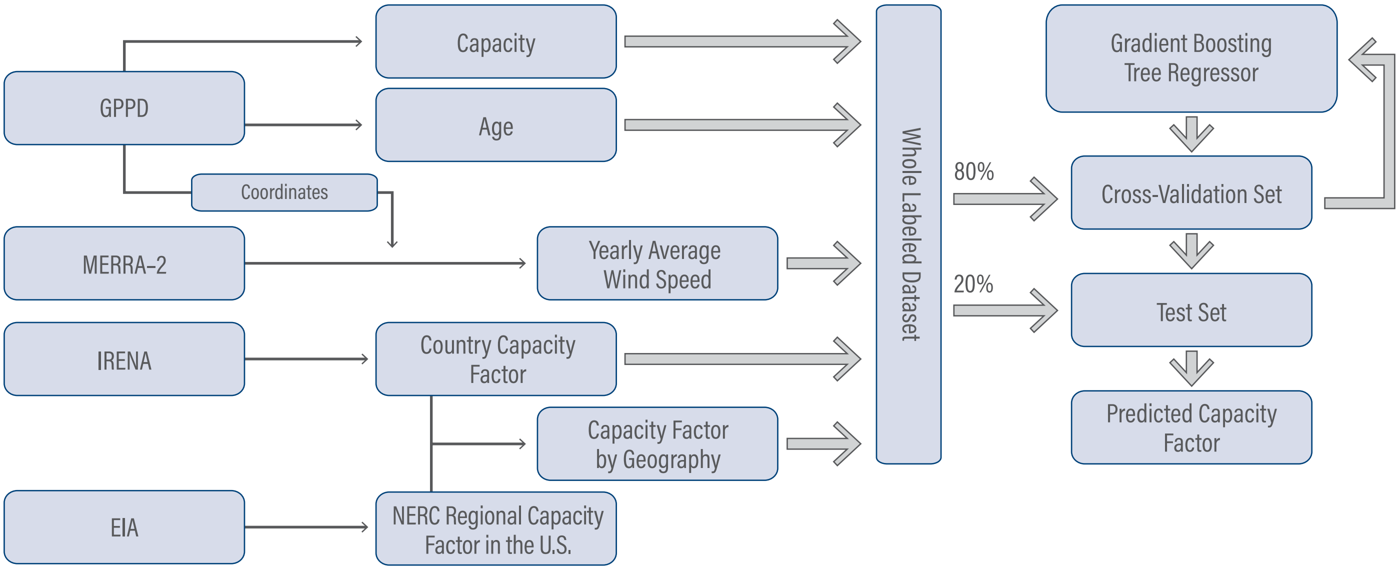
In our model, the generation of each wind farm is a function of the following predictors:

1. Capacity (in MW). The larger the wind farm, the more electricity it can generate.
2. Age, which is correlated with characteristics that we do not observe in the reported data, such as technology and efficiency, including height of the wind shaft.
3. Yearly average wind speed at a specific site, which depends on site location and year.
4. Average capacity factor by region. Adding the capacity factor by region or country helps account for unobserved factors such as curtailment at the country level. Curtailment occurs when wind farms are not allowed to export their electricity to the grid because of transmission, distribution, or electricity system constraints.

MERRA-2 by NASA provides a relatively consistent weather-corrected model that integrates measured wind data from a variety of sources with atmospheric modeling data resulting in output that reports historical wind speed for areas of 0.5 degrees (°) by 0.625°, or approximately 50 kilometers (km) by 70 km at the equator. To obtain wind speed for every wind farm location, which in GPPD is specified as a single (latitude, longitude) point, we interpolate wind speeds between adjacent model grid cells to produce a smoother spatial resolution. Grids are weighted inversely to their distance to the point of interest using inverse distance weighting. Grid-cell centers spatially closer to the target point have more influence on the interpolated value.

The workflow is depicted in Figure 7.

Figure 7 | **Wind Farm Generation Estimation Model Workflow**


GPPD
Capacity
Age
Coordinates
MERRA-2
Yearly Average Wind Speed
IRENA
Country Capacity Factor
Capacity Factor by Geography
EIA
NERC Regional Capacity Factor in the U.S.
Whole Labeled Dataset
80%
20%
Gradient Boosting Tree Regressor
Cross-Validation Set
Test Set
Predicted Capacity Factor

Notes: GPPD stands for Global Power Plant Database and MERRA-2 for Modern-Era Retrospective analysis for Research and Applications, Version 2, from the U.S. National Aeronautics and Space Administration. IRENA stands for International Renewable Energy Agency, EIA for Energy Information Administration, and NERC for North American Electric Reliability Corporation.

Source: Authors.

We also evaluated the accuracy of an alternative two-step process. We first simulated generation using the Climate based Optimization of renewable Power Allocation (COPA) model, a global, bottom-up model of electricity generation (Mosshammer 2016) based on available hourly data, in this case, wind speed. In the second step, we fed this information into a machine learning model. This model did not improve accuracy and is more complex. This lack of improvement may be due to the need for detailed information on hub height, which we do not have for all plants globally. We will revisit the approach if more information becomes available globally.

5.2 Data

MERRA-2

From MERRA-2, we collected monthly average wind speed at the surface level and used it to calculate annual average wind speed.

To be more specific, the relevant dataset from MERRA-2 is named "M2TMNXFLX" and the relevant field is "surface_wind_speed," coded as "SPEEDLML."

GPPD

As shown in Table 5, 1,910 wind farms are included in the GPPD with 944 of them reporting generation, most of them in North America. North America accounts for 47 percent of all wind plants, but 95 percent of plants with reported generation.

We eliminate observations where the capacity factor is more than three standard deviations away from the mean of all wind farms. This leads to the removal of 7 observations, leaving 937 observations.

5.3 Model Evaluation

Table 6 reports average errors for the test set, based on the holdout set of 188 samples. Predicted annual wind capacity factor per power plant has an absolute error of 0.043 and a percentage error of about 16 percent, using the proposed model, down from 0.062 and 24 percent in the baseline. Validation scores for each of the cross-validation folds are available in Appendix B.

On the left side of Figure 8 (the baseline model), we assume that all wind farms in a given region have the same capacity factor. This is likely to be more accurate

Table 5 | Distribution of Wind Plants and Wind Plants with Reported Generation in 2016 in GPPD

	NUMBER OF WIND PLANTS BY REGION	WIND PLANTS BY REGION (% OF WORLD PLANTS)	NUMBER OF WIND PLANTS WITH REPORTED GENERATION DATA (PLANTS THAT CAN BE USED FOR TRAINING)	WIND PLANTS WITH REPORTED GENERATION DATA BY REGION (% OF WORLD PLANTS WITH REPORTED GENERATION)
NORTH AMERICA	896	46.9%	896	94.9%
SOUTH AMERICA	301	15.8%	0	0.0%
EUROPE	622	32.5%	11	1.2%
AFRICA	24	1.3%	0	0.0%
ASIA	23	1.2%	0	0.0%
AUSTRALIA/OCEANIA	44	2.3%	37	3.9%
TOTAL	**1,910**	**100.0%**	**944**	**100.0%**

Source: Authors' elaboration of Global Power Plant Database data.

in smaller countries (e.g., Luxembourg) where wind farms are closer to each other than they are in larger ones. On the right-hand side we include the output of our preferred estimation method, which allows for variation in capacity factors across wind farms and yields the average percentage error rate of 16 percent. Generally, the residuals under the wind model follow a symmetrical distribution around the 45-degree line and the model does not appear to consistently over- or underestimate generation.

Figure 9 shows that the model performs better on wind farms larger than 100 MW, when the mean absolute percentage error falls below 10 percent.

Figure 10 shows that average wind speed and wind farm capacity are the relatively more important predictors in our model.

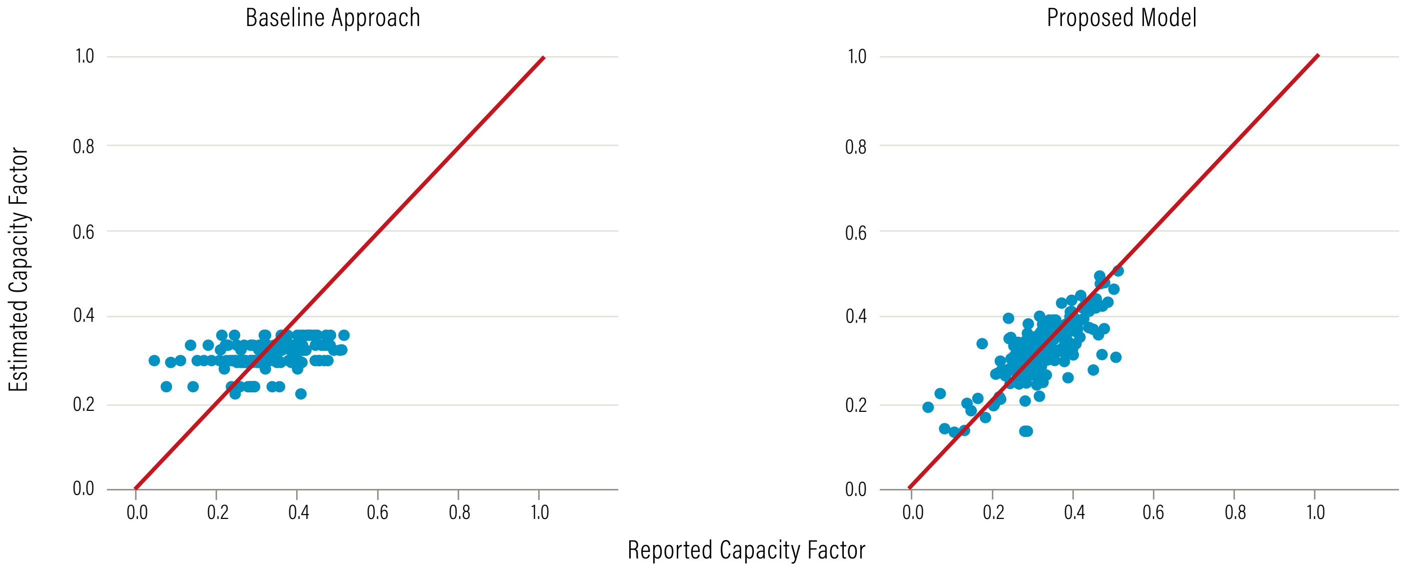
Table 6 | **Test Scores for Yearly Capacity Factor of Wind Plants (2016)**

	BASELINE MODEL	PROPOSED MODEL
MAE	0.0615	0.0427
MAPE	0.2440	0.1599

Notes: MAE stands for mean absolute error and MAPE for mean absolute percentage error.

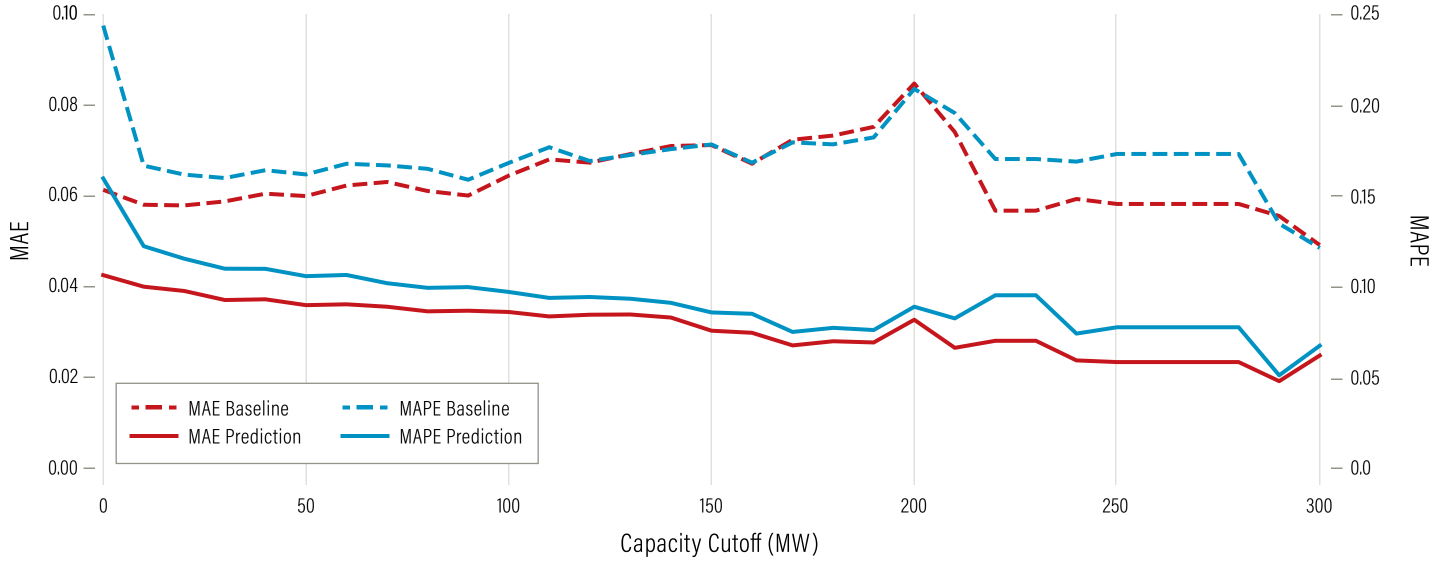
Source: Authors.

Figure 8 | **Reported versus Estimated Capacity Factors, Baseline versus Proposed Model**



Source: Authors.

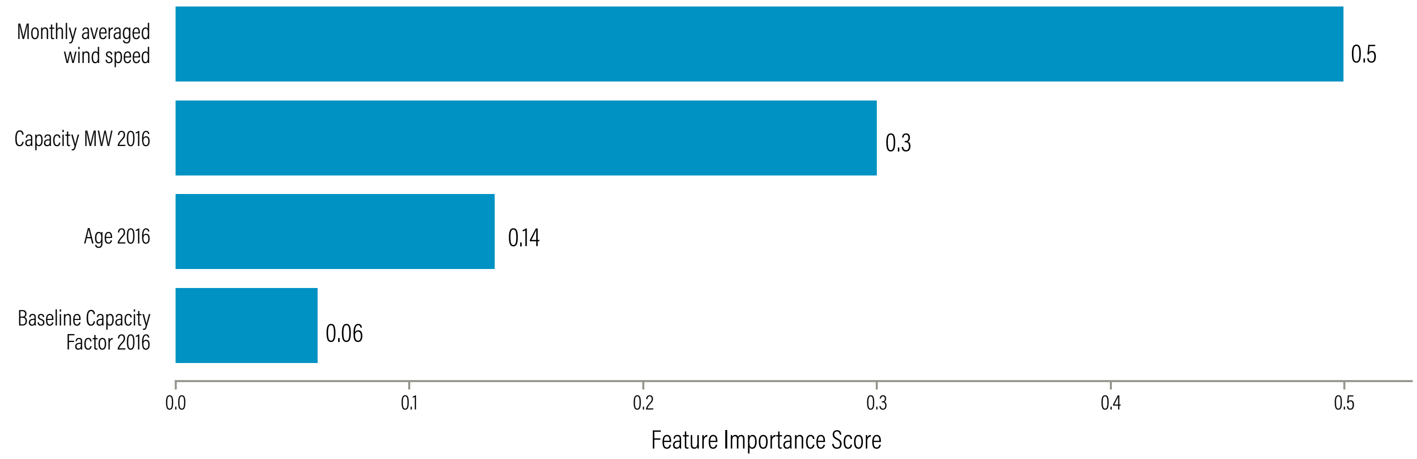
Figure 9 | Wind Model Accuracy Analysis by Plant Size



Notes: MAE stands for mean absolute error and MAPE for mean absolute percentage error. MW stands for megawatt.

Source: Authors.

Figure 10 | Wind Model Accuracy Analysis by Plant Size



Notes: MW stands for megawatt.

Source: Authors.

6. SOLAR PHOTOVOLTAIC

In this section we focus on solar photovoltaic plants. Solar thermal plants represent only 1.6 percent of total solar capacity and are excluded from this analysis.

6.1 Model Description

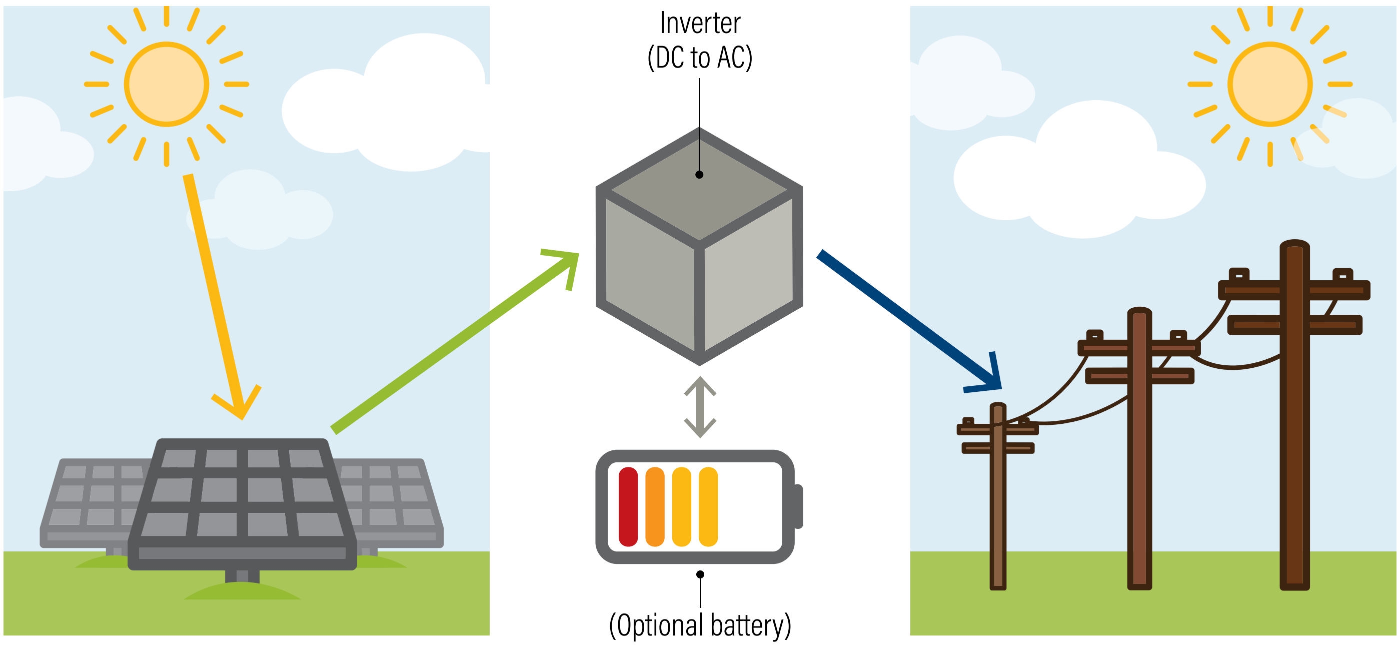
Figure 11 describes the main components of a grid-connected photovoltaic (PV) system. PV modules convert incoming solar radiation into electricity through the photovoltaic effect. Once electricity is generated, balance-of-system (BoS) components help regulate it to ensure grid-quality output. One of the most important BoS components is the inverter, which transforms the electricity from direct current (DC) to alternating current (AC), allowing it to be fed into the distribution system. Sometimes the PV system will have a battery or other storage option, which allows it to provide consistent output, instead of rapidly changing output when events such as cloud cover occur.

One of the main determinants of the amount of electricity generated per MW of installed capacity is the amount of solar radiation received by the solar panel, which can be approximated by the yearly average global horizontal irradiance (GHI) at the surface level.

Temperature also plays a role, as increasing temperature decreases the efficiency of a solar panel (Dubey et al. 2013). We therefore include the yearly average ambient temperature in the analysis.

PV performance decreases over time, by between 0.5 percent to 1 percent per year (Jordan and Kurtz 2013). Age of the PV system may capture this effect and is added as another feature. Age may also be correlated with other factors that change systematically over time but that we do not observe in the available data, such as solar panel technology. Unfortunately, age is not commonly reported, so we also explore models that exclude age.

Figure 11 | Simplified Grid-Connected Photovoltaic System



Notes: DC stands for direct current, AC for alternating current.

Source: Authors.

We collect surface-level GHI and ambient temperature data from MERRA-2 at a resolution of 0.5° by 0.625° (50 km by 70 km at the equator). Irradiance and temperature measurements have to be interpolated from nearby grid cells to create smooth spatial data, and we do so through the inverse distance weighting interpolation function introduced in section 5.1. The interpolation is purely two dimensional and does not include topographic relief or elevation.

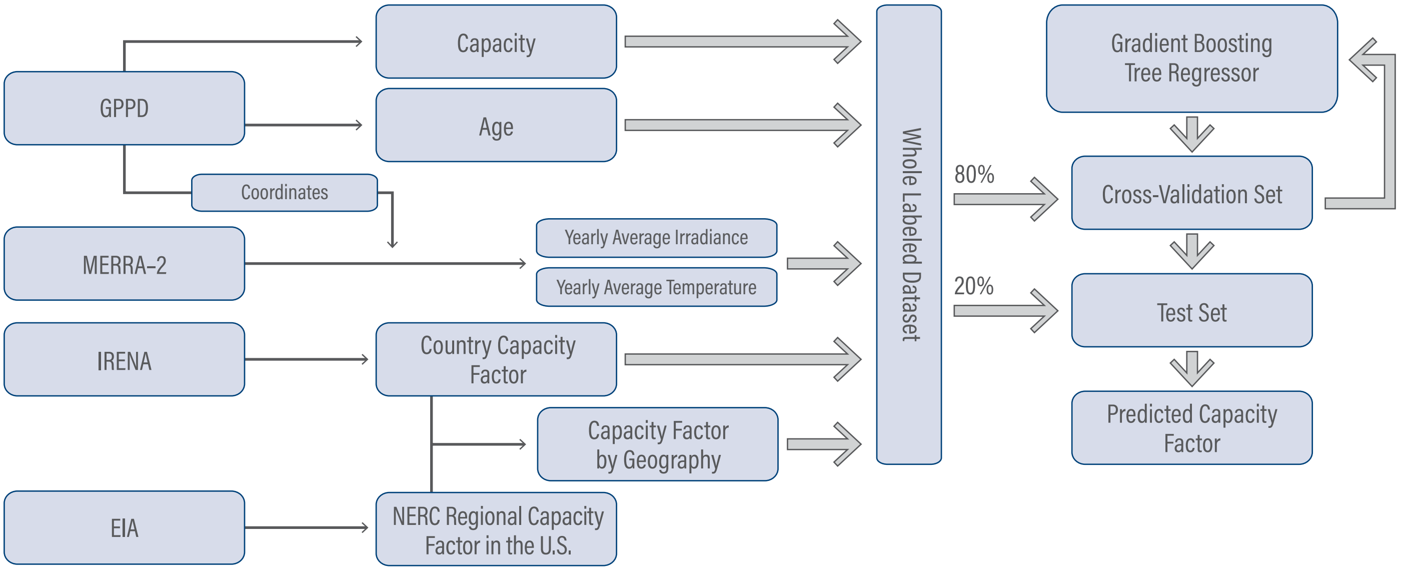
The annual capacity factor by country or region embeds variations in many inputs that we are not able to measure directly, including the average curtailment in a country or region due to transmission, distribution, or system constraints. The capacity factors by country and by U.S. NERC region are derived from IRENA and the EIA, respectively.

To summarize, in the solar model, the dependent variable is capacity factor and the independent variables are

- ground-level yearly average global horizontal irradiance;
- yearly average ambient temperature at the location of the solar farm;
- plant age;
- plant capacity; and
- average solar capacity factor by country/region.

The entire workflow is shown in Figure 12.

Figure 12 | **Solar Farm Generation Estimation Workflow**


GPPD
Capacity
Age
Coordinates
MERRA-2
Yearly Average Irradiance
Yearly Average Temperature
IRENA
Country Capacity Factor
Capacity Factor by Geography
EIA
NERC Regional Capacity Factor in the U.S.
Whole Labeled Dataset
80%
20%
Gradient Boosting Tree Regressor
Cross-Validation Set
Test Set
Predicted Capacity Factor

Notes: GPPD stands for Global Power Plant Database and MERRA-2 for Modern-Era Retrospective analysis for Research and Applications, Version 2, from the U.S. National Aeronautics and Space Administration. IRENA stands for International Renewable Energy Agency, EIA for Energy Information Administration, and NERC for North American Electric Reliability Corporation.

Source: Authors.

6.2 Data

MERRA-2

Table 7 shows which MERRA-2 database variables we collect to calculate the features that feed into the machine learning model.

GPPD

The Global Power Plant Database includes 1,489 solar farms operating across the world by 2016 with a subset of 1,324 having labeled generation, as shown in Table 8. All of the reported generation data refer to North American plants.

The capacity, age, and technical indicators come from GPPD.

In this dataset there are no observations with a capacity factor smaller than 0 or larger than 1. We eliminate any observation with a capacity factor that is more than three standard deviations away from the mean of the capacity factor across all plants. This eliminates 13 observations, leading to a labeled dataset with 1,311 observations.

Table 7 | **MERRA-2 Field Code Map**

CODE IN MERRA-2	DESCRIPTION IN MERRA-2	FEATURE IN MACHINE LEARNING MODEL
TS	Surface skin temperature	Temperature
SWGNT	Surface net downward shortwave flux	GHI

Notes: MERRA-2 stands for Modern-Era Retrospective analysis for Research and Applications, Version 2, by the U.S. National Aeronautics and Space Administration. GHI stands for global horizontal irradiance.

Source: Authors and MERRA-2.

Table 8 | **Distribution of Solar Plants and Solar Plants with Generation Information (2016)**

	NUMBER OF SOLAR PLANTS BY REGION	SOLAR PLANTS BY REGION (% OF WORLD PLANTS)	NUMBER OF SOLAR PLANTS WITH REPORTED GENERATION DATA (PLANTS THAT CAN BE USED FOR TRAINING)	SOLAR PLANTS WITH REPORTED GENERATION DATA BY REGION (% OF WORLD PLANTS WITH REPORTED GENERATION)
NORTH AMERICA	1,324	88.9%	1,324	100.0%
SOUTH AMERICA	4	0.3%	0	0.0%
EUROPE	97	6.5%	0	0.0%
AFRICA	25	1.7%	0	0.0%
ASIA	39	2.6%	0	0.0%
AUSTRALIA/OCEANIA	0	0.0%	0	0.0%
TOTAL	**1,489**	**100.0%**	**1,324**	**100.0%**

Source: Authors' elaboration of Global Power Plant Database data.

6.3 Model Evaluation

Based on the holdout set of 263 observations, Table 9 and Figure 13 show that the model results are more accurate than the baseline estimates. On average, the capacity factor for a power plant is estimated to be within 15 percent of its true capacity factor, based on the test set. Observations cluster around the 45-degree line on the right-hand side of Figure 13, suggesting that the predictors are able to explain within-region variation across solar farms well. The residuals on each example in the test set, as can be seen on the right-hand side of the figure, are symmetrically and universally distributed, which is a sign that the model is robust enough to give predictions of constant quality. Further validation could occur as we collect information on plants outside of the United States. Cross-validation scores in each of the cross-validation folds are available in Appendix B.

The accuracy of the model improves for larger plants and solar farms. Figure 14 shows that for plants over 20 MW, the mean absolute percentage error falls under 10 percent.

Figure 15 shows the relative importance of each feature in the trained model. In the solar model, weather variables measuring the irradiance and temperature at the site play a relatively more important role in the capacity factor prediction.

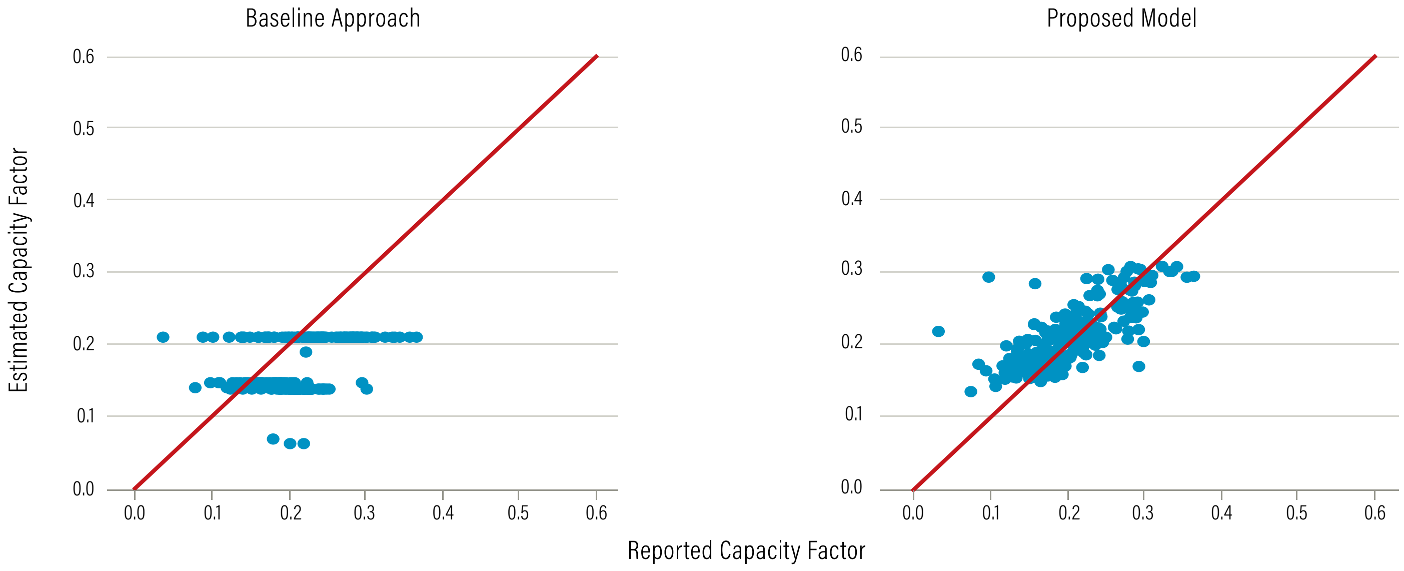
Table 9 | Test Scores for Yearly Capacity Factor of Solar Plants (2016)

	BASELINE MODEL	PROPOSED MODEL
MAE	0.053	0.026
MAPE	0.264	0.153

Notes: MAE stands for mean absolute error and MAPE for mean absolute percentage error.

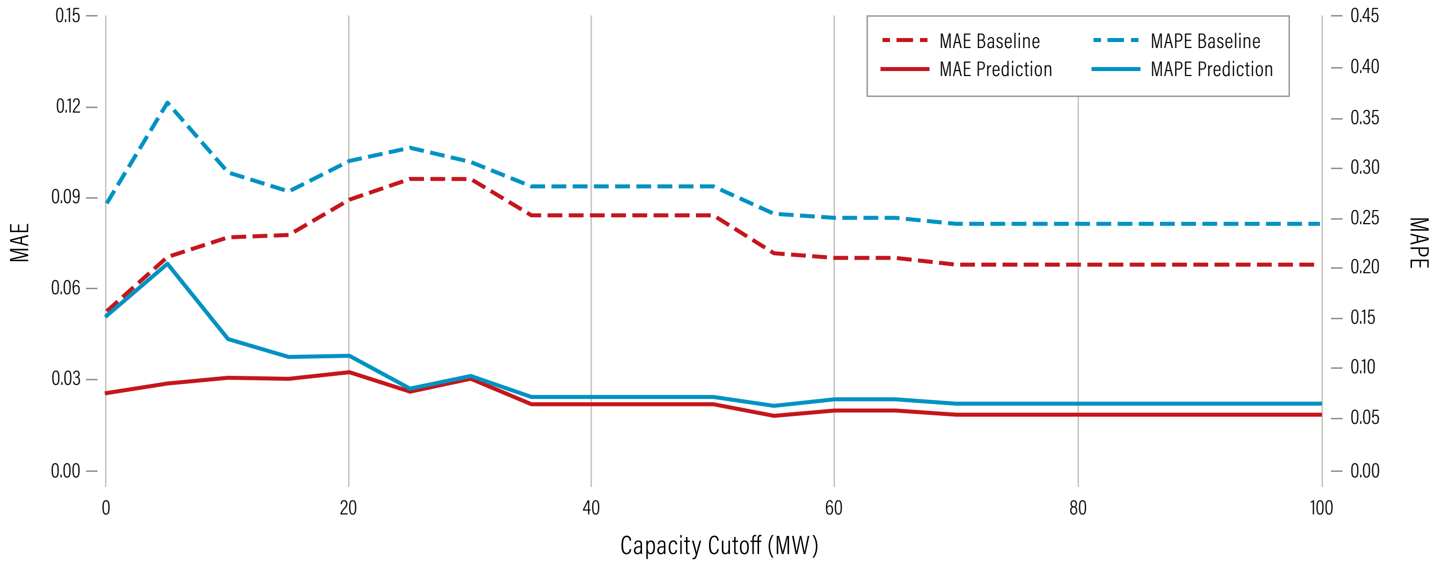
Source: Authors.

Figure 13 | Reported versus Estimated Capacity Factor, Baseline versus Proposed Model



Source: Authors.

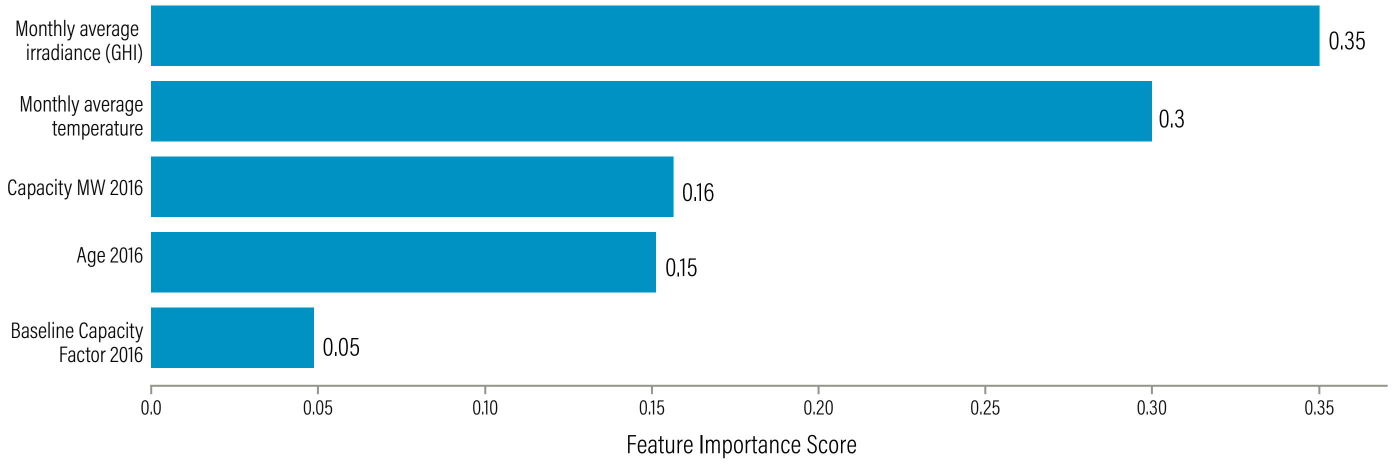
Figure 14 | Solar Model Accuracy Analysis by Plant Size



Notes: MAE stands for mean absolute error and MAPE for mean absolute percentage error. MW stands for megawatt.

Source: Authors.

Figure 15 | Relative Importance of Predictors in the Solar Model



Notes: Relative importance scores do not sum exactly to 1 due to rounding. MW stands for megawatt.

Source: Authors..

Table 10 | Test Scores for Yearly Capacity Factor of Solar Plants Excluding Age (2016)

	PROPOSED MODEL (WITH AGE)	PROPOSED MODEL (NO AGE)
MAE	0.026	0.027
MAPE	0.153	0.162

Notes: MAE stands for mean absolute error and MAPE for mean absolute percentage error.

Source: Authors.

The commissioning year of solar plants is rarely reported in the global solar PV information we have access to, which does not allow us to build the age variable. Since we will use the model to estimate plant-level generation globally, we also develop a version that does not rely on age as a predictor.

Table 10 compares the scores on the test set to the scores on the test set for the model that includes age (reported in Table 9 in the "proposed model" column). The model without age has slightly higher errors, with the mean absolute percentage error increasing from 15 percent to 16 percent. When commissioning year, and therefore age, is available, we use the model to estimate annual generation. When commissioning year is not available, the model without age is still a useful tool to estimate annual generation. Appendix C compares the scatterplot of reported versus estimated generation for the model without age.

7. HYDROPOWER

Hydroelectricity plants include large plants with dams, smaller run-of-river plants, and pumped storage. This analysis excludes generation by pumped storage hydropower (PSH) plants, as they are a storage rather than a generation technology. Annual generation for PSH plants is negative: It takes more electricity to pump the water to an upper reservoir than the amount generated when the water runs back down through the turbine. The stored electricity can be accessed quickly and used to meet utility or regional peak demand. Pumped storage plants are undoubtedly important components of some electricity grids, but their operation as essentially electricity arbitrageurs is unlikely to be captured with any certainty at the annual level.

7.1 Model Description

Hydroelectricity plants use the water flow to spin the turbines that produce electricity (Figure 16). Large hydroelectric projects have dams and a reservoir. Smaller ones are typically run-of-river plants, which have a weir to pool the water, rather than a dam. The weir allows some of the water to be diverted from the main river course and sent to a turbine (Paish 2002).

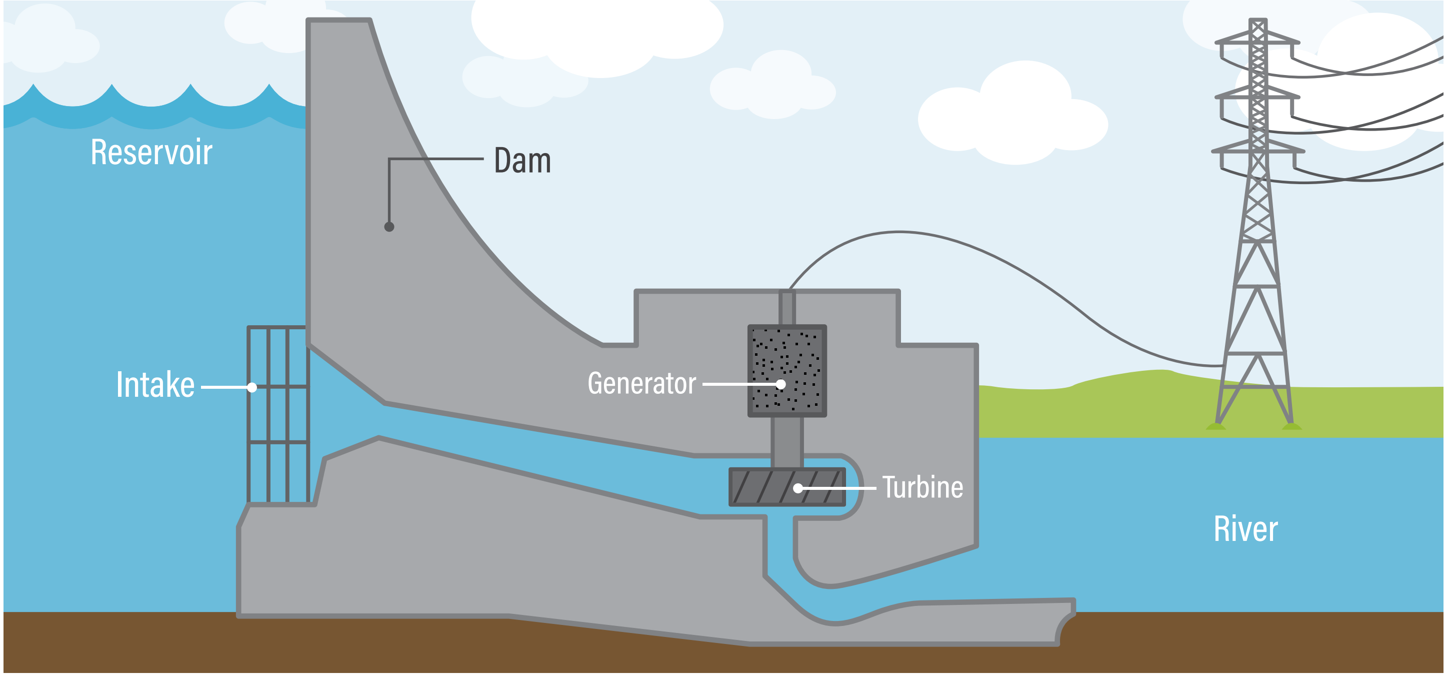
Hydroelectricity generation largely relies on precipitation and runoff accumulated across the drainage area where the plant sits.

We include the following variables to predict annual plant-level generation:

1. Capacity of the plant, which determines the maximum electricity that can be generated at any given time.
2. Average runoff (including surface and subsurface runoff) for the power plant site.
3. The size of the river that flows to the reservoir, measured by river order. Smaller orders refer to larger rivers. Order 1 represents the main stem river from sink to source; order 2 refers to all tributaries that flow into a first-order river; order 3 represents all tributaries that flow into a second-order river; and order 0 is used for conglomerates of small coastal watersheds (Lehner and Grill 2013).
4. Annual average capacity factor by country, which incorporates other country- or region-level information that we do not observe or measure directly.

Day-to-day operations of hydroelectricity plants are typically determined by full-system and legal requirements, including environmental constraints (Niu and

Figure 16 | **Simplified Hydroelectricity System**



Source: U.S. Environmental Protection Agency. 2017. "Water Energy." Web Archive. Last updated May 9. https://archive.epa.gov/climatechange/kids/solutions/technologies/water.html.

Insley 2013). Over a whole year, day-to-day requirements become less important and the amount of runoff becomes a major determinant of generation (Kao et al. 2015).

We include surface and subsurface runoff separately in the model, as surface runoff represents a fast response to precipitation events, whereas subsurface, or base flow, runoff has a lower response to precipitation (Kao et al. 2015).

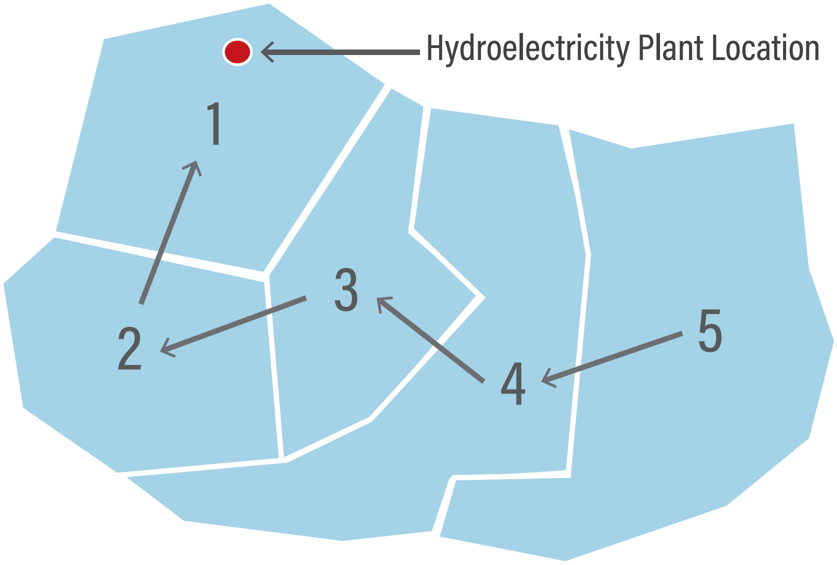
Simply measuring runoff at the reservoir location provides limited information, as water accumulates across a drainage area. We therefore sum measurements across the relevant drainage area as a predictor of generation, following Kao et al. (2015).

The ERA5 global climate data, coupled with HydroBASINS, are used to measure the drainage area and aggregate the relevant meteorological data within the area. HydroBASINS is a series of polygon layers that depicts watershed and sub-basin boundaries at a global scale. Each polygon has a unique ID and links to its upstream polygon. Using the plant coordinates, we locate the polygon where a particular plant sits (e.g., polygon 1 in Figure 17), and then backtrack all upstream polygons to identify the whole drainage area. The arrows across polygons represent the actual direction of water flow and accumulation. For every polygon, HydroBASINS identifies the upstream polygon, if any. By looking up the upstream polygon of 2, we find 3. We continue this search until we reach a polygon that doesn't have any upstream polygon (polygon 5 in this case). The drainage area of this plant is then defined as the ensemble of all polygons found in this process.

The ERA5 global climate data can be aggregated across the drainage area identified with the HydroBASINS data. Total runoff values are added as predictors to the model.

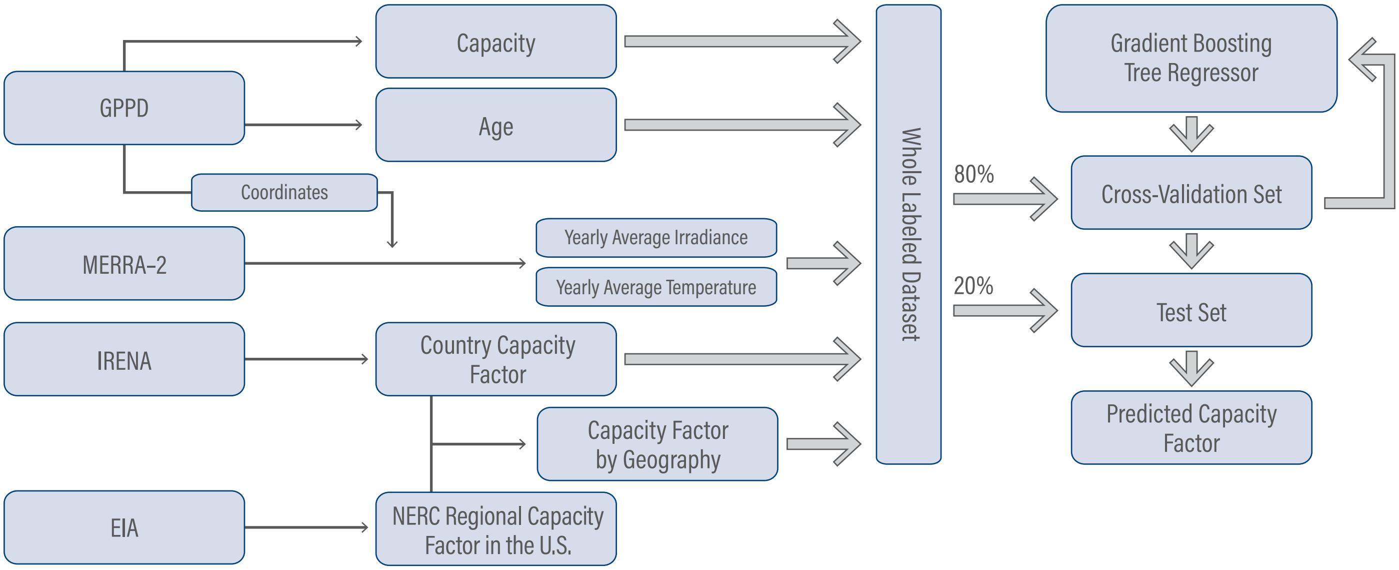
We also include capacity of the plant from GPPD, and the annual country/region capacity factor from the IRENA and EIA reports. Age is not included, since the correlation between age and capacity factor was found to be weak. Figure 18 depicts the structure of the model and the workflow from input data to estimation.

Figure 17 | **Drainage Area Delineation Process**



Source: Authors.

Figure 18 | **Workflow for Hydro Plant Generation Estimation Model**


GPPD
Capacity
Age
Coordinates
MERRA-2
Yearly Average Irradiance
Yearly Average Temperature
IRENA
Country Capacity Factor
Capacity Factor by Geography
EIA
NERC Regional Capacity Factor in the U.S.
Whole Labeled Dataset
80%
20%
Gradient Boosting Tree Regressor
Cross-Validation Set
Test Set
Predicted Capacity Factor

Notes: ERA5 is a database maintained by the European Centre for Medium-Range Weather Forecasts. GPPD stands for Global Power Plant Database, NERC for North American Electric Reliability Corporation, IRENA for International Renewable Energy Agency, and EIA for Energy Information Administration.

Source: Authors.

7.2 Data

ERA5

Measurements for the runoff variables are provided by ERA5. ERA5 is the fifth-generation ECMWF reanalysis for the global climate and weather, with data starting in 1979 to within three months of real time. The data resolution is 0.25° by 0.25° (about 31 km by 31 km near the equator). ERA5 provides similar information as MERRA-2, but with higher grid resolutions, which help us assign appropriate values to the drainage area of each plant.

Table 11 shows which ERA5 database variables we collect as inputs for the machine learning model.

GPPD

GPPD contains information on 3,736 hydroelectricity power plants commissioned before 2016 across the world, of which 1,762 include reported generation, as shown in Table 12. Hydroelectricity plants in GPPD come from multiple regions, including North America, South America, Europe, and Asia. North America and Asia are the major sources of observations with reported generation, accounting for 77 percent and 18 percent of all hydro reported generation, respectively.

Table 11 | **ERA5 Field Code Map**

CODE IN ERA5	FEATURE IN MACHINE LEARNING MODEL
SRO	Surface Runoff
SSRO	Subsurface Runoff
ORDER	Order of Related River

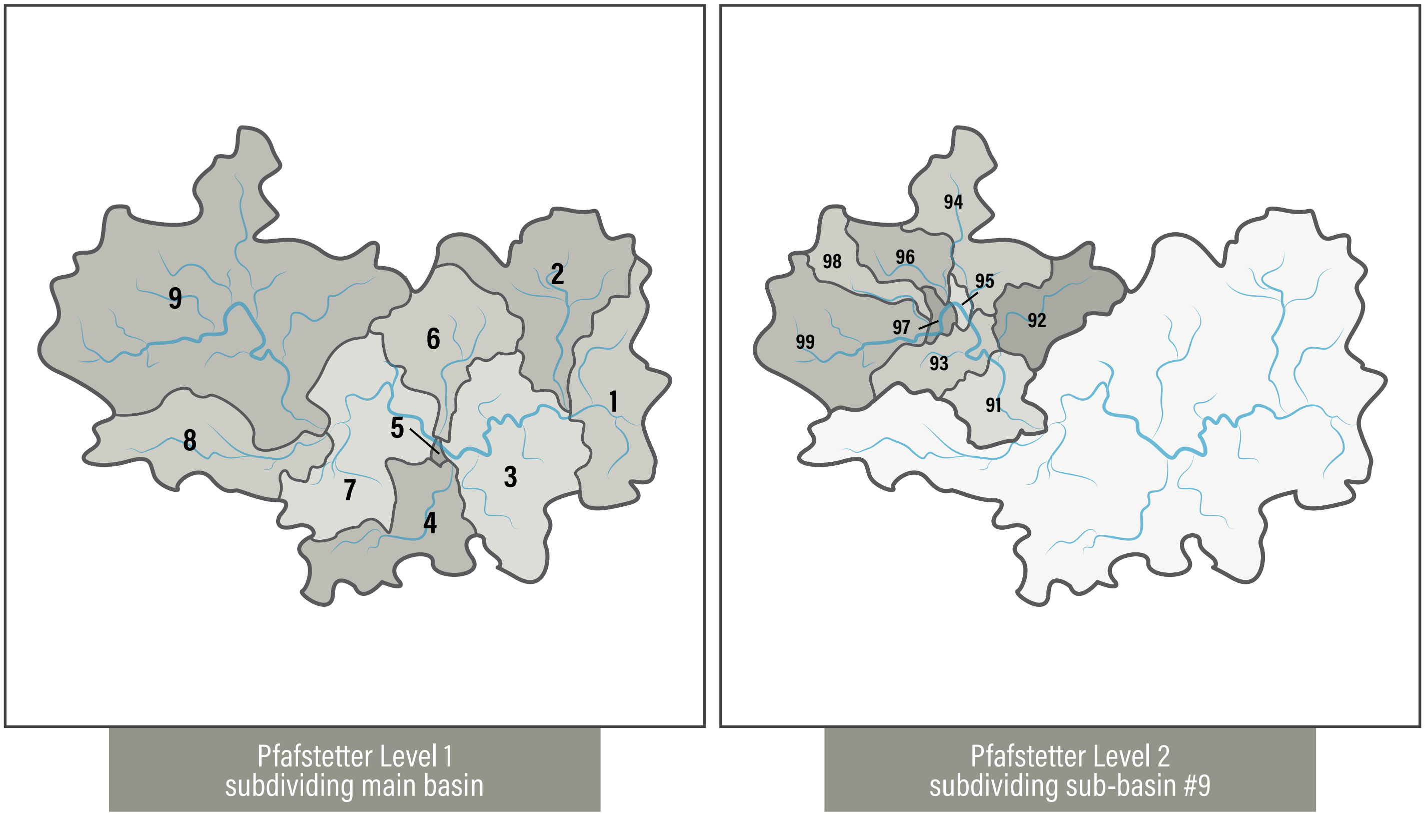
Source: Authors and European Centre for Medium-Range Weather Forecasts.

Table 12 | **Distribution of Hydro Plants and Hydro Plants with Reported Generation (2016)**

	NUMBER OF HYDROELECTRICITY PLANTS BY REGION	HYDROELECTRICITY PLANTS BY REGION (% OF WORLD PLANTS)	NUMBER OF HYDROELECTRICITY PLANTS WITH REPORTED GENERATION DATA (PLANTS THAT CAN BE USED FOR TRAINING)	HYDROELECTRICITY PLANTS WITH REPORTED GENERATION DATA BY REGION (% OF WORLD PLANTS WITH REPORTED GENERATION)
NORTH AMERICA	1,474	39.5%	1,364	77.4%
SOUTH AMERICA	633	16.9%	0	0.0%
EUROPE	841	22.5%	62	3.5%
AFRICA	80	2.1%	25	1.4%
ASIA	690	18.5%	311	17.7%
AUSTRALIA/OCEANIA	18	0.5%	0	0.0%
TOTAL	**3,736**	**100%**	**1,762**	**100%**

Source: Authors' elaboration of Global Power Plant Database data.

Figure 19 | **Different Levels in HydroBASINS**



Pfafstetter Level 1
subdividing main basin

Pfafstetter Level 2
subdividing sub-basin #9

Note: Pfafstetter coding divides each basin into nine smaller ones, starting with Level 1 for the main basin (Verdin and Verdin 1999).

Source: Lehner and Grill 2013.

We exclude 262 plants that are associated with basins small enough that ERA5 data have no grid cell center that falls within the basin. That narrows down the sample size to 1,500. Four plants have a capacity factor that differs from the mean by more than three standard deviations, and we eliminate them from the analysis, leaving 1,496 observations.

HydroBASINS

HydroBASINS provides 12 possible basin levels, from more general to more detailed, as depicted in Figure 19. We found that the correlation between runoff and generation is strongest when runoff is measured at level 12, which is the highest and most detailed level.

7.3 Model Evaluation

Based on the holdout set of 299 observations, Table 13 and Figure 20 suggest that the proposed hydro model improves accuracy compared with the assumption that all hydro plants in a country have the same capacity factor. The estimate remains noisy, with a percentage error for annual plant-level generation of about 46 percent.

The validation scores of each cross-validation fold can be seen in Appendix B.

Although the MAPE drops sharply as the capacity cutoff increases (see Figure 21), the MAE doesn't display a clearly increasing or decreasing pattern.

The relative importance of each feature among all relevant features is shown in Figure 22. The aggregated surface and subsurface runoff and the capacity of the plant are relatively more important than other predictors.

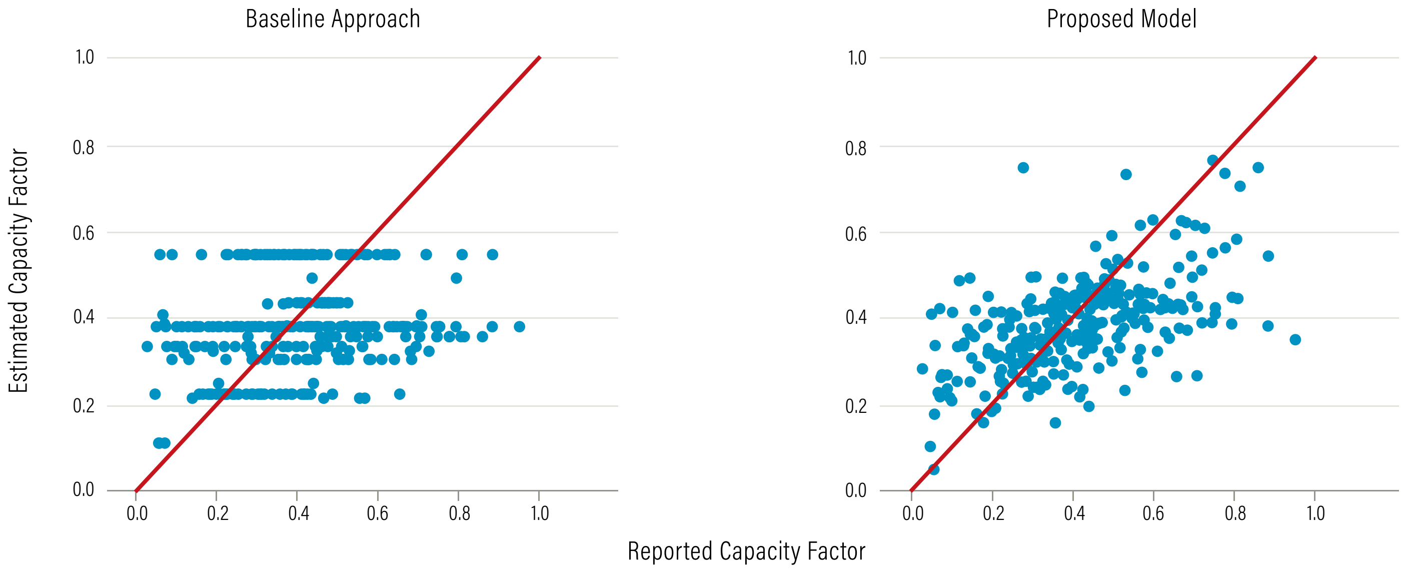
Table 13 | **Test Scores for Yearly Capacity Factor of Hydroelectricity Plants (2016)**

	BASELINE	PROPOSED MODEL
MAE	0.156	0.117
MAPE	0.596	0.459

Notes: MAE stands for mean absolute error and MAPE for mean absolute percentage error.

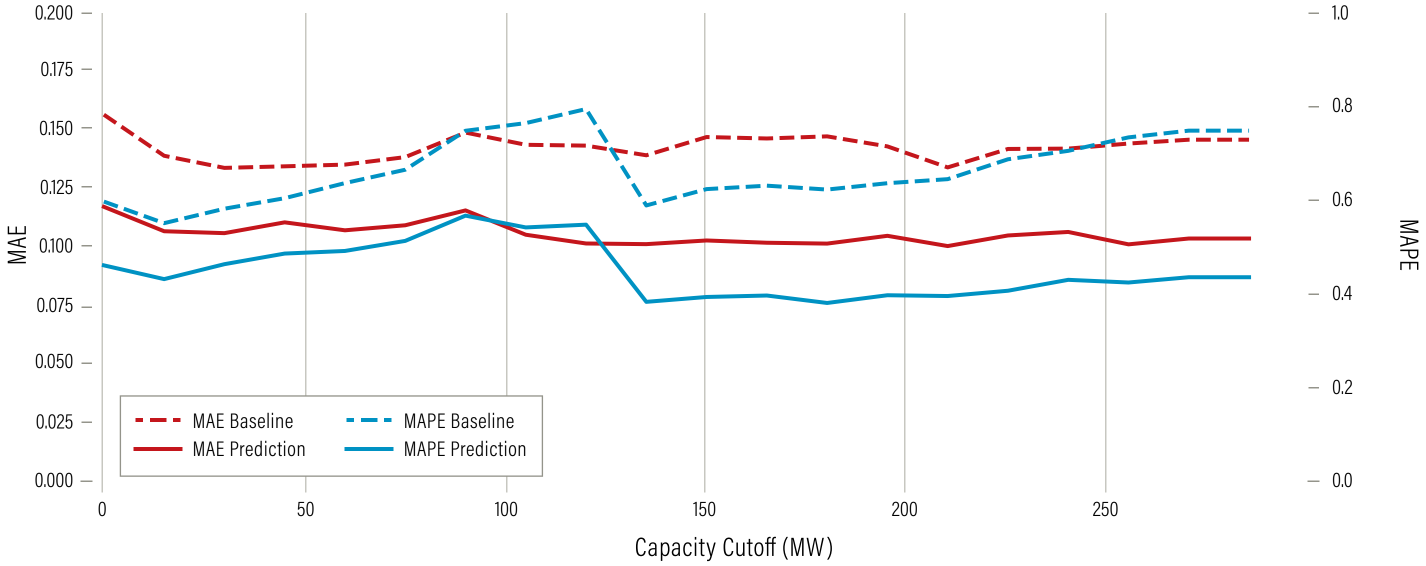
Source: Authors.

Figure 20 | **Reported versus Estimated Capacity Factor of Hydro Plants, Baseline versus Proposed Model**



Source: Authors.

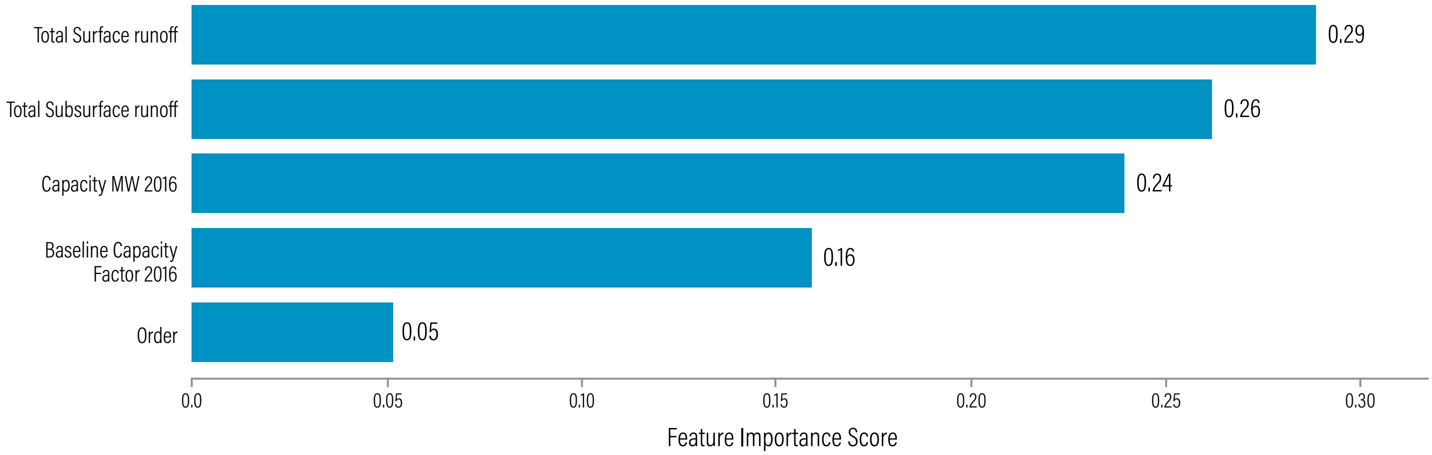
Figure 21 | Hydro Model Accuracy Analysis by Plant Size



Notes: MAE stands for mean absolute error and MAPE for mean absolute percentage error. MW stands for megawatt.

Source: Authors.

Figure 22 | Relative Importance of Predictors in the Hydropower Model



Note: MW stands for megawatt.

Source: Authors.

8. LIMITATIONS

We are not able to factor in all potentially relevant predictors. Data availability is a major constraint on producing more complex models. Our goal is to estimate annual generation for power plants located across the world, so we use predictors that are available globally. As the Global Power Plant Database continues to develop and more power plant attributes are discovered and recorded, the methods described here may co-evolve with the database. Plant technology is a necessary technical attribute to provide more context about the plant than fuel type is able to do alone. Efficiency, which may be tightly coupled to plant technology, is an additional technical metric that would likely provide an improvement in predictive power for annual generation.

The estimates from this model can be applied to plants only when we have the relevant technical and environmental characteristics (e.g., technology type for natural gas plants, runoff information for hydro plants). While an improvement over baseline GPPD estimates, this means that approximately 18 percent of the plants do not contain the requisite parameters and still rely on average capacity factors as their best estimate. We hope to overcome this limitation as we continue improving the information on the technical characteristics of each plant through additional data collection and contributions.

Most of the models were fit to training sets that are dominated by data from the United States, while the goal is to estimate generation for power plants in countries or regions outside of the U.S. The models included average annual capacity factor by country by fuel, which helps to improve generalization, but we still cannot fully avoid the risks of the model overfitting to the training data and therefore not being generalizable. In more technical terms, we cannot ensure that the data used for training and in the application of the model have similar distributions. Further data collection and advancements in openness around electricity data are needed to mitigate this concern.

While signaling an improvement over the baseline, some of the estimates are still quite noisy. This includes the estimates for natural gas generation, where the average error is 170 percent of the mean generation. Further analysis is needed to identify other features that are available globally and can help improve the estimation. Because of the flexibility of gas plants to meet less predictable variations in supply-demand balance, including additional system-level variables may be warranted.

The models proposed here are applied to a power plant database that is incomplete and therefore contains an imperfect picture of global electricity generation capacity. The generation of each power plant is estimated independently of other power plants and there is no guarantee that aggregate generation (i.e., the sum of generation for all power plants in a country) matches independently provided measures. The utility of these estimates is in providing variable capacity factors for plants within a country. As more power plants are identified, mapped, and incorporated into the database, the methods used here can be directly applied to these plants and yield annual generation estimates.

9. CONCLUSION AND FUTURE WORK

This project evaluated the accuracy of annual generation estimations for wind, solar, hydro, and natural gas plants based on geographic, environmental, and system variables. The methodology focuses on estimating each plant's deviation from the average generation of plants of its type based on detailed information on plant-level and geo-located environmental factors.

The analysis found that plant-level annual generation for wind and solar can be estimated fairly precisely given information on how much wind blows and sun shines at the plant location. The low penetration of intermittent renewables until recently means that system constraints have been limited in practice: When wind and solar resources are available, they are generally dispatched.

Annual generation from hydropower plants can be predicted less accurately, and depends significantly on water runoff. Natural gas plants were the most difficult to predict annual generation for, highlighting how system factors, for which we have limited information, are important in determining when and how they are dispatched.

There are a few areas that can be explored in future work to improve generation estimates. The first applies techniques that can improve method robustness when the target data are not necessarily similar to the training data (Pan and Yang 2010; Jean et al. 2016) or when unlabeled samples largely outnumber labeled samples (Kostopoulos et al. 2018).

A second area involves searching for data that are more closely related to generation. One option for thermal plants includes infrared information that identifies whether a plant is on or not, similar to the generation model adopted by the Carbon Tracker for coal plants (Gray et al. 2018). This would allow the models to achieve higher accuracy when system factors are important, but would rely on the existence of appropriate high-frequency information.

Finally, we continuously add new training data as they become available. While the additional data will not affect our estimation methods, they will improve the predictive models. They will also help us develop time series, either by adding annual generation estimations for different years or developing higher frequency data. In addition, we will update technical information for the plants as it becomes available, increasing the number of plants for which we can adopt the methods laid out in this technical note.

A link to the code is provided in Appendix E.

APPENDIX A. COUNTRIES WITH REPORTED PLANT-LEVEL GENERATION

Table A1 | **Countries with Reported Plant-Level Generation (2016)**

COUNTRY	NUMBER OF PLANTS WITH REPORTED GENERATION
United States	7,944
India	427
Australia	248
Vietnam*	156
Germany	124
Italy	116
Spain	81
Guatemala*	72
France	58
Portugal	42
Netherlands	42
Austria	40
Poland	37
Sweden	26
Finland	24
Romania	23
Kenya*	21
Morocco*	18
Denmark	12
Belgium	12
Czech Republic	9
Hungary	9
Lithuania	6
United Kingdom	5
Estonia	4
Slovenia	4
Latvia*	2
Montenegro*	2
Ireland	2
Egypt*	1
Luxembourg	1

Note: * Countries that are included in the "Other" category in Table 1.

Sources: See Table 1.

APPENDIX B. CROSS-VALIDATION SCORES OF EACH MODEL

This section reports the mean absolute error (MAE) and the mean absolute percentage error (MAPE) for the validation phase of each model.

Table B1 | Natural Gas Cross-Validation Error Scores by Fold

	FOLD 1	FOLD 2	FOLD 3	FOLD 4	FOLD 5	FOLD 6	FOLD 7	FOLD 8	FOLD 9	FOLD 10
MAE	0.135	0.151	0.131	0.148	0.153	0.176	0.136	0.138	0.141	0.136
MAPE	1.495	1.729	1.879	1.581	1.371	2.164	1.368	1.977	2.301	1.850

Source: Authors.

Table B2 | Wind Model Cross-Validation Error Scores by Fold

	FOLD 1	FOLD 2	FOLD 3	FOLD 4	FOLD 5	FOLD 6	FOLD 7	FOLD 8	FOLD 9	FOLD 10
MAE	0.053	0.046	0.047	0.043	0.047	0.043	0.044	0.050	0.051	0.046
MAPE	0.233	0.167	0.191	0.194	0.191	0.154	0.178	0.212	0.192	0.194

Source: Authors.

Table B3 | Solar Model Cross-Validation Error Scores by Fold

	FOLD 1	FOLD 2	FOLD 3	FOLD 4	FOLD 5	FOLD 6	FOLD 7	FOLD 8	FOLD 9	FOLD 10
MAE	0.025	0.031	0.027	0.030	0.030	0.024	0.030	0.027	0.031	0.034
MAPE	0.147	0.156	0.165	0.131	0.168	0.118	0.183	0.161	0.194	0.226

Source: Authors.

Table B4 | Hydropower Model Cross-Validation Error Scores by Fold

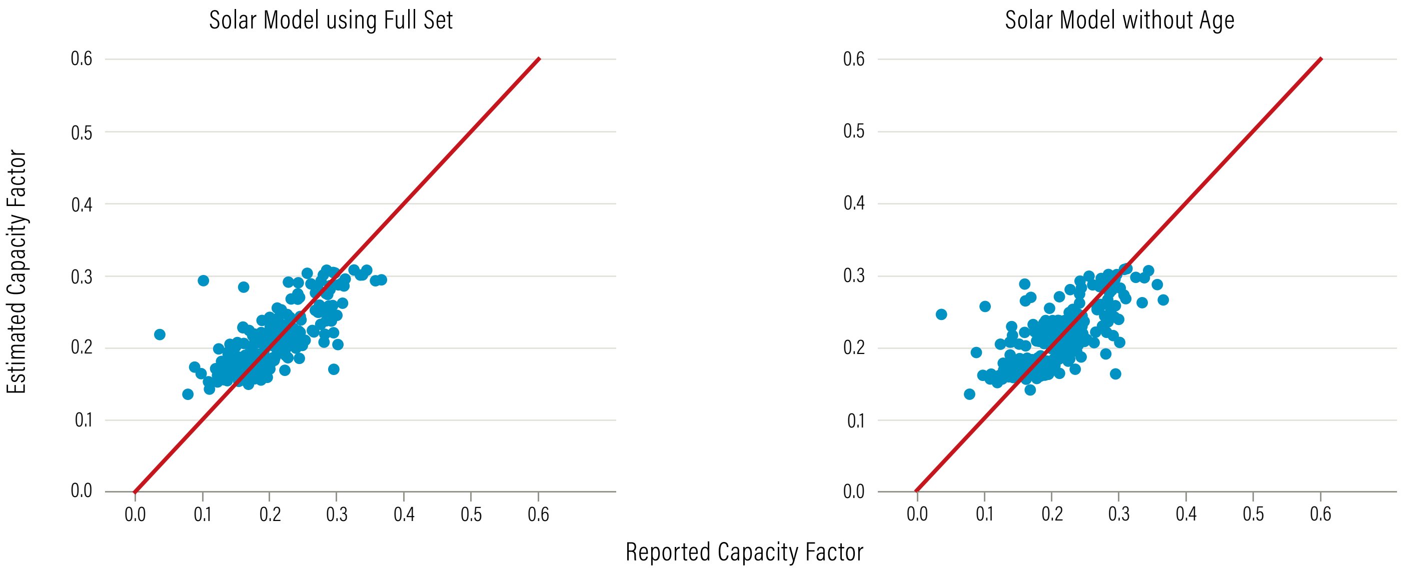
	FOLD 1	FOLD 2	FOLD 3	FOLD 4	FOLD 5	FOLD 6	FOLD 7	FOLD 8	FOLD 9	FOLD 10
MAE	0.122	0.116	0.137	0.113	0.127	0.114	0.109	0.136	0.117	0.117
MAPE	1.130	9.361	0.868	1.055	6.658	0.865	0.578	1.250	0.946	0.428

Source: Authors.

APPENDIX C. SOLAR MODEL WITHOUT AGE

Figure C1 compares the scatterplot of estimated versus reported generation for the solar model when age is available (left) to that when age is not available (right) for the test sample. The closer the points are to the 45-degree line, the more accurate the estimate. While age should be included when it is available, using the model without age is still preferable to using the naïve baseline model.

Figure C1 | Reported versus Estimated Solar Capacity Factor, Proposed Model with and without Age



Source: Authors.

APPENDIX D. EXTERNAL DATA SOURCES

1. HydroBASINS. Data available at www.hydrosheds.org.
2. U.S. Energy Information Administration. "Form EIA-923." Data available at https://www.eia.gov/electricity/data/eia923/.
3. U.S. Energy Information Administration. "Form EIA-860." Data available at https://www.eia.gov/electricity/data/eia860/.
4. International Energy Agency Statistics on Electricity Generation by Country and Fuel. Data available at https://www.iea.org/statistics/?country=WORLD&year=2016&category=Electricity&indicator=ElecGenByFuel&mode=table&dataTable=ELECTRICITYANDHEAT.
5. International Energy Agency, OECD—Net Electrical Capacity. Data available at https://www.oecd-ilibrary.org/energy/data/iea-electricity-information-statistics/oecd-net-electrical-capacity_data-00460-en.
6. Central Electricity Authority of India. Data available at http://www.cea.nic.in/reports.html.
7. National Greenhouse and Energy Reporting of Australia. Data available at http://www.cleanenergyregulator.gov.au/NGER/National%20greenhouse%20and%20energy%20reporting%20data/electricity-sector-emissions-and-generation-data.
8. European Network of Transmission System Operators for Electricity (ENTSO-E). Data available at https://transparency.entsoe.eu/generation/r2/actualGenerationPerGenerationUnit/show.
9. World Electric Power Plants Database, version March 2017. Current version available at https://www.spglobal.com/platts/en/products-services/electric-power/world-electric-power-plants-database.
10. European Centre for Medium-Range Weather Forecasts, Reanalysis 5 (ERA5). Data available at https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5.
11. U.S. National Aeronautics and Space Administration, Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2). Data available at https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/data_access/.
12. International Renewable Energy Agency statistics on capacity and generation. Data available at https://www.irena.org/Statistics/View-Data-by-Topic/Capacity-and-Generation/Statistics-Time-Series.

APPENDIX E. CODE

The code is hosted on GitHub and can be accessed at https://github.com/wri/gppd-ai4earth-api.

REFERENCES

Byers, L., J. Friedrich, R. Hennig, A. Kressig, X. Li, C. McCormick, and L. Malaguzzi Valeri. 2018. "A Global Database of Power Plants." Washington, DC: World Resources Institute. https://www.wri.org/publication/global-power-plant-database.

Darrow, K., R. Tidball, J. Wang, and A. Hampson. 2017. "Catalog of CHP Technology." Washington, DC: U.S. Environmental Protection Agency Combined Heat and Power Partnership Program.

Dubey, S., J.N. Sarvaiya, and B. Seshadri. 2013. "Temperature Dependent Photovoltaic (PV) Efficiency and Its Effect on PV Production in the World—a Review." *Energy Procedia*, PV Asia Pacific Conference 2012, 33 (January): 311–21. doi:10.1016/j.egypro.2013.05.072.

EirGrid. 2016. "Generation Capacity Statement 2016–2025." http://www.eirgridgroup.com/site-files/library/EirGrid/Generation_Capacity_Statement_20162025_FINAL.pdf.

Elith, J., J.R. Leathwick, and T. Hastie. 2008. "A Working Guide to Boosted Regression Trees." *Journal of Animal Ecology* 77 (4): 802–13. doi:10.1111/j.1365-2656.2008.01390.x.

Gray, M., L. Watson, S. Ljungwaldh, and E. Morris. 2018. "Nowhere to Hide: Using Satellite Imagery to Estimate the Utilisation of Fossil Fuel Power Plants." Carbon Tracker. https://www.carbontracker.org/reports/nowhere-to-hide/.

IEA and NEA (International Energy Agency and Nuclear Energy Agency). 2015. *Projected Costs of Generating Electricity*. 2015 ed. Paris: IEA.

Iglewicz, B., and D. Hoaglin. 1993. *How to Detect and Handle Outliers*. Milwaukee, WI: American Society for Quality Control.

Jean, N., M. Burke, M. Xie, W.M. Davis, D.B. Lobell, and S. Ermon. 2016. "Combining Satellite Imagery and Machine Learning to Predict Poverty." *Science* 353 (6301): 790–94. doi: 10.1126/science.aaf7894.

Jordan, D.C., and S.R. Kurtz. 2013. "Photovoltaic Degradation Rates—An Analytical Review." *Progress in Photovoltaics: Research and Applications* 21 (1): 12–29. doi: 10.1002/pip.1182.

Kanellopoulos, K., M. De Felice, I. Hidalgo, A. Bocin, and A. Uihlein. 2019. "The Joint Research Centre Power Plant Database (JRC-PPDB)—Version 0.9." JRC Scientific Information Systems and Databases, Publications Office of the European Union. doi:10.2760/5281.

Kao, S.-C., M.J. Sale, M. Ashfaq, R. Uria Martinez, D.P. Kaiser, Y. Wei, and N.S. Diffenbaugh. 2015. "Projecting Changes in Annual Hydropower Generation Using Regional Runoff Data: An Assessment of the United States Federal Hydropower Plants." *Energy* 80: (February) 239–50. doi: 10.1016/j.energy.2014.11.066.

Kostopoulos, G., S. Karlos, S. Kotsiantis, and O. Ragos. 2018. "Semi-supervised Regression: A Recent Review." *Journal of Intelligent & Fuzzy Systems* 35 (2): 1483–1500. doi:10.3233/JIFS-169689.

Lehner, B., and G. Grill. 2013. "Global River Hydrography and Network Routing: Baseline Data and New Approaches to Study the World's Large River Systems." *Hydrological Processes* 27 (15): 2171–86. doi:10.1002/hyp.9740.

Mosshammer, S. 2016. "Assessing the Validity of MERRA Reanalysis Data for Simulation of Wind Power Production." Master's Thesis, University of Natural Resources and Life Sciences, Vienna. https://homepage.boku.ac.at/jschmidt/TOOLS/Endfassung_masterarbeit_sm%20030716_I.pdf.

Niu, S., and M. Insley. 2013. "On the Economics of Ramping Rate Restrictions at Hydro Power Plants: Balancing Profitability and Environmental Costs." *Energy Economics* 39: (September) 39–52. doi:10.1016/j.eneco.2013.04.002.

Olden, J.D., J.J. Lawler, and N.L. Poff. 2008. "Machine Learning Methods without Tears: A Primer for Ecologists." *Quarterly Review of Biology* 83 (2): 171–93. doi:10.1086/587826.

Paish, O. 2002. "Small Hydro Power: Technology and Current Status." *Renewable and Sustainable Energy Reviews* 6 (6): 537–56. doi:10.1016/S1364-0321(02)00006-0.

Pan, S.J., and Q. Yang. 2010. "A Survey on Transfer Learning." *IEEE Transactions on Knowledge and Data Engineering* 22 (10): 1345–59. doi:10.1109/TKDE.2009.191.

Shulga, D. 2018. "5 Reasons Why You Should Use Cross-Validation in Your Data Science Projects." *Medium* (blog). September 27. https://towardsdatascience.com/5-reasons-why-you-should-use-cross-validation-in-your-data-science-project-8163311a1e79.

Ummel, K. 2012. "CARMA Revisited: An Updated Database of Carbon Dioxide Emissions from Power Plants Worldwide." *Center for Global Development, Working Paper 304.* August 23. doi:10.2139/ssrn.2226505.

Varian, H.R. 2014. "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives* 28 (2): 3–28. doi:10.1257/jep.28.2.3.

Verdin, K.L., and J.P. Verdin. 1999. "A Topological System for Delineation and Codification of the Earth's River Basins." *Journal of Hydrology* 218 (1–2): 1–12.

ACKNOWLEDGMENTS

Funding for this research was provided by a joint WRI and Microsoft AI for Earth grant. This work builds upon the WRI Global Power Plant Database. Venkata Dusi contributed to the project during his time as a WRI intern. The authors received helpful comments from WRI and external reviewers. WRI reviewers included David Rich, John Brandt, Emily Cassidy, Deepak Krishnan, Dimitrios Mentis, Robi Robichaud, Samantha López, and Gregory Taff. External reviewers included Paul Deane, research fellow at University College Cork, and Kevin Ummel, president of Greenspace Analytics and research affiliate at the University of Pennsylvania. All remaining errors are the authors'.

We are pleased to acknowledge our institutional strategic partners, who provide core funding to WRI: Netherlands Ministry of Foreign Affairs, Royal Danish Ministry of Foreign Affairs, and Swedish International Development Cooperation Agency.

ABOUT THE AUTHORS

Luotian Yin is a research assistant with WRI's Climate Program.
Contact: terry.yin@wri.org

Logan Byers is a research assistant with 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

Laura Malaguzzi Valeri is deputy 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.