CARMA Revisited: An Updated Database of Carbon Dioxide Emissions from Power Plants Worldwide

Kevin Ummel

Abstract

The Carbon Monitoring for Action (CARMA) database provides information about the carbon dioxide emissions, electricity production, corporate ownership, and location of more than 60,000 power plants in over 200 countries. Originally launched in 2007, CARMA is provided freely to the public at www.carma.org and remains the only comprehensive data source of its kind. This paper documents the methodology underpinning CARMA v3.0, released in July, 2012. Comparison of CARMA model output with reported data highlights the general difficulty of precisely predicting annual electricity generation for a given plant and year. Estimating the rate at which a plant emits CO2 (per unit of electricity generated) generally faces fewer obstacles. Ultimately, greater disclosure of plant-specific data is needed to overcome these limitations, particularly in major emitting countries like China, Russia, and Japan. For any given plant in CARMA v3.0, it is estimated that the reported value is within 20 percent of the actual value in 85 percent of cases for CO2 intensity, 75 percent for annual CO2 emissions, and 45 percent for annual electricity generation. CARMA's prediction models are shown to offer significantly better estimates than more naïve approaches to estimating plant-specific performance.

CARMA v3.0 also includes a significant upgrade in the quantity and quality of geographic data, including standardized geopolitical information for nearly all facilities. High-precision coordinates are now available for 10 percent of plants (covering 30% of global CO2 emissions) and approximate coordinates are available for an additional 70 percent of facilities. The new version also lays the technical groundwork for future expansion to green house gases other than CO2, offering potential improvement in continental-scale modeling of the environmental and health consequences of conventional pollutants.

JEL Codes: Q50, Q53, Q54, Q55

Keywords: CARMA, carbon monitoring, greenhouse gases, global warming.

Center for Global Development
www.cgdev.org

Working Paper 304
August 2012

CARMA Revisited: An Updated Database of Carbon Dioxide Emissions from Power Plants Worldwide

Kevin Ummel
CARMA Project Manager

CGD is grateful to its funders and board of directors for support of this work.

Kevin Ummel . 2012. "CARMA Revisited: An Updated Database of Carbon Dioxide Emissions from Power Plants Worldwide." CGD Working Paper 304. Washington, D.C.: Center for Global Development.
http://www.cgdev.org/content/publications/detail/1426429

Center for Global Development
1800 Massachusetts Ave., NW
Washington, DC 20036

202.416.4000
(f) 202.416.4050

www.cgdev.org

Contents

Foreword

The Center for Global Development launched the Carbon Monitoring for Action (CARMA) database (www.carma.org) in 2007 to make available to the public data on the carbon dioxide emissions of more than 50,000 power plants. The motivations were, and still are, to lay the informational groundwork for any market-based system of emissions regulation, as well as to apply the pressure of public disclosure on power companies to reduce emissions. It remains the only comprehensive database of its kind.

In this paper, former CGD research assistant Kevin Ummel explains the most recent updates to the database, CARMA v3.0. CARMA now includes information from public databases from India, Canada, South Africa, the United States, and the European Union, which account for more than a third of global power-sector CO2 emissions and a quarter of global electricity generation. Version 3.0 quantifies the maximum likely errors in measurement and includes more and better information on the location of power plants. It also begins the technical work for future versions to monitor other greenhouse gases such as methane, sulfur oxides, and nitrogen oxides.

David Roodman
Senior Fellow
Center for Global Development

1. Introduction

While national carbon dioxide emissions are regularly published for most countries, data specific to individual sectors, companies, geographic regions, or facilities are more difficult to obtain – if available at all. This is unfortunate, because disaggregated data are especially useful to educators, policymakers, academics, investors, and environmental activists in need of information about the carbon footprint of particular entities. Aggregate totals reveal general trends, but disaggregated data facilitate specific actions.

Power generation suits the development of detailed data, because it relies heavily on stationary point sources for which information is more likely to be available. The power sector is also a major contributor to global climate change, producing 40% of energy-related CO_2 emissions worldwide (IEA 2011). Governmental efforts to collect, process, and disclose power plant emissions vary widely. Relatively few countries mandate public disclosure of greenhouse gas (GHG) emissions from power plants; still fewer make this information easily accessible. Actors requiring comprehensive, global information (for example, investors allocating resources among power companies in global capital markets) are often simply out of luck, given the unevenness of disclosure efforts worldwide.

The Carbon Monitoring for Action (CARMA) database was created in 2007 to help facilitate the disclosure, consolidation, and public dissemination of information about the CO_2 emissions and electricity generation of individual power plants and companies worldwide. In cases where disclosed data is unavailable, CARMA provides estimates. Wheeler and Ummel (2008) provide a description of the original rationale and methodology. A public version of the CARMA database is made available through the CARMA website (www.carma.org).

Over the past year, the CARMA database has been upgraded to include new data sources and statistical techniques. The latest version, CARMA v3.0, now includes:

- Publicly disclosed plant-level CO_2 emissions for more than 6,200 power plants in the United States, European Union, Canada, India, and South Africa.
- Publicly disclosed plant-level electricity generation data for more than 5,700 power plants in the United States, India, and South Africa, as well as nuclear power plants worldwide.
- Estimates of plant-level CO_2 emissions and electricity generation for an additional ~49,000 power plants in over 200 countries.
- Aggregate generation and emissions data for ~22,000 power companies and utilities.
- Aggregate generation and emissions data for ~13,000 geographic regions (countries, states/provinces, cities, etc.).
- High-resolution geographic coordinates for ~6,200 power plants and approximate coordinates for an additional ~39,000 facilities.
- Analysis of the likely prediction error for facilities with statistically-modeled (i.e. estimated) generation and emissions data.

2. Summary of methodology

This section provides an overview of the CARMA v3.0 data sources and methodology. Technical details can be found in subsequent sections.

CARMA's principal task is to collect and consolidate plant-level generation and emissions data made public by national or international authorities. Table 1 shows the status of known efforts to disclose plant-level CO_2 emissions and/or electricity generation. Links to the original databases are listed in Footnote 1.

These data are made public in various formats and with varying levels of detail, requiring that the data be cleaned and standardized before incorporation into CARMA. Worldwide, only about 10% of CO_2-emitting power plants regularly disclose their emissions to the public. However, since disclosure efforts often focus on larger plants, these facilities are collectively responsible for ~35% of global power sector CO_2 emissions.

Table 1: Summary of power plant CO_2 and electricity generation disclosure databases[1]

	Facility CO_2 emissions?	Facility electricity generation?	Emissions coverage
United States	Yes	Yes	~100%
European Union	Yes	No	63%
Canada	Yes	No	91%
India	Yes	Yes	78%
South Africa	Yes	Yes	96%
IAEA (nuclear units only)	N/A	Yes	~100%

Note: Emissions coverage gives the approximate percentage of total power sector emissions disclosed at the plant level.

[1] The original databases can be found at the following URLs:
Canada: http://www.ec.gc.ca/pdb/ghg/onlineData/dataSearch_e.cfm
European Union: http://www.eea.europa.eu/data-and-maps/data/member-states-reporting-art-7-under-the-european-pollutant-release-and-transfer-register-e-prtr-regulation-4
India: http://www.cea.nic.in/reports/planning/cdm_co2/cdm_co2.htm
International Atomic Energy Agency: http://www.iaea.org/pris/
South Africa: http://www.eskom.co.za/c/article/236/cdm-calculations/
United States: http://205.254.135.7/cneaf/electricity/forms/datamatrix.html *and* http://ampd.epa.gov/ampd/

Australia and New Zealand disclose plant-level power generation (not emissions) through the national grid operator, but these data are not yet incorporated into CARMA. If you are aware of other relevant disclosure databases or efforts, please notify CARMA at: carma@cgdev.org

For facilities where no public data are available, it is necessary to estimate CO_2 emissions and electricity generation. A comprehensive listing of the world's power plants is provided by the commercial World Electric Power Plant (WEPP) database maintained by Platts, Inc.[2] WEPP provides geographic, corporate, and engineering data for individual generating units in over 200 countries. This information is used as the basis for estimating plant performance in the absence of public data.

When necessary, CARMA estimates power plant performance using statistical models fitted to a detailed dataset of U.S. facilities. These models predict key variables like plant capacity factor and heat rate as a function of the plant's size, vintage, technology, and other engineering and operating characteristics. In addition, national generation, heat rate, and CO2 emissions data from the International Energy Agency (IEA) are used to constrain initial model estimates to ensure accurate aggregate totals for the year in question.[3] When IEA national-level data are not available, country-specific generation totals from the U.S. Energy Information Administration (EIA) are used instead.[4] Details of the model fitting and estimation process are described in Section 3.

CARMA v3.0 also includes internal calculation of generation and emissions data for all U.S. power plants (even when disclosed emissions data are not available), relying on detailed datasets from the EIA and Environmental Protection Agency (EPA). This means CARMA is potentially capable of disclosing preliminary U.S. power plant data about 3-6 months after the end of the calendar year. However, disclosure and estimation of international data face greater delays, typically dictated by the release of annual IEA and EIA national-level datasets. Figure 1 illustrates the full CARMA v3.0 processing chain from inputs to final product.

One objective of this paper is to describe the likely error associated with estimated data. CARMA effectively reports two quantities: plant capacity factor (i.e. rate of utilization) and CO_2 intensity (i.e. rate of CO_2 emission per unit electricity). Wheeler and Ummel (2008) note that plant capacity factor is highly variable – even from year to year for the same facility. This variability is driven, in part, by financial, regulatory, and maintenance considerations that are largely unobservable. This makes prediction of electricity generation for a given plant and year the largest source of error within CARMA. Estimation of CO_2 intensity faces fewer obstacles, as this is driven largely by observable fuel and engineering characteristics. Assessment of model skill and analysis of year-to-year variability are provided in Sections 4 and 5. The effect of plant aggregation on prediction error is discussed in Section 6.

After constructing plant-level CO_2 emissions and electricity generation data from a combination of disclosed data and model estimates, geographic data are added. CARMA v3.0 includes substantial advances in geocoding of individual facilities. Specifically, geopolitical data

---

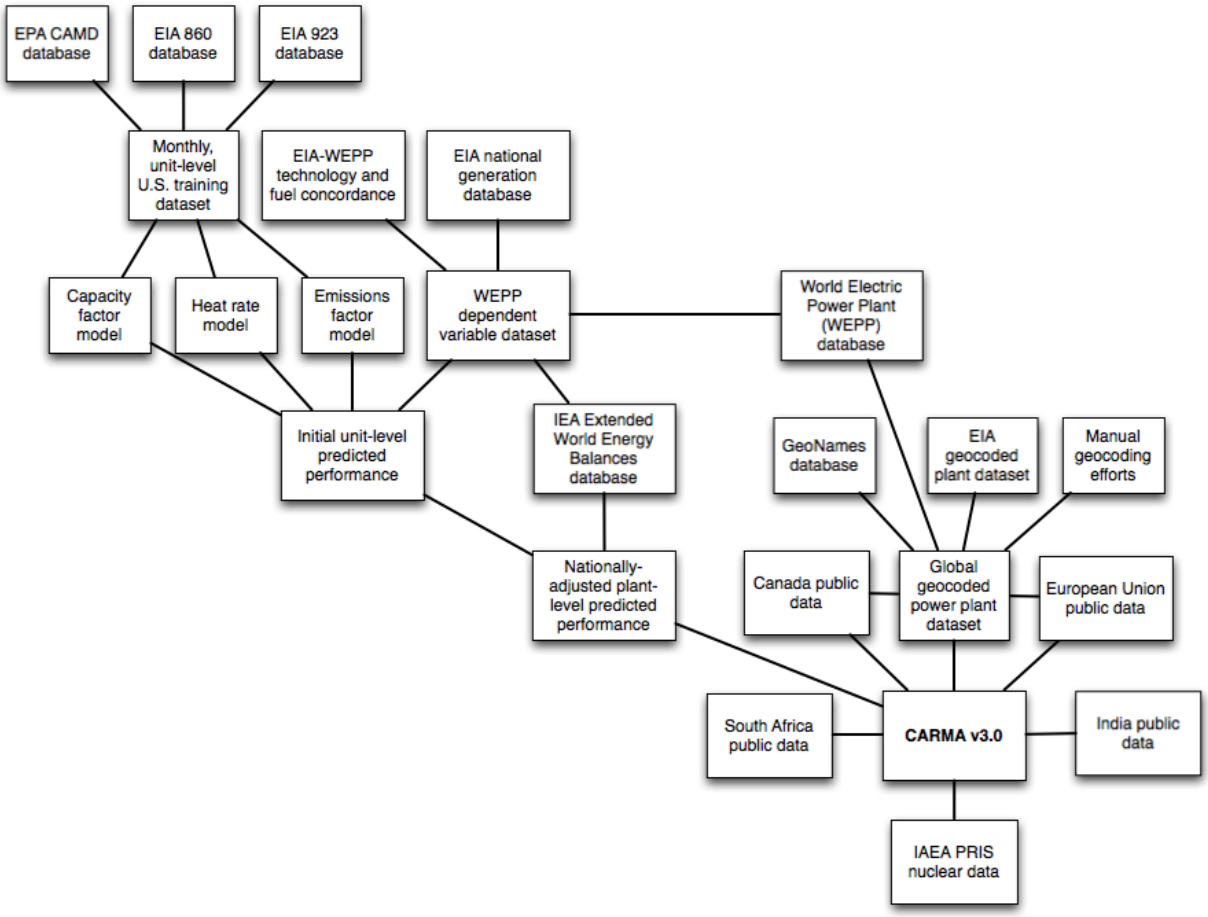
[2]http://www.platts.com/Products/worldelectricpowerplantsdatabase

[3]http://data.iea.org/ieastore/wedproduct.asp?dept_id=101&pf_id=205
http://data.iea.org/ieastore/product.asp?dept_id=101&pf_id=305

[4]http://www.eia.gov/cfapps/ipdbproject/IEDIndex3.cfm?tid=2&pid=2&aid=12

provided by WEPP (city, region, etc.) have been processed with "fuzzy string" algorithms to standardize spellings, extract maximum information, and create concordance with the open-source GeoNames database (www.geonames.org). The final CARMA database includes this information along with approximate plant coordinates for ~39,000 plants and high-resolution coordinates for another ~5,900 (see Section 7).

Figure 1: CARMA v3.0 data processing chain



Each plant in CARMA is also assigned corporate ownership data. CARMA attempts to report the ultimate, primary owner of each facility (i.e. the highest entity in the corporate hierarchy). The data comes primarily from WEPP, which attempts to track ownership relationships and hierarchies in the power sector. The ultimate owner may differ from the local operator/owner. For example, the Scherer coal power plant in Juliette, Georgia is operated by Georgia Power Co., but the ultimate owner is Southern Company. If an ultimate owner cannot be identified, the plant operator (often a utility company) is reported instead.

3. Detailed methodology

For plants that do not disclose electricity generation or CO_2 emissions, it is necessary to estimate values. Note that estimation is necessary only for a subset of power plants *outside* the U.S. since the electricity generation and CO_2 emissions of all U.S. facilities are effectively disclosed. The basic approach, as outlined in Section 2, is to use statistical models fitted to U.S. data to predict the performance of non-disclosing plants elsewhere in the world (i.e. those not disclosed through the databases listed in Table 1). This involves five steps:

1. Create concordance between variables found in U.S. datasets and those in the global WEPP database.

2. Process U.S. datasets to extract maximum information regarding PGU operation at monthly time scales.

3. Fit regression models to U.S. monthly data that predict PGU performance as a function of variables available in WEPP.

4. Predict the annual performance of PGU's using fitted models, applying national-level adjustments to restrain initial model estimates.

5. Integrate any disclosed data and replace or update model estimates as necessary.

Before describing methodological details, it is helpful to explain the nomenclature used to describe power plants and their characteristics. Many power plants are actually a collection of power generation units (PGU's), each typically consisting of a generator and, for certain combustion technologies, one or more boilers. In most cases, a generator produces electricity when its turbine is moved by a working fluid (e.g. steam, gas, water, wind, etc.). In a conventional steam turbine, fuel is burned in a boiler to produce the required steam. In a combustion turbine, the pressure produced by direct gas combustion moves the turbine. In combined cycle systems, heat is recovered from combustion turbine exhaust gas to produce steam that moves an additional turbine, resulting in higher overall efficiency.

3.1 Creating concordance across databases

It is useful to classify a PGU by a combination of its *prime mover technology* and *primary fuel*. For example, a conventional coal power plant's prime mover is a steam turbine and primary fuel is coal. In CARMA, the PGU's associated with this plant are classified on this basis using a set of codes that have been standardized to create concordance across all of CARMA's input databases. In this case, the PGU *type* is "ST_COAL" (i.e. prime mover_primary fuel). A conventional hydroelectric dam is classified as "HY_WAT" (hydraulic turbine and water). Table 2 describes the suite of PGU types classified in CARMA, whether they emit CO_2, and the share of global generating capacity in 2009. Notice that units with biofuels as the primary fuel are not listed as CO_2 emitters, as the carbon comes from recent biological fixation.

Table 2: Description of PGU "Type" variable

PGU Type	Prime Mover	Primary Fuel	CO2 Emitter?	% of Global Capacity
ST_COAL	Steam turbine	Coal	Yes	32.853%
HY_WAT	Hydraulic turbine	Water	No	19.035%
CC_FGAS	Combined cycle	Fossil fuel gas	Yes	13.557%
ST_NUC	Steam turbine	Nuclear	No	7.980%
ST_FGAS	Steam turbine	Fossil fuel gas	Yes	7.423%
GT_FGAS	Combustion (gas) turbine	Fossil fuel gas	Yes	5.837%
ST_FLIQ	Steam turbine	Fossil fuel liquid	Yes	4.975%
WT_WIND	Wind turbine	Wind	No	2.404%
GT_FLIQ	Combustion (gas) turbine	Fossil fuel liquid	Yes	2.187%
IC_FLIQ	Internal combustion engine	Fossil fuel liquid	Yes	1.301%
ST_BSOL	Steam turbine	Biomass	No	0.795%
CC_FLIQ	Combined cycle	Fossil fuel liquid	Yes	0.550%
IC_FGAS	Internal combustion engine	Fossil fuel gas	Yes	0.325%
OT_EMIT	Other	N/A	Yes	0.286%
ST_GEO	Steam turbine	Geothermal	No	0.224%
ST_WSTH	Steam turbine	Waste heat	No	0.082%
IC_BGAS	Internal combustion engine	Biogas	No	0.073%
PV_SUN	Photovoltaic	Sun	No	0.061%
ST_SUN	Steam turbine	Sun	No	0.015%
IC_BLIQ	Internal combustion engine	Bioliquid	No	0.009%
ST_BGAS	Steam turbine	Biogas	No	0.008%
GT_BGAS	Combustion (gas) turbine	Biogas	No	0.007%
CC_BLIQ	Combined cycle	Bioliquid	No	0.006%
CC_BGAS	Combined cycle	Biogas	No	0.004%
OT_NOEMIT	Other	N/A	No	0.002%
GT_BLIQ	Combustion (gas) turbine	Bioliquid	No	0.001%
CC_BSOL	Combined cycle	Biomass	No	0.000%
GT_WSTH	Combustion (gas) turbine	Waste heat	No	0.000%

U.S. EIA Form 860 reports annual data on generator capacity, age, and design steam flow, as well as a suite of variables describing the status of pollution control technologies at U.S. PGU's. The WEPP database provides an additional set of data about PGU's worldwide (including U.S. units) from which it is often possible to extract variables similar those found in the EIA-860. For example, it is possible to extract information about the presence of pollution control technologies from PGU descriptions in WEPP and create code concordance with similar data from EIA-860. Table 3 reports the suite of PGU-specific variables for which concordance was created between EIA datasets and WEPP. Most of the variables are 0/1 dummies indicating the absence of presence of a particular technology.

Table 3: Description of additional PGU variables

PGU Variable	Code	Notes
Installed capacity	cap	Gross electric generating capacity.
Unit age	age	Years since reported initial year of operation.
Generator status	sb	Operational or standby.
Business type	bustype	Utility, manufacturing, etc. (based on NAICS code).
Electricity production type	electype	Utility, auto-producer, or private generator.
Combined heat and power (CHP) status	chp	District heating, heat recovery for desalinization, etc.
Design maximum steam flow (steam-based units only)	sflow	If unavailable in WEPP, imputed using PGU type, capacity, age, status, design steam pressure, and design steam temperature.
Supercritical combustion status (sub-, super-, or ultra-)	super	Supercritical defined by reported or imputed design steam pressure > 240 bar. Ultra-supercritical defined by reported or imputed design steam temperature above 593ºC.
Fluidized bed technology	fbt	
Solid fuel gasification	sfg	
Activated carbon injection	aci	Mercury control.
Carbon capture technology	ccap	None operational. Included for future use.
Lime injection system	lis	
Wet flue gas desulfurization	fgd.wet	Jet bubbling reactor, tray, Venturi type, etc.
Dry flue gas desulfurization	fgd.dry	Spray, circulating, dry powder injection, etc.
Baghouse flue gas particulate (FGP) control	fgp.bag	Shake and deflate, reverse, pulse, etc.
Electrostatic precipitator FGP control	fgp.esp	Cold or hot side.
Cyclone FGP control	fgp.cyc	Single or multiple.
Other FGP control	fgp.oth	

PGU Variable	Code	Notes
Staged combustion NO_x control	nox.stag	Overfire air, biased firing, etc.
Flue gas recirculation NO_x control	nox.fgr	
Steam/water injection NO_x control	nox.stm	
Low-NO_x burner NO_x control	nox.lnb	
Selective (non)catalytic reduction NO_x control	nox.fgr	Includes ammonia injection.
Burner management system NO_x control	nox.ctrl	Low excess air, boiler optimization, etc.
Fuel reburning NO_x control	nox.burn	
Other NO_x control	nox.oth	

3.2 Extracting monthly performance for U.S. units

The PGU variables described in Tables 2 and 3 reveal no *direct* information about the actual operation of units (e.g. electricity generation, CO_2 emissions, etc.). Operational data for U.S. PGU's come from two sources: EIA Form 923 and the EPA Clean Air Markets (CAM) Program.

EIA-923 collects detailed monthly data on the operation of steam-based, organic-fired (i.e. fossil fuel) PGU's at U.S. facilities, as reported by the plant operators. Collected metrics include net generation, fuel input energy, and fuel characteristics (e.g. fuel heat content, sulfur content, etc.). For other facilities (i.e. non-steam and non-fossil fuel), EIA-923 provides similar monthly data disaggregated by prime mover technology and primary fuel at the plant level.

The initial challenge is to process the data to extract maximum information about the operation of individual PGU's at monthly time scale, or (second-best) information regarding small groups of PGU's at a single facility. Because the EIA-923 generation data are specified for individual generators while fuel consumption and fuel characteristics are specified for individual boilers, it is necessary to appropriately link generators and boilers at a given facility to generate aggregate totals for the unit as a whole. Using linkage data provided by the EIA-860, generator(s)-boiler(s) relationships are deduced at the highest possible resolution.[5] The

[5]In the simplest case, a generator is linked to a single boiler. In more complex cases, a generator may share a boiler with other generators (or a generator may be shared across multiple boilers). CARMA's algorithms group boilers and/or generators at individual facilities so that aggregate electricity generation and fuel consumption are aggregated at the highest possible level of specificity.

results are merged with the EIA-923's generation and fuel consumption data specified by prime mover technology and fuel type for non-steam and non-fossil fuel facilities.

An algorithm then combines the operational data from EIA-923 with EIA-860 engineering specifications. The end result is a dataset containing monthly electricity generation and fuel consumption and characteristics specified at either the individual PGU-level (in the case of steam-based facilities) or by PGU "type" at a given facility.

Table 4 shows representative results for DTE Energy's power plant in Monroe, Michigan for the months of September and October, 2009. Data for the "ST_COAL" observations comes from the EIA-923 generator- and boiler-specific data (hence, multiple ST_COAL "units"). Data for the other "units" come from the plant-level totals at the prime-mover level. Matching against the EIA-860 allows inclusion of the generating capacity (MW) variable along with others from Table 3 (not shown in Table 4). Note that the fuel characteristics (heat content, sulfur content, and ash content) are, in fact, MMBtu-weighted averages of the various coal types in use at the plant.

Table 4: Example results from processing of U.S. EIA 923 and 860 datasets

Monroe, Michigan power plant for September and October, 2009

Month	Type	MW	Net MWh	Fuel input (MMBtu)	Primary fuel	Primary fuel %	Secondary fuel	Secondary fuel %	Fuel heat content	Fuel sulfur content	Fuel ash content
9	IC_FLIQ	13.5	-12	215	FLIQ	100.0	NA	NA	59,442	NA	NA
9	ST_COAL	817.2	470,370	4,314,968	COAL	100.0	FLIQ	0.0%	21,673	0.61	6.2
9	ST_COAL	822.6	327,364	3,224,202	COAL	100.0	FLIQ	0.0%	21,579	0.61	6.3
9	ST_COAL	817.2	457,673	4,513,724	COAL	100.0	FLIQ	0.0%	21,685	0.61	6.2
9	ST_COAL	822.6	398,677	3,652,176	COAL	99.6	FLIQ	0.4%	21,709	0.62	6.3
10	IC_FLIQ	13.5	-41	0	FLIQ	100.0	NA	NA	59,546	NA	NA
10	ST_COAL	822.6	469,309	4,019,502	COAL	99.6	FLIQ	0.4%	21,556	0.64	6.3
10	ST_COAL	817.2	359,102	3,089,002	COAL	99.9	FLIQ	0.1%	21,579	0.64	6.3
10	ST_COAL	817.2	366,959	3,418,194	COAL	99.9	FLIQ	0.1%	21,944	0.70	6.2
10	ST_COAL	822.6	340,206	2,763,018	COAL	100.0	FLIQ	0.0%	21,662	0.65	6.2

Finally, CO_2 emissions data are added. The EPA's Clean Air Markets (CAM) database collects data reported by U.S. power plants under various regulatory and emissions trading programs. The emissions are recorded for individual units (typically boilers) at hourly resolution (often from stack measurements) and released quarterly. CARMA aggregates emissions to monthly totals and merges with the EIA-derived generation and fuel data.

This is complicated by the fact that facility and unit codes are not fully standardized across EIA and EPA databases. Consequently, CARMA employs algorithms that analyze fuel input and gross generation at the unit-level for both EIA and EPA data to determine when high-quality matches can be made across databases. Emissions data are not available for all plants – or even all PGU's within a given plant. Reporting is determined by the requirements of federal regulations, typically restricting observed emissions to larger facilities.

Continuing with the example data above, Table 5 shows a subset of results after matching the EIA-derived results against the EPA CAM emissions database for the Monroe, Michigan plant. We can see that only the larger coal units report CO_2 in the EPA CAM database. The table also shows the results of CARMA's unit-matching algorithm, which links generators and boilers on the basis of EIA-860 linkage data and then proceeds to link boiler-generator pairings with the emissions reporting units in the EPA CAM data ("EPA CAM Unit ID"), using fuel energy input and gross generation data from both sources to identify high-quality matches.

Table 5: Example results from processing of U.S. EPA CAM emissions database

Monroe, Michigan power plant for September and October, 2009

Month	Type	EIA-923 Generator ID	EIA-923 Boiler ID	EPA CAM Unit ID	CO2 emissions (tons)
9	IC_FLIQ	NA	NA	NA	NA
9	ST_COAL	1	1	13	488,598
9	ST_COAL	3	3	68	365,224
9	ST_COAL	4	4	80	511,080
9	ST_COAL	2	2	57	413,491
10	IC_FLIQ	NA	NA	NA	NA
10	ST_COAL	2	2	57	455,352
10	ST_COAL	1	1	13	349,908
10	ST_COAL	4	4	80	386,644
10	ST_COAL	3	3	68	312,879

Once the full suite of EIA and EPA operational data are merged, it is possible to calculate variables summarizing the critical processes to be modeled. The key quantities to be estimated by statistical models are annual electricity generation and annual CO_2 emissions at the plant level. Calculation of these quantities requires knowledge of a plant's generating capacity, capacity factor (i.e. rate of utilization), heat rate (i.e. technical efficiency), and emission factor (i.e. carbon-intensity of the fuel):

Generating Capacity x *Capacity Factor* = Electricity Generation

Electricity Generation x *Heat Rate* x *CO2 Emission Factor* = Total CO2 Emissions

Capacity factor (*cf*), heat rate *(hr),* and CO_2 emission factor (*co2.rate*) can be calculated from the merged EIA/EPA operational data for each unit-month. The capacity factor (*cf*) for a given unit and period is:

$$cf = \frac{net.mwh}{(cap * h)}$$

where *h* is the number of hours during the month in question. In practice, *net.mwh* can be negative if a generator is online and consuming electricity but not producing. It can also be unusually high if *cap* is unexpectedly expressed at the net rather than gross installed capacity. Consequently, *cf* is assigned minimum and maximum bounds:

cf = *0,* when $\frac{net.mwh}{(cap*h)} < 0$ and *cf* = 1.2, when $\frac{net.mwh}{(cap*h)} > 1.2$

The heat rate (*hr*) describes the quantity of fuel energy (*mmbtu*) required to produce a unit of electrical energy *(net.mwh)*:

$hr = \frac{mmbtu}{net.mwh} * 1.0551$, when *cf* > 0 for units where *type* is identified as a CO_2 emitter

Multiplying by 1.0551 converts *hr* to units TJ/GWh. The CO_2 emission factor (*co2.rate*) describes the amount of CO_2 released to the atmosphere (*co2.mass)* per unit of input fuel energy (*mmbtu*):

$co2.rate = \frac{co2.mass}{mmbtu} * 859.86$, for units where *type* is identified as a CO_2 emitter

Multiplying by 859.86 converts *co2.rate* to units tCO_2/TJ.

For both *hr* and *co2.rate*, the entirety of the U.S. data are analyzed to determine allowable ranges for feasible values in order to remove outliers likely to have resulted from data input errors (which are rare, but present, in the EIA data). For each unit *type*, the lowest feasible value is set to 2 times the interquartile range below the 25th percentile; the highest feasible value is set to 2 times the interquartile range above the 75th percentile. Observations with *hr* or *co2.rate* values outside this range are considered erroneous and excluded from further analysis.[6]

[6]Technical note: The EIA 923 reports gross calorific value (i.e. GCV or gross heat content) of input fuels, while the IEA national data report net calorific value (NCV). The EIA values are converted to NCV using default derate values from the IEA, or, in the case of coal, the following approximation derived from the U.S. coal analysis of Quick et al. (2005): NCV = GCV(1-X), where X = (-0.43*GCV+16)/100

The complete monthly, unit-level U.S. dataset contains approximately 65,000 observations for the year 2009. These observations are used to fit statistical models that can then estimate the performance of non-disclosing plants outside the U.S. Table 6 reports the *cf*, *hr*, and *co2.rate* variables calculated above for the Monroe, Michigan power plant.

Table 6: Example results after calculation of key performance variables

Monroe, Michigan power plant for September and October, 2009

Month	Type	Capacity factor (*cf*)	Heat rate (*hr*, TJ/GWh)	CO2 Emissions Rate (*co2.rate*, tCO2/TJ)
9	IC_FLIQ	0.00	NA	NA
9	ST_COAL	0.80	9.68	97.36
9	ST_COAL	0.55	10.39	97.40
9	ST_COAL	0.78	10.40	97.36
9	ST_COAL	0.67	9.67	97.35
10	IC_FLIQ	0.00	NA	NA
10	ST_COAL	0.77	9.04	97.41
10	ST_COAL	0.59	9.08	97.40
10	ST_COAL	0.60	9.83	97.26
10	ST_COAL	0.56	8.57	97.37

3.3 Fitting regression models to U.S. training data

This section describes the construction of models for predicting the *capacity factor, heat rate,* and *CO_2 emission factor* for non-disclosing power plants. Although the ultimate objective is to predict annual, plant-level performance, the models are fit to the previously described dataset of monthly, unit-level U.S. "training data" and predictions made using unit-level independent variables from WEPP. The unit-level predicted values are then aggregated to the plant-level to create the final estimates. The use of highly-disaggregated U.S. training data provides a wide range of operating conditions against which to fit the predictive models.

Before fitting models to the U.S. training data, a number of variables are constructed to mimic national-level independent variables available when predicting the annual performance of non-disclosing plants outside the U.S. For example, if the goal is to predict the capacity factor of a given coal plant in China in 2009, it is very useful to know the *average* capacity factor of *all* coal plants in China in 2009. And, indeed, this information can be derived by combining fuel-specific electricity generation from the IEA's Extended Energy Balances with WEPP's data on fuel-specific operational generating capacity.

The analogous value for a given coal plant in the U.S. training dataset is the average capacity factor for all coal plants within the boundaries of the associated North American Electric Reliability Corporation's (NERC) regional entity. The NERC regions are largely autonomous power grids with limited electricity trading and are, therefore, approximately analogous to the national boundaries used to tabulate IEA and EIA country-specific aggregate data on electricity generation, CO2 emissions, and fuel consumption. In other words, NERC regions in the U.S. training data are treated similarly to countries in the IEA, EIA, and WEPP data.

The following "grid-specific" variables are constructed for each unit in both the U.S. and WEPP data, using NERC regions and countries, respectively, as the "grid". In the case of U.S. data, the variables are calculated for each grid-*month*:

grid.cap : percentile of a given unit's capacity relative to all other units in the grid

grid.age : percentile of a given unit's age relative to all other units in the grid

plant.cap : percentile of total capacity for a given plant relative to all other plants in the grid

fuel.cf : average grid-wide capacity factor for a given unit's primary fuel

As with any modeling exercise that attempts out-of-sample predictions or forecasts, there is the risk that the training data is unrepresentative of the population for which predictions will be generated. There is also the related risk that the models themselves could be "over-fitted" (in terms of variable selection and functional form) and, therefore, modeling noise rather than underlying relationships that are applicable beyond the training data.

To address these concerns, two steps are taken. First, the models are fit a stratified random sample (n = 50,000) drawn from the complete monthly, unit-level U.S. dataset. The sampling weights are empirically derived from the set of non-U.S. units in the WEPP database, using unit *type*, *cap, grid.cap,* and *chp* as the stratifying variables. This ensures that the models are fit to a dataset where the likelihood of various unit types generally resembles that found in the out-of-sample predictor dataset.

Second, the model fitting process itself makes use of a multivariate adaptive regression splines (MARS) algorithm that allows piecewise non-linear relationships and includes a generalized cross-validation approach for deciding which independent variables and functional form to include in the final model. This approach reduces the risk of over-fitting while allowing the selection of independent variables and functional forms to vary across models.

MARS is implemented via *R*'s *earth* package (Milborrow 2012), using the approach of Friedman (1991). In short, MARS allows for piece-wise "hinge" functions to be adaptively fit to training data – no functional form must be specified beforehand. After adding hinge terms until the model's residual error is stable, a "backward pass" uses a generalized cross validation criterion to discard terms in an attempt to produce a parsimonious model that avoids over-fitting. Given the general absence of *a priori* information regarding appropriate func-

tional form – and the significant risk of over-fitting in the presence of so many observations and independent variables – MARS provides a good model-fitting process for this context.

3.3.1 Capacity factor (cf) model

Two capacity factor models are constructed. One for units where *grid.cap*<=0.5 (i.e. a model predicting the performance of relatively small units), and one for units where *grid.cap*>0.5. This is designed to capture different relationships between variables for small and large plants. The dependent variable for the CF models (*cf.ind*) consists of a [0,1] variable whereby each month-unit's *cf* is linearly scaled such that *fuel.cf* for the unit in question is assigned a value of 0.5:

$cf.ind = 0.5 + 0.5\frac{(cf - fuel.cf)}{(1.2 - fuel.cf)}$, when *cf* >= *fuel.cf*

$cf.ind = \frac{0.5\text{cf}}{cf.fuel}$, when *cf* < *fuel.cf*

In other words, *cf.ind* is constructed to measure the extent to which a unit's observed capacity factor deviates from the grid-wide capacity factor for units with the same fuel type. This approach allows cross-grid variation in the utilization of different unit types to be directly incorporated into the modeling framework via the observable *fuel.cf* variable. Both capacity factor models have the same dependent and independent variables (though the MARS algorithm may choose to leave certain dependent variables out of the final model); the only difference is the subset of data used to fit the model. The form is:

$$cf.ind = b_0 + b_1B_1type + b_2B_2bustype + b_3B_3electype + b_4B_4plant.cap + b_5B_5grid.age + b_6B_6sb + b_7B_7chp + \varepsilon$$

where b_0 is the intercept, ε is the error term, b_i are fitted coefficients, and B_i are basis functions that may consist of a constant, a hinge term, or a product of two hinge terms (i.e. a 2-degree model is allowed).

3.3.2 Heat rate (hr) model

A total of six heat rate models are fitted: two for each of the three dominant emitting fuel types (i.e. coal, fossil gas, and fossil liquid), stratified into large and small units by a fuel-specific cutoff point for *cap* (200 MW for coal, 50 MW for gas fossil fuel, and 25 MW for liquid fossil fuel). All six heat rate models have the same dependent and independent variables; the only difference is the subset of data used to fit the model. The form is:

$$hr = b_0 + b_1B_1type + b_2B_2bustype + b_3B_3electype + b_4B_4cf + b_5B_5\ln(cf) + b_6B_6\ln(hcon) + b_7B_7\ln(age) + b_8B_8\ln(cap) + b_9B_9\ln(sflow) + b_{\dots}B_{\dots}x_{\dots} + \varepsilon$$

where$b_{...}B_{...}x_{...}$denotes estimation for the suite of engineering and pollution control variables described in Table 3. Note that capacity factor (*cf*) is a dependent variable in the heat rate model, since units in regular operation will tend to exhibit greater thermal efficiency.

3.3.3 CO_2 emission factor (co2.rate) model

A total of three CO_2 emission factor models are fitted: one for each of the three dominant emitting fuel types (i.e. coal, fossil gas, and fossil liquid). All three CO_2 emission factor models have the same dependent and independent variables; the only difference is the subset of data used to fit the model. The form is:

$$co2.rate = b_0 + b_1B_1type + b_2B_2\ln(hcon) + b_3B_3age + b_4B_4cap + b_{...}B_{...}x_{...} + \varepsilon$$

where$b_{...}B_{...}x_{...}$denotes estimation for the suite of engineering and pollution control variables described in Table 3.

3.4 Predicting values for non-disclosed plants

Each of the models is applied to the WEPP predictor dataset to estimate *cf*, *hr*, and *co2.rate* for units worldwide. The prediction algorithm includes country- and fuel-level adjustments to ensure that the aggregate totals match those reported in the IEA Extended Energy Balances. For example, after the models make an initial prediction of capacity factor, all of the coal units in China are aggregated and a uniform adjustment is applied to ensure that the aggregate predicted capacity factor matches the value implied by the IEA data. Similar adjustments are applied for the heat rate and CO_2 emission factor, using IEA input fuel energy and CO_2 emissions data. Predictions are further constrained so as not to exceed the minimum and maximum feasible values as determined by analysis of U.S. data (see Section 3.2).

A manual heat rate adjustment is also included for supercritical and ultra-supercritical (USC) coal-fired units. Such technology is increasingly common in newly-built and planned coal plants worldwide, but there are few (or none, in the case of USC) operational units in the U.S. training dataset. Analysis of an EPRI (2008) engineering study suggests that each 1% increase in steam pressure results in a 0.28% decrease in net heat rate. This relationship is linear over the range studied, which includes steam pressure up to 352 bar for future, advanced USC. This manual adjustment is applied to all coal units with a reported or imputed design steam pressure greater than 240 bar (the approximate minimum steam pressure for supercritical designation).

3.5 Integration of disclosed plant data

Once *cf, hr, and, co2.*rate predictions are made the unit-level, it is easy to calculate estimated annual electricity generation and CO_2 emissions using the generating capacity (*cap*) data provided by WEPP. The unit-level values are then aggregated to the plant-level. At this point, it

is possible to calculate plant-level CO_2 intensity, which is the ratio of total CO_2 emissions to net electricity production (measured as $kgCO_2$/MWh).

Whenever plant-level CO_2 emissions or electricity generation are disclosed by a verified, public source, it is CARMA's policy is to replace model estimates with the actual data. Outside the U.S., such information is currently available for a subset of power plants in Europe, Canada, India, and South Africa, as well as electricity generation for worldwide nuclear power plants from the IAEA (see Table 1). Disclosure is not universal, which requires that the plants in the public databases be matched with WEPP in order to determine which plants are *not* disclosed and in need of estimated values. A limited exception is the U.S., for which disclosure coverage is universal, eliminating the need to match EIA plants with WEPP plants. However, *corporate ownership* data is *not* universal in the EIA data, which means EIA and WEPP must, in practice, be matched in order to assign WEPP corporate data to U.S. plants.

Matching observations in public databases against WEPP can be laborious and error-prone, particularly for smaller plants. Power plants names are not standard across databases. In some cases, distinct plants in WEPP or a public database must be consolidated to create an accurate match. In the EU and Canada, the disclosure databases include emitters from outside the power sector (like refineries and factories), making it difficult to narrow the set of potential matches. CARMA makes use of algorithms that attempt to identify potential matches on the basis of geographic data. Confirmation of matches is usually done manually. Efforts to date have identified WEPP matches for ~840 plants in Europe, Canada, South Africa, and India and a further ~2,700 U.S. plants.

For plants with a good match between WEPP and a disclosure database, the disclosed value replaces the model estimate(s). If *only* CO_2 emissions are reported (as in the EU and Canada), then the CO_2 intensity implied by the model estimates is applied to the actual emissions to back-out a new estimate of electricity generation.

Due to the universal plant coverage provided by the EIA, U.S. plant totals in CARMA bypasses WEPP altogether (except for assignment of corporate ownership data). That is, CARMA effectively reports EIA data directly for U.S. facilities. The one exception is small, CO_2-emitting plants for which emissions are not recorded in the EPA CAM database. In such cases, the average *co2.rate* across observable U.S. plants with similar fuel type is used to estimate total CO_2 emissions from reported fuel energy consumption.

4. Comparison of model estimates and reported values

To help assess the ability of CARMA's models to estimate plant performance, a special dataset was constructed containing the full set of likely matches between WEPP and public disclosure databases, including the U.S. This provided a set of ~3,500 plants (~800 from outside the U.S.) for which it is possible to compare CARMA's model estimates against reported values.

Note that analysis of the individual fitted models themselves (e.g. analysis of individual model residuals) would not provide useful information about CARMA's overall predictive skill. The only *practically useful* assessment of skill is to compare the annual, plant-level predictions produced by the suite of models to real-world observed values for the same plants. In addition, the models are, in practice, only used to make out-of-sample predictions (i.e. estimate performance of non-disclosing power plants *outside* the U.S.). Consequently, the *most* relevant conclusions regarding overall, applied skill are provided by assessment of predictions made for power plants outside the U.S.

An initial, rough assessment of model performance is provided in Figures 2 through 7. The graphs illustrate the relationship between estimated and reported values for the full matched set of plants both inside and outside the U.S. For total CO_2 emissions (Fig. 2 and 3) and electricity generation (Fig. 5 and 6), both linear and log scale plots are given in order to examine the relationship for both small and large plants. Fig. 4 plots estimated and actual CO_2 intensity and distinguishes plants on the basis of the primary fuel source. Fig. 7 reports the R^2 value for each plot and plant type (emitting versus non-emitting).

The plots suggest that CARMA's model estimates are, indeed, capturing broad differences among plants of various types and sizes. The R^2 values in Fig. 7 suggest that, *in a relative sense*, agreement between estimates and reported values across the full set of plants is *generally* good. Similar plots and summary statistics are were provided in Wheeler and Ummel (2008) when assessment the performance of the original CARMA methodology and models.

However, as noted above, it is more relevant to focus on performance outside of the U.S., as this is the population of power plants for which CARMA's model predictions will actually be needed. Further, summary statistics like the coefficient of determination (R^2) are misleading when calculated for variables with such a wide range of values. The R^2 statistic is dominated by a relatively small number of large facilities, capturing little information about model performance across the full set plants. The practical value of interest to CARMA users – and one that treats plants of all sizes equally – is the percentage error associated with estimates for individual plants.

Figures 2 through 7: Comparison of estimates and reported values for matched sample (2009 data; n = 3,500)

Fig. 2

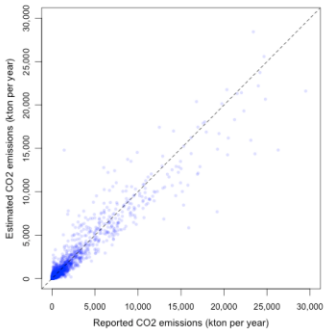


Fig. 3

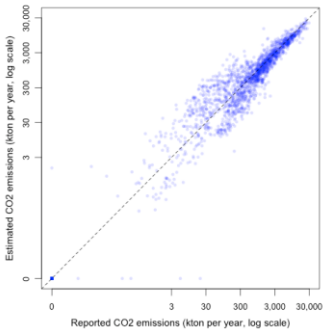


Fig. 4

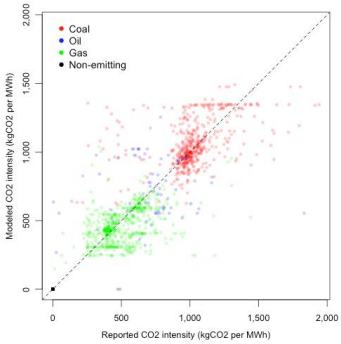


Fig. 5

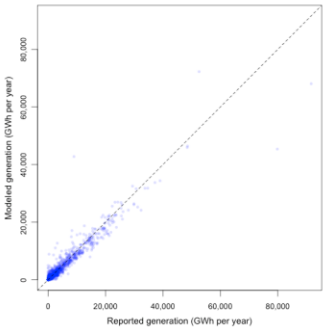


Fig. 6

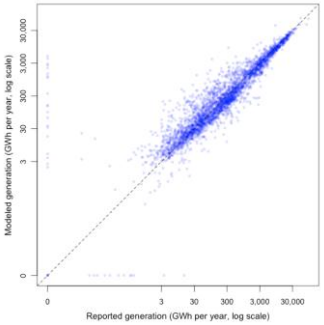


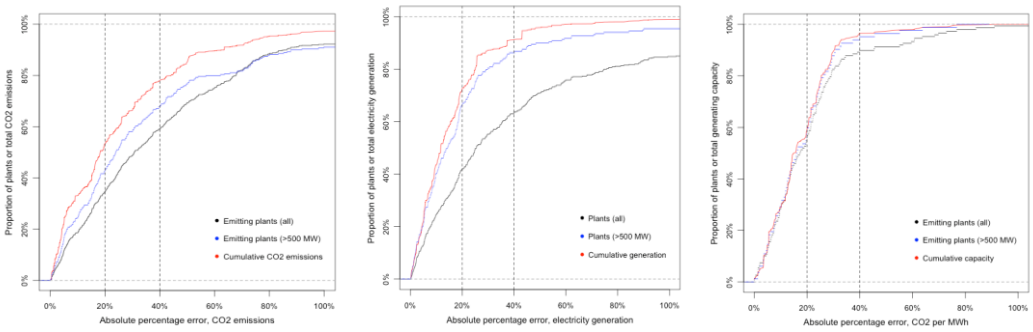
Fig. 7

R^2	All plants	Emitting plants	Non-emitting
CO_2 emissions	0.91	0.91	~1.0
Electricity generation	0.90	0.88	0.92
CO_2 intensity	0.90	0.66	~1.0

Figures 8 through 10 report the absolute percentage error (APE) associated with modeling estimates for *emitting power plants outside the U.S.* (n = 630 for Fig. 8 and 10; n = 832 for Fig. 9). Each figure reports three cumulative distributions for the APE. For example, Fig. 10 reveals that ~60% of emitting plants in the sample have an estimated CO_2 intensity that is within 20% of the true value.

Figures 8 through 10: Modeling error associated with plants outside the U.S. (2009 data)

Fig. 8 Fig. 9 Fig. 10



Figures 8 and 9 illustrate the difficulty in generating precise estimates for electricity generation and, for emitting units, CO_2 emissions. They also show that the quality of the estimates tends to improve for larger plants (blue lines). The *cumulative* amount of CO_2 or electricity generation associated with a given APE level is shown with the red lines. For example, Fig 9. reveals that ~90% of electricity generation in the sample comes from plants with an APE of less than 40%. Figure 10 confirms that modeling of CO_2 intensity results in considerably lower errors. About 60% of emitting plants in the sample have an APE for CO_2 intensity less than 20% and very few emitting plants exhibit an APE above 40%.

Plots of the cumulative error distribution are used to convey the range of APE values across different samples. Calculating *mean* APE is not particularly useful in this case, because the value is dominated by a few, extremely large errors (for example, the maximum APE for generation in the sample is 3,761%). Errors of extreme magnitude are likely the result of mis-matching of plants between WEPP and the public disclosure database or, alternatively, cases where a plant is effectively taken offline for most of the year (an unobservable event) but still treated as fully operational by the models. Perhaps the only defensible summary statistic is the *median* APE, which, for this sample of emitting power plants outside the U.S., is 25.5% for electricity generation, 17.3% for CO_2 intensity, and 30.1% for total CO_2 emissions. A more substantive measure of overall model skill is presented in Section 5.

Taking the results of Fig. 8 through 10 and extrapolating to the full set of power plants in CARMA, it is possible to make general statements about the *likely* error across the entire database. These statements take into account the fact that the CO_2 intensity and CO2 emis-

sions of many power plants can be deduced with certainty from the primary fuel source (i.e. APE for CO_2 emissions from hydroelectric dams is effectively zero).

Electricity generation: Across *all* plants in the CARMA database, it is estimated that more than 45% of facilities report annual electricity generation with absolute percentage error < 20% and more than 65% of plants are below 40% APE. About 80% of global electricity generation comes from plants with APE < 20% and more than 90% comes from plants with APE < 40%.

CO_2 emissions: Across *all* plants in the CARMA database, it is estimated that at least 75% of facilities report annual CO_2 emissions with absolute percentage error < 20% and more than 85% of plants are below 40% APE. Nearly 70% of global CO_2 emissions comes from plants with APE < 20% and more than 85% come from plants with APE < 40%.

CO_2 intensity: Across *all* plants in the CARMA database, it is estimated that nearly 85% of facilities report average CO_2 intensity with absolute percentage error < 20% and more than 95% of plants are below 40% APE. Nearly 80% of global generating capacity comes from plants with APE < 20% and nearly 100% comes from plants with APE < 40%.

5. Effects of year-to-year variability on model skill

The preceding section made clear that estimating electricity generation (i.e. capacity factor) for a given plant in a given year is particularly problematic. Wheeler and Ummel (2008) show that a given plant's capacity factor exhibits considerable year-to-year variation. The same trend is evident in more recent data. It is possible to examine the change in annual capacity factor between 2009 and 2010 for a set of ~5,000 U.S. power plants. The chosen facilities exhibit no change in generating capacity or engineering characteristics, yet significant changes in the rate of utilization are not uncommon.

Figure 11 gives the distribution of percentage change in annual generation for identical U.S. plants between 2009 and 2010. Note that nearly *half* of the plants exhibit a change of at least 20% and about 30% see a change of more than 40%. This variability suggests that estimating generation for a specific plant and year faces unavoidable difficulties. Specifically, it implies that even a "pseudo-optimal" model that is able to use last year's observed, plant-specific capacity factor to predict this year's capacity factor cannot be expected to exhibit an APE distribution significantly better than in Figure 11. For practical purposes, Figure 11 gives something like the "best case" error distribution when predicting plant- and year-specific electricity generation.

Figure 12 shows how different types of power plants exhibit quite different degrees of year-to-year variability in generation. Nuclear power plants tend to be large and highly-utilized, with limited inter-annual variability. Gas and oil-fired plants show much higher variability. These results are specific to the U.S., largely reflecting the relative operating prices and prior-

itization of fuels in the U.S. power sector. For example, gas is often (and oil almost exclusively) used in "peaking" operations that are inherently volatile.

To the extent that the U.S. experience is relevant to other countries, one can expect the accuracy of CARMA's generation models to mimic the trends found in Figures 11 and 12. Specifically, larger facilities are more stable and, therefore, more easily and reliably modeled that smaller facilities. In addition, nuclear and coal power plants – in part owing to their predominant use as base-load providers – should enjoy greater model accuracy (though this is a moot point in the case of nuclear plants, since they are all effectively disclosed through the IAEA). Conversely, smaller and/or gas- or oil-based units are likely to see higher prediction errors. Hydroelectricity exhibits a moderate amount of inter-annual variability in the U.S. for 2009 and 2010, but it should be noted that the operation of a given dam in any given year is highly dependent on local weather conditions that are not observed by CARMA's models.

In comparison to Figure 11, CARMA's prediction errors for plants outside the U.S. (Figure 9) are not unreasonable. A "pseudo-optimal" model would likely predict annual generation with APE<20% for about 55% of plants. Figure 9 suggests that CARMA's models currently achieve this level of accuracy for slightly more than 40% of plants. And whereas an ideal model might be expected to achieve APE<40% for about 70% of plants, CARMA does so in more than 60% of cases.

Figures 11 and 12: Variability in year-to-year electricity generation for identical U.S. Plants

Fig. 11

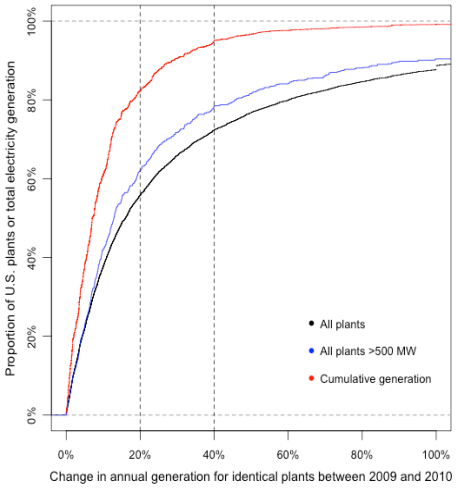
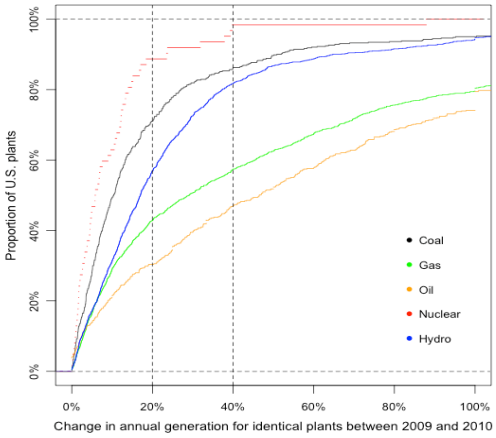


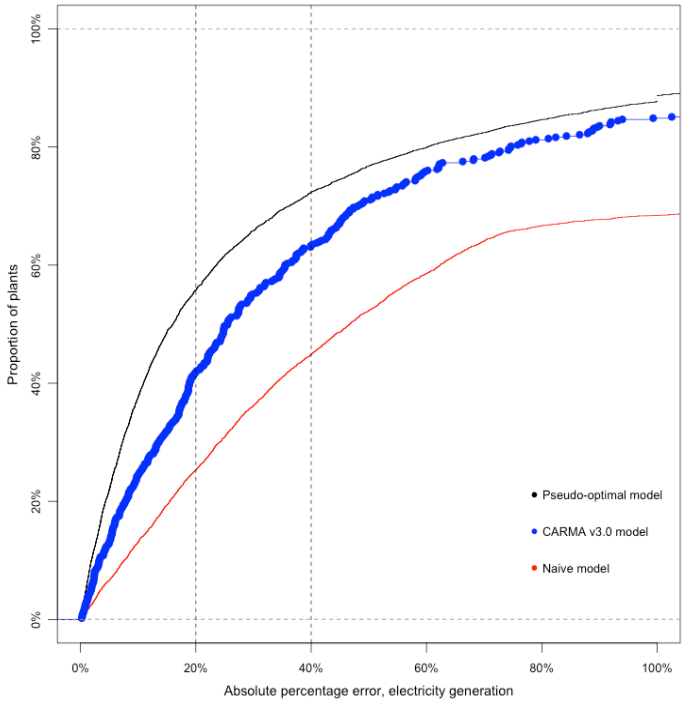
Fig. 12



We can also imagine a "naïve" model that assumes each plant's capacity factor is simply equal to the average capacity for the fuel type in question. Figure 13 shows how CARMA model estimates compare to the hypothetical "naïve" and "pseudo-optimal" models . The comparison is not exact, because the "pseudo-optimal" and "naïve" APE distributions come

from a large sample of U.S. power plants, while CARMA's error distribution is the same as that in Figure 9 (i.e. APE for a set of emitting power plants outside the U.S.). Still, the results are informative and provide probably the single best assessment of practical model skill. Figure 13 suggests that the CARMA v3.0 models eliminate about two-thirds of the reducible error, compared to a "naïve" model.

Figures 13: CARMA v3.0 electricity generation error distribution compared to alternative models



While precise, plant- and year-specific estimates are obviously preferable, they are likely impossible given inherent variability in rates of plant utilization. The factors driving the significant year-to-year variation are too site-specific to be sufficiently observed and modeled. Larger facilities do tend to operate by more regular and predictable rules, and this is confirmed by a general reduction in estimation error for larger plants. However, only increased disclosure of plant-specific data can really help overcome the low signal-to-noise ratio that hampers modeling efforts. Encouragingly, CARMA's models do appear to provide a significant improvement over more simplistic approaches to estimating plant-specific electricity generation.

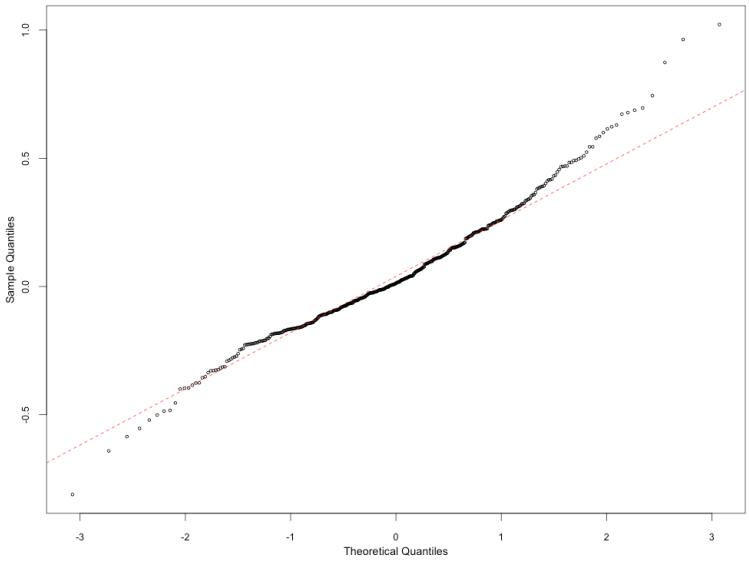
6. Aggregation effects

In addition to plant-specific figures, the CARMA database publishes aggregate totals for a wide variety of geographic regions and power companies. The electricity generation and CO_2 emissions totals for these entities are summed from the individual plant data. It is important to note that the typical prediction errors associated with aggregated plant data are significantly lower than for individual plants. This is due to the fact that prediction errors will tend to offset each other and "zero out" as plant-specific data are added together.

To confirm this, total electricity generation, CO_2 emissions, and CO_2 intensity were calculated for each U.S. state using 1) disclosed plant data and 2) estimated plant data for 2009. The APE between the actual and estimated state totals were considerably smaller than those observed for individual plants. Specifically, the median APE across states was 9.4% for electricity generation, 8.5% for CO_2 intensity, and 17.1% for total CO_2 emissions. These median APE values are ~ 45-65% lower than those reported for individual plants in Section 4.

This is not surprising, given that the prediction errors show no evidence of systematic bias. Figure 14 provides a "Normal Q-Q" plot for assessing the distribution of capacity factor prediction errors for emitting plants outside the U.S. (same observations as used in Figures 9 and 13). The plot suggests the errors are approximately normally distributed. The median error is just 1.2%, confirming the absence of bias.

Figure 14: Normal Q-Q plot of capacity factor prediction errors



Although not readily testable at present, it is very likely that prediction errors are offset *over time* for a specific facility. In other words, it's likely CARMA's estimates can be fairly interpreted as reasonable long-term, average performance metrics. While estimates for any *particular* year may exhibit significant error, the long-term performance of most plants is likely con-

sistent with the model predictions. This is especially true of larger plants. Measures of typical, long-term performance for larger facilities (existing and planned) are, perhaps, the most relevant information for many real-world uses of CARMA.

7. Geocoding of global power plants

In addition to plant performance, CARMA provides information regarding the location of individual power plants. CARMA v3.0 offers a number of significant advances in this area. The WEPP database includes variables for country, state/province, and city, though the coverage is inconsistent. An algorithm was developed to use the open-source GeoNames place names database and API to help standardize, fill-out, and expand the geographic data provided by WEPP (for example, entity spellings vary widely in the raw WEPP data).[7] This process also resulted in city-center latitude and longitude for about 70% of power plants worldwide. Over 6,000 additional high-resolution plant coordinates were obtained from disclosure databases in the U.S., Europe, and Canada and manual geocoding.

For users interested in using CARMA's data in geospatial applications, use of the approximate (city-center) coordinates may be necessary. For the set of ~6,200 plants for which the city-center *and* precise coordinates were obtained, it is possible to calculate the typical distance error. The results show that for about 50% of the sample, the approximate coordinates are within 5 km of the actual location and 70% are within 10 km. Among the closest 90% of pairs, the average distance is 7 km.

8. Conclusion

The CARMA database attempts to provide comprehensive information about the state of power plants worldwide, using a combination of public and private data and model estimates. Recent advances have expanded the amount of public data incorporated in CARMA, improved model estimates, and quantified the likely error. It is hoped that CARMA's continued presence will act as an impetus for further disclosure of plant-level CO_2 emissions, whether by governments or corporations. CARMA's policy is to integrate data from companies that provide verified, plant-specific emissions reports for the entirety of their generating fleet.

Importantly, CARMA v3.0 also lays the technical groundwork for an expansion to non-CO_2 compounds (both GHG's and conventional pollutants). CARMA's initial development was inspired by the threat of global climate change and need to reduce power plant CO_2 emissions, but many other pollutants of interest are often correlated with CO_2. Some of these compounds have discernible local impacts (e.g. acid rain, urban air pollution, etc.). Combined with CARMA's extensive geographic data, an expansion to other pollutants would enable education, research, and activism at local scales for a wider range of environmental

[7] Linkage of CARMA's *R* processing scripts and the GeoNames API was provided by the *geonames* package (Rowlingson 2011).

and human health threats. This would complement CARMA's existing focus on CO_2 emissions, providing educators, policymakers, researchers, investors and activists with an even richer suite of detailed information about the environmental footprint of power plants worldwide.

Works cited

EPRI. 2008. Engineering and economic evaluation of 1300ºF series ultra-supercritical pulverized coal power plants: phase 1. Electric Power Research Institute, Technical Update, September.

Friedman, J.H. 1991. Multivariate adaptive regression splines. *Annals of Statistics* 19: 1-141.

IEA. 2010. World energy outlook 2010. International Energy Agency, Paris.

Milborrow, S. 2012. earth: multivariate adaptive regression spline models. R package version 3.2-2. Available at: http://CRAN.R-project.org/package=earth

Quick, J.C., Tabet, D.E., Wakefield, S., and Bon, R.L. 2005. Optimizing technology to reduce mercury and acid gas emissions from electric power plants. U.S. Department of Energy, October.

Rowlingson, B. 2011. geonames: interface to www.geonames.org web service. R package version 0.99. Available at: http://CRAN.R-project.org/package=geonames

Wheeler, D. and Ummel, K. 2008. Calculating CARMA: global estimation of CO_2 emissions from the power sector. Center for Global Development Working Paper 145, May.