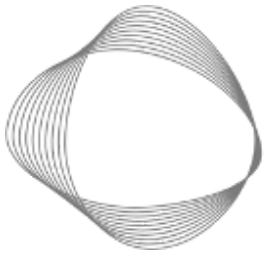
Power sector: Emissions from Electricity Generation



Jeremy Freeman[1,5], Ali Rouzbeh Kargar[1,5], Heather D. Couture[2,5], Jeyavinoth Jeyaratnam[1,5], Jordan Lewis[1,5], Madison Hobbs[1,5], Hannes Koenig[1,5], Tiffany Nakano[3,5], Charmaine Dalisay[3,5], Aaron Davitt[1,5], Lee Gans[1,5], Christy Lewis[1,5], Gabriela Volpato[1,5], Colin McCormick[1,4,5], and Gavin McCormick[1,5]

1) WattTime, 2) Pixel Scientia Labs, 3) Global Energy Monitor, 4) Georgetown University 5) Climate TRACE

1. Introduction

Responding to climate change requires timely and accurate measurement of greenhouse gas (GHG) emissions, especially CO_2. The Paris Agreement, adopted in 2015, set goals to limit global temperature rise and established frameworks for nations to report on GHG emissions and steps to reduce them (Paris Agreement, 2015). The energy sector is responsible for the majority of GHG emissions – 76% globally as of 2019, representing 38 gigatons of CO_2 ($GtCO_2$) (Ge et al., 2020). Nearly half of the energy sector's GHG emissions result from electricity generation, accounting for 16 $GtCO_2e$ in 2019 or over 31% of total global GHG emissions (Ge et al., 2020).

Power plant emissions monitoring and reporting currently includes continuous emissions monitoring systems (CEMS) and bottom-up approaches (Liu et al., 2020). CEMS measures emissions at individual plants, providing reliable measurements. However, CEMS are costly, have limited deployment globally, and require continual calibration (Cusworth et al., 2021). A more common approach is bottom-up self-reporting, which quantifies emissions at power plants using fuel consumption, fuel quality, and emission factors (Kuhlmann et al., 2019; Vaughn et al., 2018). Yet, this approach tends to have uncertainties in fuel properties which leads to uncertainty in power plant emissions (Kuhlman et al., 2019; Cusworth et al., 2021). In addition, the recency and spatiotemporal resolution of self-reporting varies by region, posing challenges for policymakers to design sustainability strategies (Nassar et al., 2017; Cusworth et al., 2021). Alternatively, GHG monitoring satellites and aerial surveys have also been used to monitor power plant emissions (Nassar et al., 2017; Cusworth et al., 2021). However, existing approaches lack spatiotemporal resolution or have limited deployment. Alternatively, multi-spectral satellites that observe Earth's surface provide an opportunity to improve monitoring. Combining imagery from these satellites provides persistent global coverage and can identify more granular power plant activity.

Currently, there is no comprehensive, third-party source of facility-level measured GHG emissions data for the global power sector. The Climate TRACE power sector aims to complement current monitoring and reporting approaches with emissions estimates derived

from remote sensing and machine learning. The power sector data includes power generation emission estimates from combustion power plants with fuel sources including coal, natural gas, fuel oil, diesel, biomass, and waste. Note however that biomass is not included in the country totals and is listed separately at the source level (see Table S2) to avoid double counting with the Climate TRACE forestry and land use sector. We use a combination of existing country- and region-level data, satellite data, and machine learning models to generate a comprehensive set of country-level annual emissions estimates, which we have published for the period 2015-2022 inclusive, alongside source-level (i.e. facility-level, individual-power-plant-level) annual emissions estimates for all power plants with capacity 100 MW or greater for which we know the latitude and longitude. We also have a good number of plants under 100 MW capacity but for which we have capacity, location, and plant name information, so we also published these at the source-level as long as their capacity was at least 10 MW. This totaled 8333 unique plants between 2019-2022 from across 169 countries, accounting for 12.34 GtCO2 in 2022 or 96% of total global power emissions. This is 14 times more power plants reported at the source-level, compared to last year's 2022 Climate TRACE release which reported 603 power plant annual emissions estimates at the source-level. This is one of several improvements and updates since the 2022 Climate TRACE release one year ago, summarized below.

Table 1 Summary of changes since last year's 2022 Climate TRACE release:

2022 Climate TRACE Release	2023 Climate TRACE Release
Country level: Jan. 1, 2015, to Dec. 31, 2021 Source (power plant) level: Jan. 1, 2019, to Dec. 31, 2021	Country level: Jan. 1, 2015, to Dec. 31, 2022 Source (power plant) level: Jan. 1, 2019, to Dec. 31, 2022
603 power plant annual emissions estimates reported at the source-level	8,333 power plant annual emissions estimates reported at the source-level (96% of total global fossil/waste power emissions)
Directly published reported electricity generation data for plants used in the training set from CAMPD (U.S.), ENTSO-E (Europe), and NEM (Australia)	No reported electricity generation data was directly published as Climate TRACE's estimates.
Conservative temperature and humidity filters for FGD machine learning model	Expanded temperature and humidity filters for FGD machine learning models (see Figure 7)
Power plant capacity, capacity factor, generation, and CO_2 emissions, and GWP published by Climate TRACE include the total across all fossil, waste, and biomass fuel types.	Fossil and waste fuel and biomass capacity, capacity factor, generation, and CO_2 emissions, and GWP are broken out separately to avoid double-counting emissions from the Climate TRACE forestry and land use sector. For plants that use both fossil/waste fuel and biomass, the emissions are appropriately broken out based on the power plant's capacity of each fuel type. Country-level reporting includes only fossil and waste fuel based emissions, while source-level reporting includes both, as separate columns (see Table S2).

2. Materials and Methods

2.1 Overview of Approach

We combined multiple complementary approaches to generate source-level emissions estimates for essentially all major emitting power plants, encompassing 96% of fossil/waste emissions, or 93% of emissions including biomass. For a given plant, we first estimated its total generation with a combination of the following approaches:

1. Satellite-derived estimates of plant generation using satellite imagery and machine learning (ML). Our current ML approach focuses on the water vapor plumes which act as a proxy signal for power generation and CO_2 emissions. This approach is possible only for plants with structures that emit such vapor plumes: natural draft wet cooling towers (which we abbreviate NDT) and wet flue gas desulfurization (FGD) stacks, described later in section 2.3.1. Thus, we used satellite- and machine-learning-derived estimates for 3% of combustion (fossil and waste, excluding biomass) power plants worldwide, which accounts for 30% plants with >500 MW capacity, 27% of global fossil and waste fuel combustion power plant capacity, and 37% of Climate TRACE estimated global CO_2 from fossil fuel and waste combustion power plants for 2022. Our machine learning approach consists of two steps:
    a. Our image-level (also called "sounding-level") models predict the activity of a power plant from a single satellite image. We train separate models using PlanetScope, Sentinel-2, and Landsat 8 satellite imagery, matched to hourly or sub-hourly reported generation data from individual power plants in the United States, Europe, and Australia.
    b. Our generation model ingests the predictions from the sounding-level models in step 1 to estimate the average hourly capacity factor of a plant over the last 30 days.
2. Country- and fuel-specific average capacity factors derived from country-level data to estimate generation for all remaining plants (described later in section 2.3.8).

For plants where both estimation approaches were possible, for the final annual source- and country-level estimates, we averaged the two methods together because there was evidence to suggest the satellite and ML-derived estimates were biased low (see Section 3.3.1 for further details). This is identical to what was done in last year's 2022 Climate TRACE data release.

Finally, we used country-, fuel-, and prime-mover-specific average carbon intensities to convert source-level generation estimates to emissions estimates and aggregate source-level estimates to the country level. This process is explained in more detail in section 2.3.9.

2.2 Datasets Employed

2.2.1 Global Combustion Power Plant Inventory

In order to form a set of power plants for training our model and running inference, we partnered with Global Energy Monitor to create a complete-as-possible, harmonized inventory of global combustion power plants that are currently operational. To use a plant in our ML modeling, we required the following inventory data:

1. An accurate plant location for our satellite imagery.
2. The location of FGD flue stacks and NDT cooling towers.
3. Attributes of the power plant, including type, fuel, cooling technology, and air pollution control equipment, to determine whether a plant is suitable for our models.
4. Local weather data to decide if temperature and humidity are conducive to vapor plume visibility.
5. Plant capacity to determine whether the plant is of sufficient size to be modeled and to calculate the generation from the modeled capacity factor. This is used in conjunction with unit operating dates to find the plant capacity on any given date.
6. Fuel and prime mover type to estimate the emissions factor.

In addition, plant-level electricity generation data is needed as labels for model training. Plants without generation data are used for inference only. To perform inference, we required the capacity, fuel type, prime mover type, and less precision on the location data (province/country level).

Unfortunately, all existing power plant inventories have shortcomings, including missing, outdated, conflicting or incomplete information. Therefore, we developed our own harmonized global inventory of power plants by assimilating data from as many sources as we could. Each dataset contains different, complementary data that were merged together and standardized. Table 2 below describes how we use each of the datasets, including which data we republish.

Table 2 Datasets employed to create a harmonized Global Combustion Power Plant Inventory

Dataset Name	Plant Metadata Used	Published (use varying by source/country)
US Energy Information Administration EIA-860, EIA-860m	Plant Name, Unit Fuel Type, Location, Unit Capacity, Unit Operating Dates, Unit Cooling type, Unit Pollution Control Tech SO_2	Source Level Data, Aggregated into Country Level Data
World Resources Institute (WRI) Global Power Plant Database (GPPD)	Plant Name, Plant Fuel Type, Location, Plant Capacity, Plant Operating Dates	Source Level Data, Aggregated into Country Level Data
S&P Global/Platts World Electric Power Plant (WEPP) database	Unit Fuel Type, Unit Capacity, Unit Operating Dates, Unit Cooling type, Unit Pollution Control Tech SO_2	The source level dataset is proprietary and is used internally only.

Global Energy Monitor (GEM) Global Coal Plant Tracker (GCPT), Global Oil and Gas Plant Tracker (GGPT), and Global Bioenergy Power Tracker (GBPT)	Plant Name, Unit Fuel Type, Location, Unit Capacity, Unit Operating Dates	Source Level Data, Aggregated into Country Level Data
Other Sources (e.g., press releases, newspaper articles, company websites)	All	Source Level Data, Aggregated into Country Level Data

The US Energy Information Administration (EIA) dataset is the only one that provides all relevant data points for every US-based plant. Therefore, we primarily used EIA for the US. For the rest of the world, we used a combination of the other datasets.

To harmonize our datasets and get all the required information for every plant, we mapped units and plants between datasets. Global Energy Monitor provides unit- and plant-level mappings to World Electric Power Plant (WEPP), while Global Power Plant Database (GPPD) contains plant level mappings to WEPP. For those plants missing linkages, we matched them ourselves.

GPPD, WEPP, and Global Energy Monitor have overlapping information, such as the capacity of many plants. Plants with discrepancies for overlapping values were investigated and validated via primary sources such as newspaper articles, press announcements, etc.

In addition, some datasets are more up-to-date than others. Global Energy Monitor, for example, contains recently built plants not found in other datasets. Comparisons and validation of the base datasets were done to ensure the most up-to-date plant information was included in our final dataset.

2.2.2 Plant Validation and Infrastructure Mapping

To validate and augment our plant-level data, we used OpenStreetMap (OSM), a publicly-available and free geographic database. First, we manually cross-referenced and corrected the geolocation of power plants in our harmonized dataset. Second, OSM enabled us to annotate ("tag") physical features of power plants. We used tags to label parts of the plant from which we expect to see vapor plumes: NDT cooling towers and FGD flue stacks.

Cooling towers can be easily identified by their large, hyperbolic structure, such as those in Figure 1.

Figure 1 Cooling towers at Camden Power Station in South Africa as seen from side-view (left; source) and overhead imagery, where plumes are being emitted (right; source).

We identified FGD flue stacks by manually looking for the presence of FGD scrubbers adjacent to the stack in Google Maps and OSM imagery. An example is shown in Figure 2.

Figure 2 In this Google Maps image of Togtoh power station (Inner Mongolia, China), the FGD scrubbers are denoted in red boxes, adjacent to the flue stack emitting a vapor plume. Annotators

use these structures to confirm the stack has FGD. In wet FGD stacks, flue gas exits the boiler/generation unit of a plant and is first intercepted by "scrubber" units adjacent to the flue stacks that scrub SO_2 and inject water, then send the resulting vapor/exhaust mixture up and out of the flue stack. The scrubber units are most often two circular buildings placed on either side of the flue stack with ducting entering them from the generation buildings and exiting them into the flue stack. While not nearly as tall as the flue stack, the scrubbers are large multistory buildings/chambers themselves and are easily visible in high-resolution imagery as provided by Google Maps.

For every power plant on which we run ML training or inference, we completed the following manual tasks using OSM:

1. Confirmed that there is a power plant at the provided coordinates.
2. Verified that it is the correct plant by checking that plant information and visible technology (e.g., cooling equipment, coal piles) on the ground matches our information about the plant.
3. Annotated all FGD flue stacks.
4. Annotated all NDT cooling towers.

We created our own annotations for specialized tags that are not relevant for OSM, including labeling of flue stacks with FGD technology. More information on our activities on OSM can be found on the Climate TRACE OSM wiki page (https://wiki.openstreetmap.org/wiki/Organised_Editing/Activities/Climate_TRACE).

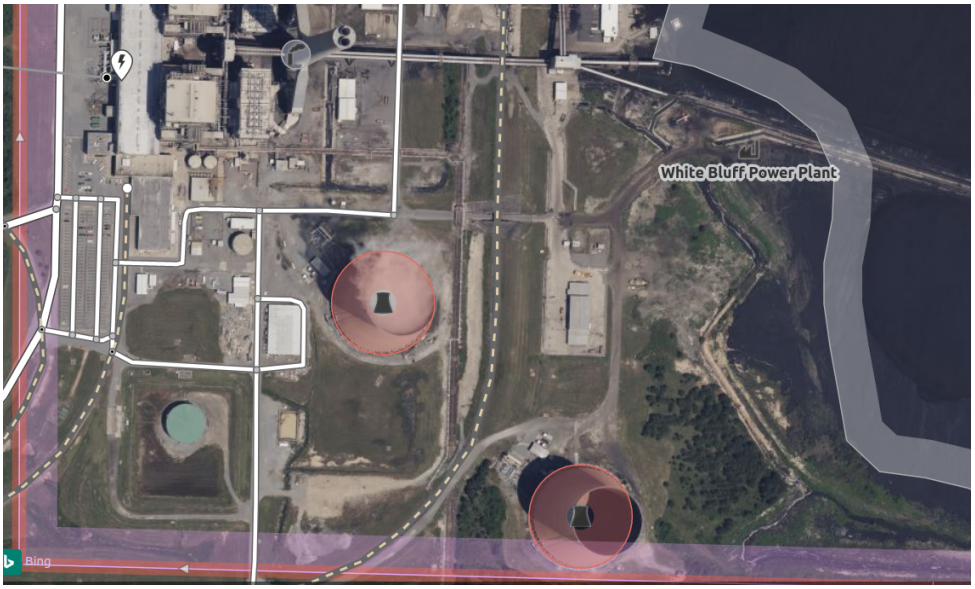


Figure 3 White Bluff power station as shown on OpenStreetMap (top) and in OpenStreetMap edit mode (bottom). We used this aerial imagery to annotate locations of FGD flue stacks (translucent white circle) and NDT cooling towers (red circles)

2.2.3 Weather data

Our ML models were trained to observe visible vapor plumes to predict power plant activity. However, we observed that visible vapor plume formation was reduced at high ambient temperature and low relative humidity, particularly for FGD structures which emit a fainter plume than NDT. In order to focus our models on weather conditions in which we expect to see a signal, we applied a set of empirically-derived filters, as detailed in Section 2.3.2. We obtained historical weather data from 2015-2022 for all of our plants from World Weather Online, available at https://www.worldweatheronline.com**.**

2.2.4 Plant-Level Electricity Generation Data

To train our ML models, we used multiple sources of reported high time-resolution (hourly to sub-hourly) plant-level generation data in MWh for plants in regions where this was available. While many datasets are available that provide low time-resolution generation data (days to months) or generation aggregated across a large number of power plants, these are not usable in our ML model training set. Our datasets include the US EPA Clean Air Markets Program Data (CAMPD), European Network of Transmission System Operators for Electricity (ENTSO-E), and Australia National Electricity Market (NEM). These datasets provide us with generation at hourly or sub-hourly intervals for several thousand power plants, from 2015 to the present. We matched each source's time series to the power plants in our database, resulting in plants with reported generation data in 23 countries.

This reported generation data must be complete and accurate for our models to be useful. To train models that predict on/off and capacity factor on each satellite image, we require hourly (or more frequent) reported generation to match to satellite imagery to avoid letting too much time elapse and risk the power plant generation value changing by the time imagery is captured. Images that cannot be matched to reported generation within the hour prior to capture are not used to train models. In power plants that have clearly visible activity-related signals, the power generation values can be validated, either through hand-labeling studies or by inspecting samples where our models have particularly confident errors. We reviewed a selection of false negative predictions from our models, i.e., cases where the reported data claims that the power plant is active, but the models predicted that it's off. This may be due to cooling towers following a dry (no plume) cooling process while our harmonized powerplant inventory incorrectly shows them as wet (plume-producing) or due to plants with inefficient or non-operating FGD pollution controls. For false positives, i.e., power generation is reported as zero, but there is an obvious plume coming out of a cooling tower or flue stack, it is more likely to be because the time gap is too great between reported generation and when the image was captured or there exists a generation reporting error. While reviewing our models, we came across a couple hundred images that showed visible NDT or FGD plumes but with generation reported as zero and a handful of others which showed no visible signal but reported generation; we excluded images from model training. For plants with an abundance of issues

with reported generation data (e.g., failure to report generation for operating units, reporting generation several months past retirement, or insufficient or inconsistent reported generation data), we excluded the entire plant from the training set. These tactics helped us avoid "garbage in, garbage out;" i.e., prevented the machine learning models from learning incorrect patterns due to erroneous data. This process also helped us identify and correct data issues (e.g. dry vs. wet cooling towers).

Various additional datasets are used for country-level estimates of generation, carbon intensity, and for baseline generation data. These include the US Energy Information Administration, Ember Electricity Generation Data and IEA Electricity Generation Data, all providing global information.

For model validation and confidence calibration, as described in Section 2.4, we also gathered reported electricity generation data from three additional countries for as many NDT or FGD power plants as possible:

- Taiwan – Taiwan Power Company, also known as Taipower (台灣電力公司). Reports electricity generation data every 10 minutes.
- Türkiye – Enerji Piyasaları İşletme A.Ş. (EPİAŞ), also known as Energy Exchange Istanbul (EXIST). Reports hourly electricity generation data.
- India – National Power Portal (NPP). Reports daily electricity generation data.

We were not able to train on power plants in India because we require hourly (or more frequent) reported data. Although the data from Türkiye and Taiwan meet this requirement, we have not yet quality controlled this data to the same extent as the rest of our training data (which has been in our database for multiple years). We are prioritizing including Turkish and Taiwanese power plants in training going forward. Although we did not use them for model training, all three datasets were included in our calibration set described in Section 2.4, which we used to estimate model error and uncertainty and generalize model performance to all plants. By augmenting our calibration dataset using power plants in Taiwan, Türkiye, and India for validation in this way, alongside out-of-sample predictions on plants from training regions, we are able to avoid overfitting our model performance and error estimates to training regions (which would amount to underestimating our error).

2.2.5 Satellite Data and Processing

Remote sensing imagery from the PlanetScope constellation, Sentinel-2A/B, and Landsat 8 satellites were employed in our ML modeling approach to infer a power plant's operational status through the identification of emitted visible water vapor plumes. A sample image of a power plant from each satellite is shown in Figure 4. A description of each satellite and imagery processing steps is provided below.

PlanetScope. Planet Labs' PlanetScope satellite constellation consists of approximately 130 individual satellites, called "Doves," with the first launch of this constellation in 2014 (Planet, 2022). The PlanetScope constellation provides daily revisits with an equator crossing time between 7:30 and 11:30 am (Planet, 2022). Each PlanetScope satellite images the earth's surface in the blue, green, red, and near-infrared (NIR) wavelengths (~450 nm – ~880 nm), with the exception of the "SuperDove" instrument which includes additional wavelengths (Dos Reis et al., 2020; Moon et al., 2022). We downloaded the PlanetScope PSScene via the Planet Labs API, providing a spatial resolution of ~3m, and including 8 additional Usable Data Mask (UDM2) image quality bands.

Sentinel-2. The European Space Agency's Sentinel-2 mission comprises two satellites: Sentinel-2A launched in 2015 and Sentinel-2B launched in 2017 (Main-Knorn et al., 2017). Each Sentinel-2 satellite has a 10-day revisit time with a 5-day combined revisit and an equatorial crossing time of 10:30am (Shikwambana et al., 2019). Both satellites are equipped with a multispectral (MSI) instrument that provides 13 spectral band measurements, blue to shortwave infrared (SWIR) wavelengths (~442 nm to ~2202 nm) reflected radiance and, depending on the band, provides measurements at 10m to 60m spatial resolution (Main-Knorn et al., 2017; Shikwambana et al., 2019). We downloaded Harmonized Sentinel-2A/B Level-1C Top of Atmosphere (TOA) products from Google Earth Engine (GEE). More information on the Sentinel-2 mission and data products can be found at the ESA Sentinel-2 mission website (sentinel.esa.int/web/sentinel/missions/sentinel-2).

Landsat 8. The Landsat 8 mission is jointly managed by NASA and U.S. Geological Survey (USGS; Marchese et al., 2019). Landsat 8 was launched in 2013 and has a 16-day revisit with an equatorial crossing time of 10 am (+/- 15 minutes). Landsat 8 is equipped with two instruments: Operational Land Imager (OLI) and the Thermal Infrared Sensor (TIRS). Together, these instruments provide 11 spectral band measurements, blue to thermal infrared (TIR) wavelengths (~430 nm to ~1250 nm) and, depending on the band, provide 30m and 100m spatial resolutions (Mia et al., 2017). We downloaded the Landsat 8 Collection 2, Tier 1 TOA from GEE. More information on the Landsat 8 mission and data products can be found at the USGS website (https://www.usgs.gov/landsat-missions/landsat-8).

Figure 4 The Ninghai power station in China as seen from Landsat 8 at 30m spatial resolution (left), Sentinel 2 at 10m spatial resolution (center), and PlanetScope at 3m spatial resolution (right). These images only represent the visible bands (red, green, and blue).

Each satellite has multiple bands with different spatial resolutions. We upsampled all lower resolution bands to match the highest resolution band for that satellite. Each band also has a different distribution of pixel values. We standardized each band to a mean of 0.5 and a standard deviation of 0.25, placing most pixel values between 0 and 1.

When possible, our models use all satellite bands available, plus some additional bands we created through post-processing and used whenever possible:

1. Haze-optimized transformation (HOT), a linear combination of blue and red bands
2. Whiteness
3. Normalized difference vegetation index (NDVI)
4. Normalized differences between the two SWIR or thermal bands

These additional bands provided beneficial features for some of our models, for example HOT and whiteness can act as a basic plume mask. Our gradient boosted tree models use all these bands, while the neural networks are more limited because of transfer learning, as described in Section 2.3.4.

For all satellite datasets, a region of interest (ROI) was produced for each power plant by setting an outer boundary that envelops the plant itself, all associated facilities, and any other affected areas of interest. Filters were applied to ensure image quality and low cloud cover. With all these image and plant filters taken into account, our sounding models ran inference on 1060 plants across 42 countries from 2015-2022 inclusive.

2.3 Modeling Pipeline

Observing the visible emitted water vapor plumes in multi-spectral satellite imagery, we trained ML models to infer a power plant's operational status. Specially, we designed models to perform two tasks:

1. Sounding-level model: to estimate if a plant was running or not (on/off) or to what extent it was being utilized, given a satellite image of that plant at a certain point in time.
2. Generation model: to aggregate the predictions from the sounding-level models into estimates of utilization over the preceding 30 days.

Each sounding model is trained on satellite images paired with the reported generation status. After predicting the utilization, we multiplied by capacity to infer generation. We trained our models on plants in countries for which we have hourly or sub-hourly generation data, then we applied them globally using country-, fuel-, and prime-mover-specific average carbon intensities to convert plant-level generation estimates to emissions estimates. Figure 5 provides an overview of how these different models are integrated to estimate emissions.

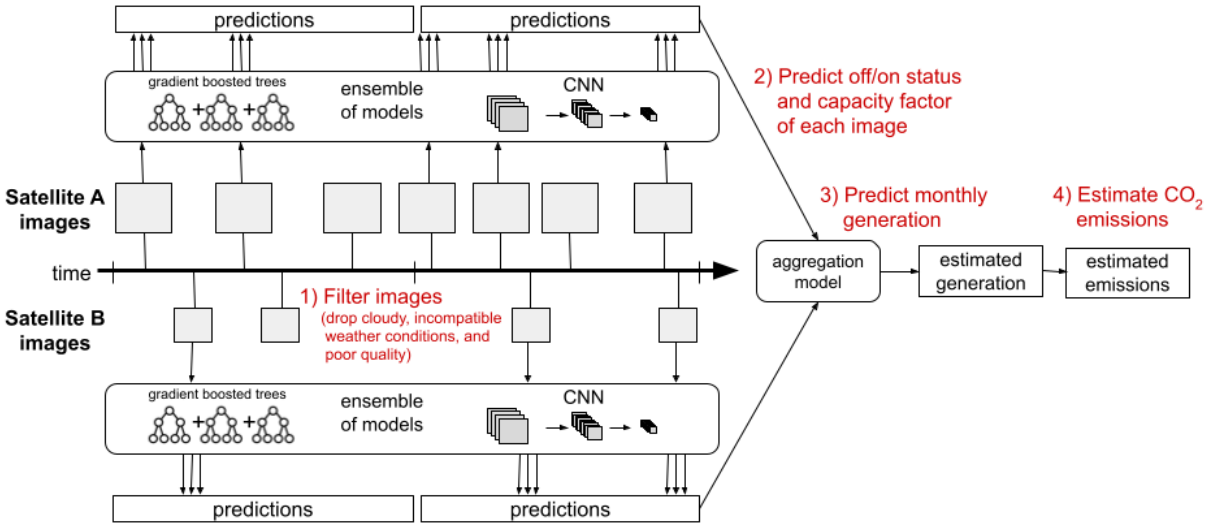


Figure 5 Overview of how satellite images are turned into CO_2 emissions estimates. First, a set of filters are applied to select modelable plants and images (Section 2.3.2). Then an ensemble of sounding models are applied to predict plant on/off status and capacity factor from cropped images around the cooling towers and flue stacks (Section 2.3.4). Sounding predictions are combined with a generation model to estimate monthly generation (Section 2.3.5) before a final CO_2 emissions estimate is calculated (Section 2.3.9).

2.3.1 Plant Selection

Power plants emit GHGs through a chimney called the flue stack, producing a flue gas plume. Plants that are more efficient or have air pollution control equipment generally have flue gas plumes that are difficult to see. Further, fuel characteristics and power plant equipment varies, impacting the size and visibility of these plumes. For these reasons, directly inspecting the visible flue gas only provides a weak indicator of emissions. A better indicator of emissions are the vapor plumes from two primary sources:

1. Natural draft towers (NDT): Plants using NDT have a large hyperbolic structure that allows vapor plumes formed during cooling to be clearly seen.

2. Wet flue gas desulfurization (FGD): After desulfurization, the flue gas becomes saturated with water, increasing the visibility of plumes from the flue stack.

A power plant may have one, both, or neither of these technologies. We created separate models for NDT and FGD due to differences in the size and shape of resulting plumes. Examples of both types of plumes are shown in Figure 6.

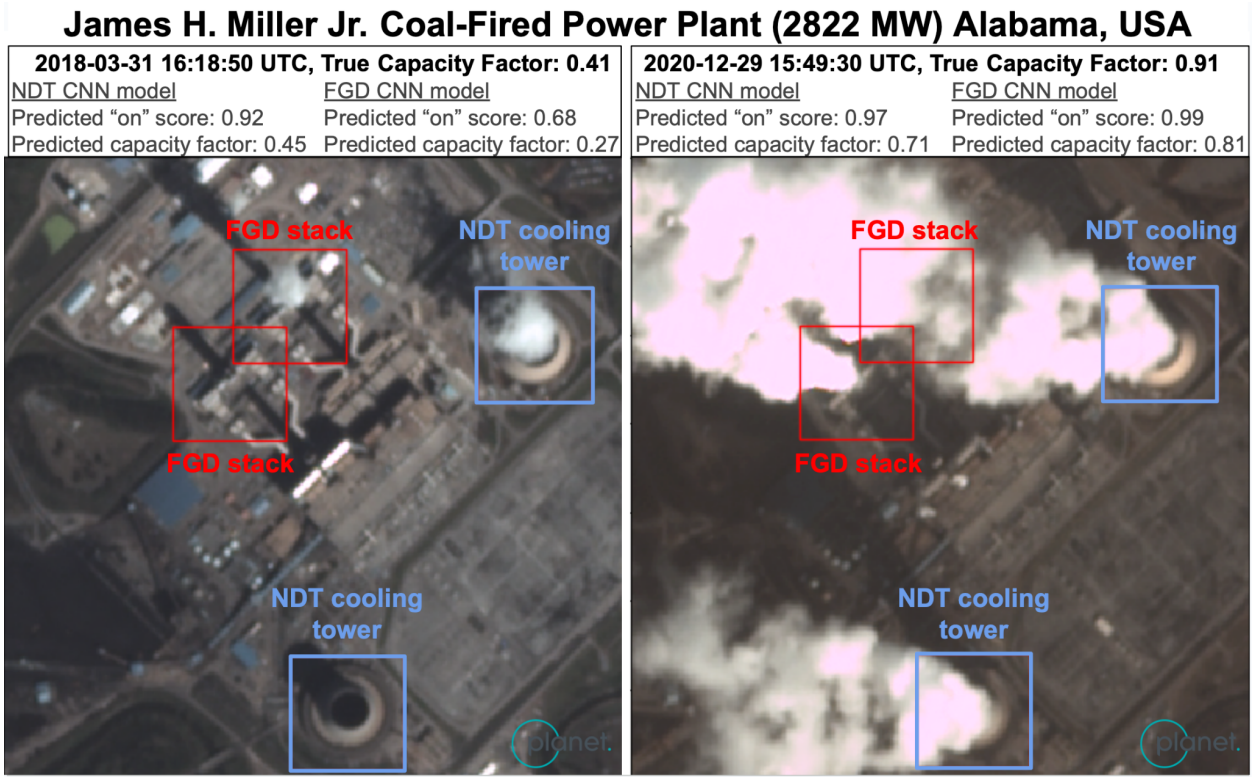


Figure 6 PlanetScope CNN predictions on the James H. Miller Jr. power plant at low vs. high generation. Separate NDT and FGD models predicted on NDT cooling tower (blue) and FGD stack (red) patches, respectively. NDT plumes are generally wider than FGD plumes due to the different size and shape of the vapor outlets. These predictions are ingested by subsequent models to estimate generation, then CO_2, for the plant. Image taken from Hobbs et al. (2023).

Separate models were built for these two sets of plants. We applied further criteria to these plants for training models:

1. Require that coal account for at least 50% of the plant's operating capacity.
2. Require that for training NDT models, ≥70% of the plant's cooling was NDT. Require that for training FGD models, ≥90% of total generation and 100% of coal capacity was associated with wet FGD for FGD models at least some point in our dataset.
3. Require that at least one NDT tower or FGD-enabled flue stack has been annotated in OpenStreetMap or in our in-house annotations database.
4. Exclude plants with generating capacity less than 500 MW.

5. Exclude from training any plants with incomplete or erroneous generation and/or retirement date, for example: failure to report generation for operating units, reporting generation several months past retirement, insufficient or inconsistent reported generation data. This is only an issue for training models, it is not a concern for inference; therefore, such plants are included in inference.
6. Because our modeling approach assumes plants with wet FGD always run these pollution controls, we removed from training any plants which repeatedly exhibited sporadic or no visible FGD usage when the plant was reported "on." For inference, this is harder to assess, but we flagged operational plants with FGD where FGD plumes were never seen under the expected plume-favorable weather conditions and/or which exhibited other signals of operating (i.e., NDT plumes) with no FGD signal.

These filters resulted in a training set with a total 100 plants to train the FGD models and 77 plants to train the NDT models.

We relaxed some of these filters for inference with the understanding that plants that passed our stricter training filters would likely have more accurate estimates, but that some model estimate is likely better than the country- and fuel-specific average imputation method for those plants that fail the strict filter but pass a looser one. We decided to not publish ML-estimated results using the looser filter unless model performance was estimated (using the calibration dataset as described in Section 2.4) to be better, on average, than imputing the country- and fuel-specific average. We further used the calibration dataset to estimate the performance gap between plants meeting the strict and the loose filters, as described in Section 2.4. The looser inference filter can be summarized as follows:

1. NDT models ran inference on plants with any positive amount of NDT capacity.
2. FGD models ran inference on plants with any positive amount of wet FGD capacity. However, post-hoc analysis on the inference set revealed that FGD models applied to the looser filter performed worse, on average, than using the country- and fuel-specific average. For this reason, we only published data that met the stricter filter used during training (≥90% of total generation and 100% of coal capacity was associated with wet FGD).
3. We required that at least one NDT tower or FGD-enabled flue stack had been annotated in OpenStreetMap or in our in-house annotations database.
4. We excluded plants with generating capacity less than 50 MW.
5. We filtered out operational plants with FGD where FGD plumes were never seen under the expected plume-favorable weather conditions and/or exhibited other signals of operating (i.e., NDT plumes) with no FGD signal.
6. As our current filtering process naïvely assumes unchanging plants, we ended up with plants that passed the inference filter at some point during the time period 2015-2022 but failed the filter at other times. Therefore, we also filtered out such plants post-inference to

avoid reporting misleading results, and we are in the process of converting our filtering system to be time-sensitive to avoid this issue going forward.

2.3.2 Satellite Image Selection

The following satellite images were excluded from our models:

1. The cloud mask over the plant indicated ≥ 20% cloud coverage. This threshold is set relatively high to avoid falsely excluding images containing large plumes, which are easily misclassified as clouds.
2. Images in which our flue stack and cooling tower annotations are not fully contained within the satellite image.
3. For PlanetScope, we used the post-2018 UDM2 bands to exclude images with ≥ 80% heavy haze.
4. For all PlanetScope images, we calculated mean brightness and developed a cloudiness score based on HOT, whiteness, and NDVI to respectively filter out excessively dark or cloudy images.
5. Images with known data quality issues were discarded, e.g. exhibiting plumes when generation has been zero for at least an hour. Section 2.2.4 details the scenarios in which we excluded images due to quality issues.

In addition, images were excluded from FGD models when ambient weather conditions were unfavorable for plume visibility. At high temperature and/or low relative humidity, the water vapor in the flue stack does not readily condense, plume visibility is reduced, and our models have no signal to detect. The warmer the temperature, the more humid it needs to be for water vapor plumes to be visible, eventually becoming very faint at high temperatures no matter the humidity. While at colder temperatures, even very dry conditions will still result in a visible plume. Therefore, we used empirically-derived cutoff rules for plume visibility:

1. Exclude images in which the ambient temperature is ≥ 14°C and relative humidity is ≤ 26%.
2. Exclude images in which the ambient temperature is ≥ 24°C and relative humidity is ≤ 36%.
3. Exclude images in which the ambient temperature is ≥ 32°C.

This filter can be visualized in the two plots in Figure 7 from our empirical analysis. They demonstrate how models are unreliable in the excluded temperature-humidity region (lower right). The filter is the same in both plots; the left plot displays classification model performance while the right plot shows regression performance.

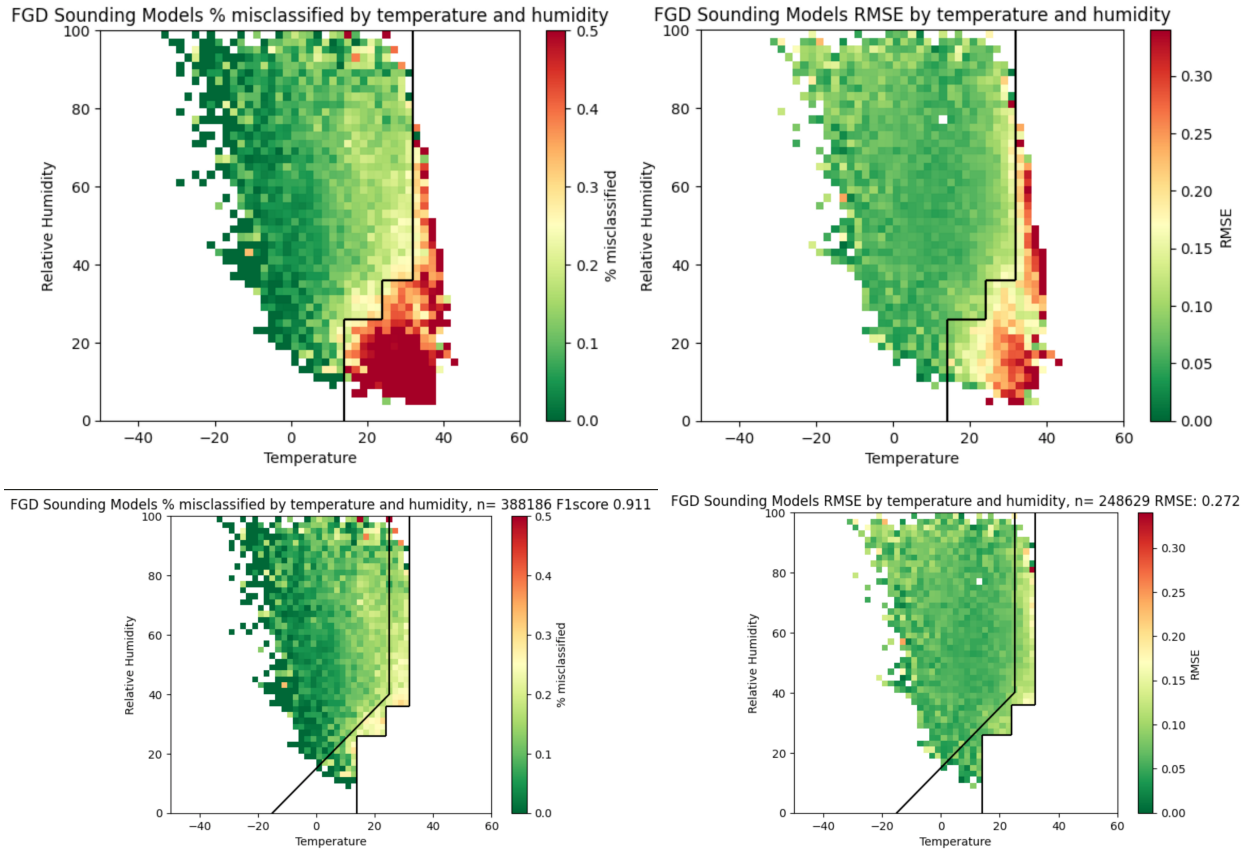


Figure 7 Empirically-derived temperature and humidity filters for FGD models. Upper two plots: model predictions in red, as seen in the lower right portion of the plots, are unreliable due to the high temperature and/or low humidity; the new "stair-step" filter is empirically set. Lower two plots: unreliable predictions have been filtered out and the previous, more conservative, FGD temperature and humidity filter has been overlaid to show the expanded prediction space this new filter has introduced since last year's Climate TRACE release.

These filters collectively resulted in a training set spanning 2015-2022 with 7201 Landsat 8 images, 15106 Sentinel-2 images, and 54247 PlanetScope images for NDT; and for FGD with 8282 Landsat 8 images, 18548 Sentinel-2 images, and 69942 PlanetScope images.

Note: this temperature and humidity filter represents a slightly more generous, expanded filter as compared to our release in November 2022. The motivation for this was to minimize unnecessarily discarding data, particularly for hotter or drier regions, which were particularly sparse due to the overly-strict filter.

2.3.3 Labels for Model Training

To train our ML models, we needed our dataset of satellite images to be linked to plant-level generation data. We used satellite image timestamps to match each image to the nearest record of

plant-level generation data, described in Section 2.2.4. Our training plants are located in the US, Europe, and Australia because these were the three regions from which we could source hourly-reported generation data and had enough time to quality-control. In recent months, we were able to also pull in hourly reported data from Türkiye and include this in our calibration set (described in Section 2.4). Once we have completed quality control checks, we will add Türkiye to our training set.

For regression models, we labeled each image with the capacity factor in that image: the generation of the plant divided by its capacity at the given timestamp. For classification, we labeled plants with greater than 5% capacity factor as "on" and everything else as "off." We used this nonzero threshold because there are a handful of plants reporting very low levels of generation that can functionally be considered "off."

2.3.4 Sounding Models

To estimate power plant generation and CO_2 emissions, we built models to identify power plant activity. These models performed the following tasks given a satellite image at a certain point in time:

1. Classification to predict if a plant was generating or not (on/off)
2. Regression to estimate the capacity factor (generation divided by capacity)

We included both sets of models because the on/off task is simpler and can be predicted more accurately, while the regression task is essential for differentiating low from high generation. We multiplied the predicted capacity factor by capacity to infer generation.

To focus our models on the most relevant parts of the plant, we used the annotated NDT cooling tower and FGD stack patches as the inputs. This strategy was previously shown to produce more accurate classification models than a single image of the entire plant (Couture, 2020). We trained two different types of models: gradient-boosted decision trees and convolutional neural networks (CNNs). We built separate models for NDT and FGD, as well as for each satellite dataset. Figure 8 illustrates the structure of both model types.

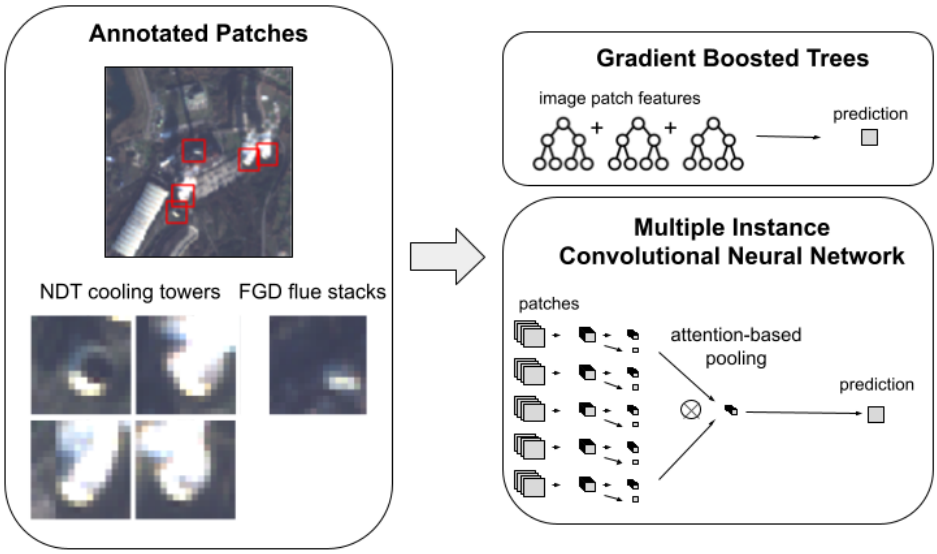


Figure 8 Diagram of the ML models used to estimate power plant operating status and capacity factor from satellite imagery. It showcases how relevant parts of power plants are cropped from satellite imagery, in this case forming patches of NDT cooling towers and FGD flue stacks, which were then used by machine learning models to classify the power plant's operating status as on or off and/or to regress on the capacity factor (i.e., what fraction of the power plant's potential is being used to generate power). These models can either be based on convolutional neural networks (CNNs), which combine the patches through an attention layer, or gradient boosted trees, that aggregate statistical features of the images of each infrastructure type. Image adapted from Couture et al. (2020).

The gradient-boosted decision tree models used XGBoost. Features were calculated by taking the mean, standard deviation, and 90th percentile of the pixel values in each imagery band of patches of varying sizes centered on FGD or NDT structures, producing a vector of statistics which we aggregated to the image level using mean, min, and max operations. We used multiple patch sizes (4 to 32 pixels for Landsat 8, 4 to 32 pixels for Sentinel-2, and 8 to 64 pixels for PlanetScope).

To better handle plants with white buildings near the cooling tower or flue stacks, we also included features from background-subtracted images. Background images were calculated as the median pixel value across a random set of 32 images of the plant. The background images were then subtracted from the current image and the same set of statistical features were calculated and concatenated to the previous set.

The CNN models used a multiple instance framework to combine the patches through an attention layer to aggregate features from the different patches for a plant (Couture, 2020). We encoded each patch with a CNN backbone (truncated after some of the convolutional blocks). We used transfer learning to initialize model weights in one of two ways:

1. RESISC: with a ResNet50 CNN (He et al., 2016) pre-trained on the RESISC dataset, which consists of aerial RGB images labeled by land cover and scene classes (Cheng et al., 2017). The RESISC dataset is particularly relevant because it includes a cloud class, enabling the model to capture distinguishing features of clouds – and, likely, plumes. This model uses the RGB channels only.
2. BigEarthNet: with a VGG16 CNN (Simonyan et al., 2014) pre-trained on the BigEarthNet dataset, which consists of Sentinel-2 images labeled by land cover class (Sumbul et al., 2019). This model uses 10 bands from Sentinel-2, excluding the lowest resolution bands 1, 9, and 10. We are not able to apply this model to PlanetScope but do adapt it for Landsat 8 by matching the band wavelengths as closely as possible and pairing the remaining bands even if the wavelengths are different. While this dataset enables the model to learn a more diverse set of spectral characteristics, it contains only cloud-free images; our model must learn plume features during fine-tuning.

After the siamese CNN backbone, the attention layer combined patch features as a weighted sum, with the weights determined by the model itself (Ilse et al., 2018). A dense layer was applied to make the final prediction.

For the classification CNNs, we used a softmax layer and cross-entropy loss. These models were trained using class weighting so that the on and off classes were represented equally during training. This is necessary because plants were "on" in about 80% of the training images.

For the regression CNNs, we used a sigmoid layer with either mean squared error or Huber loss based on performance. The regression models had a more difficult time converging than the classification ones. We found the simplest solution was to train it as a multi-task model that performs both classification and regression, with weights of 0.02 and 0.98 applied to each, respectively.

The patch size around each tower or stack was optimized as a hyperparameter for each model type and imagery product. Patch sizes ranged from 8 to 64 pixels, with larger patch sizes selected for regression models where the model can benefit from a full view of the plume.

We trained our CNN models using the AdamW optimizer that uses weight decay for regularization. We also regularized with dropout and image augmentation, including transformations for random flipping, rotation, brightness, contrast, darkness, Gaussian blur, translation, and zooming. The brightness, contrast, darkness, and Gaussian blur transformations simulate some of the image quality issues that we saw. Many of these issues are caused by natural phenomena like haze and various lighting conditions. As we trained models on a single satellite at a time with a fixed spatial resolution, the amount of translation and zooming augmentation is relatively small but does provide a benefit. We tuned the magnitude of image augmentation during hyperparameter optimization.

Our XGBoost and CNN approaches can be equally applied to our classification and regression problem, and to NDT or FGD models. In practice, we build at least one of each of these models for all distinct combinations of problem class (classification, regression) and plant type (NDT, FGD). We independently tuned the hyperparameters of both XGBoost and CNN models using cross-validation (see Section 2.3.6) to optimize for validation mean average precision (mAP) and root mean squared error (RMSE) for classification and regression, respectively. We selected mAP due to the class imbalance.

2.3.5 Generation Ensemble Model

The sounding-level models described in Section 2.3.4 give us an instantaneous estimate of the power plant activity in a single image. Collecting these instantaneous estimates creates an irregular time series of classification and regression estimates for each plant (Figure 9). In order to estimate emissions of a plant over a given period of time, we built a set of second-stage models ("generation models") responsible for aggregating the sounding model time series into features and predicting a rolling 30-day average capacity factor for each plant (predicting one value for each day, that value being the average hourly capacity factor over the preceding 30 days). We then multiplied this capacity factor with each plant's capacity to obtain the plant's estimated generation. Afterwards, the estimated generation was multiplied by an emissions factor, as described later in Section 2.3.9, to estimate emissions.

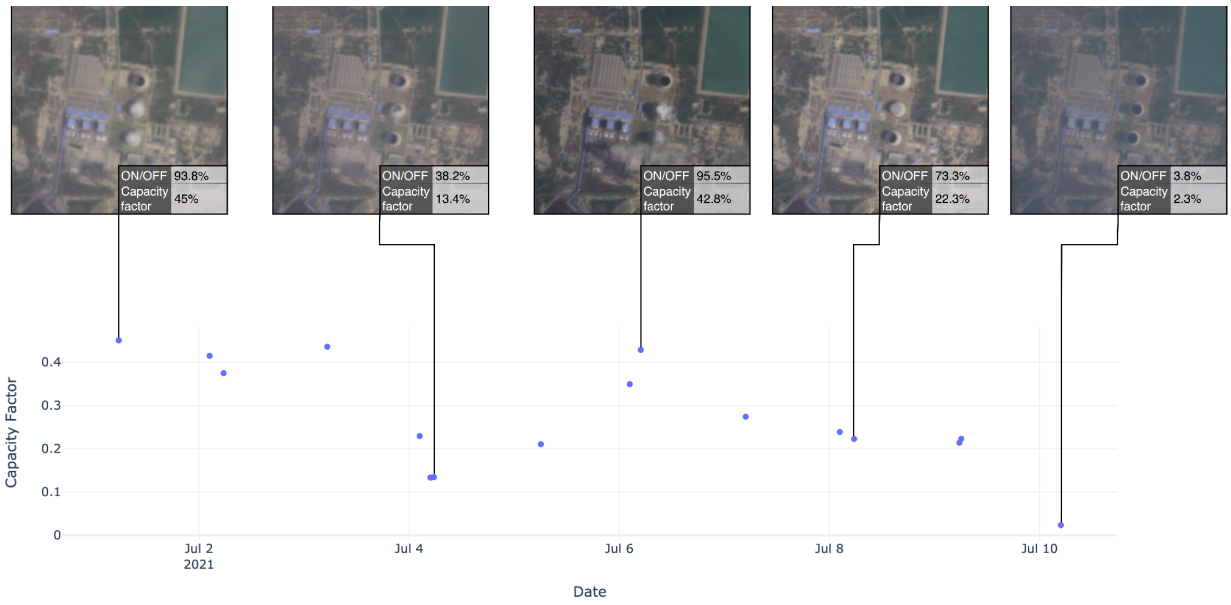


Figure 9 Depiction of the capacity factor time series, alongside on/off classification, for the power plant Talwandi Sabo Power Project in India, using PlanetScope satellite imagery and a CNN-based model.

We trained separate generation models for NDT and FGD plants. Each generation model is a L1-regularized linear regression model with features based on the predictions obtained from models applied to individual images. To calculate features for each plant at each point in time, we aggregated the predictions from the sounding-level models over multiple lookback windows. Our equations below use the variables defined as follows:

- A sounding prediction p from the set of soundings P within a lookback window; |P| represents the number of sounding predictions in the lookback window
- A sounding model m_s from the set of sounding models M_s associated with a satellite s; $|M_s|$ represents the number of sounding models for satellite s
- A classification sounding from sounding model m_s: y_{pm_s}
- A regression sounding from sounding model m_s: z_{pm_s}

If a feature value was missing, i.e., when there were no soundings for a plant during the lookback window, we imputed the value by calculating the average of the feature across all plants within the generation training fold. The generation models had access to the following feature sets calculated within each lookback window:

1. Model-averaged regression & classification soundings: We averaged each sounding model's capacity factor predictions and, separately, the ON-scores during the lookback window:

$$\underline{y}_{m_s} = \frac{1}{|P|}\sum_{p\epsilon P} y_{pm_s}$$

$$\underline{z}_{m_s} = \frac{1}{|P|}\sum_{p\epsilon P} z_{pm_s}$$

   This produced a feature for each sounding model for each satellite.

2. Satellite-averaged regression & classification soundings: We averaged the capacity factor predictions and, separately, the ON-scores from all sounding models associated with a satellite. This resulted in one ensembled capacity factor estimate and ON-score per image in the lookback window. These values were then averaged over the images to obtain a single value per lookback window:

$$\underline{y}_s = \frac{1}{|P|\,|M_s|}\sum_{p\epsilon P}\sum_{m_s\epsilon M_s} y_{pm_s}$$

$$\underline{z}_s = \frac{1}{|P|\,|M_s|}\sum_{p\epsilon P}\sum_{m_s\epsilon M_s} z_{pm_s}$$

   This produced a feature for each satellite.

3. Weighted-average regression soundings: We weighed the capacity factor related predictions based on the corresponding classification soundings. First, we averaged the classification soundings from all sounding models associated with a satellite:

$$\underline{y}_{ps} = \frac{1}{|M_s|} \sum_{m_s \epsilon M_s} y_{pm_s}$$

This produced one ensembled ON-score per image in the lookback window. These values were then used to weigh the capacity factor related predictions. The further away from 0.5 the ensembled ON-score, the higher the weight, with a maximum weight of 1 and a minimum weight of 0. The resulting weighted regression scores were then averaged within the lookback window to obtain a single value. This was done for each model and for each satellite:

$$w_{m_s} = \frac{1}{\sum_{p \epsilon P} \underline{y}_{ps}} \sum_{p \epsilon P} 2\,|\underline{y}_{ps} - 0.5|\, z_{pm_s}$$

$$w_s = \frac{1}{\sum_{p \epsilon P} \underline{y}_{ps}} \sum_{p \epsilon P} 2\,|\underline{y}_{ps} - 0.5| \frac{1}{|M_s|} \sum_{m_s \epsilon M_s} z_{pm_s}$$

This produced a feature for each model and one for each satellite.

4. Mean thresholded classification soundings: These features indicate the percentage of ON-scores in the lookback window that were above 0.5:

$$b_{m_s} = \frac{1}{|P|} \sum_{p \epsilon P} I(y_{pm_s} > 0.5)$$

$$b_s = \frac{1}{|P|} \sum_{p \epsilon P} I(\frac{1}{|M_s|} \sum_{m_s \epsilon M_s} y_{pm_s} > 0.5)$$

where I is an indicator function mapping to 1 if the condition is true and 0 otherwise. This resulted in a feature for each model and one for each satellite.

5. Missing feature indicator (FGD only): This value indicates if a feature was imputed, 1 if imputed and 0 otherwise. Imputation was used more often for the FGD model due to the stricter temperature and humidity filter.

In total, the NDT model had access to 162 features, while the FGD model had access to 324 features. All features were standardized to zero mean and unit standard deviation prior to model training. The same standardization factors were applied for inference.

The lookback windows for the NDT generation model were 30, 60, and 180 days. Due to the lower number of soundings for the FGD generation model, its lookback windows were 30, 60, and 365 days. In addition to including more plant-specific information, these longer windows also allowed our models to account for the longer-term behavior of the plant when making a prediction. The generation models were trained with L1-regularization weights of 0.01 and 0.005 for FGD and NDT, respectively.

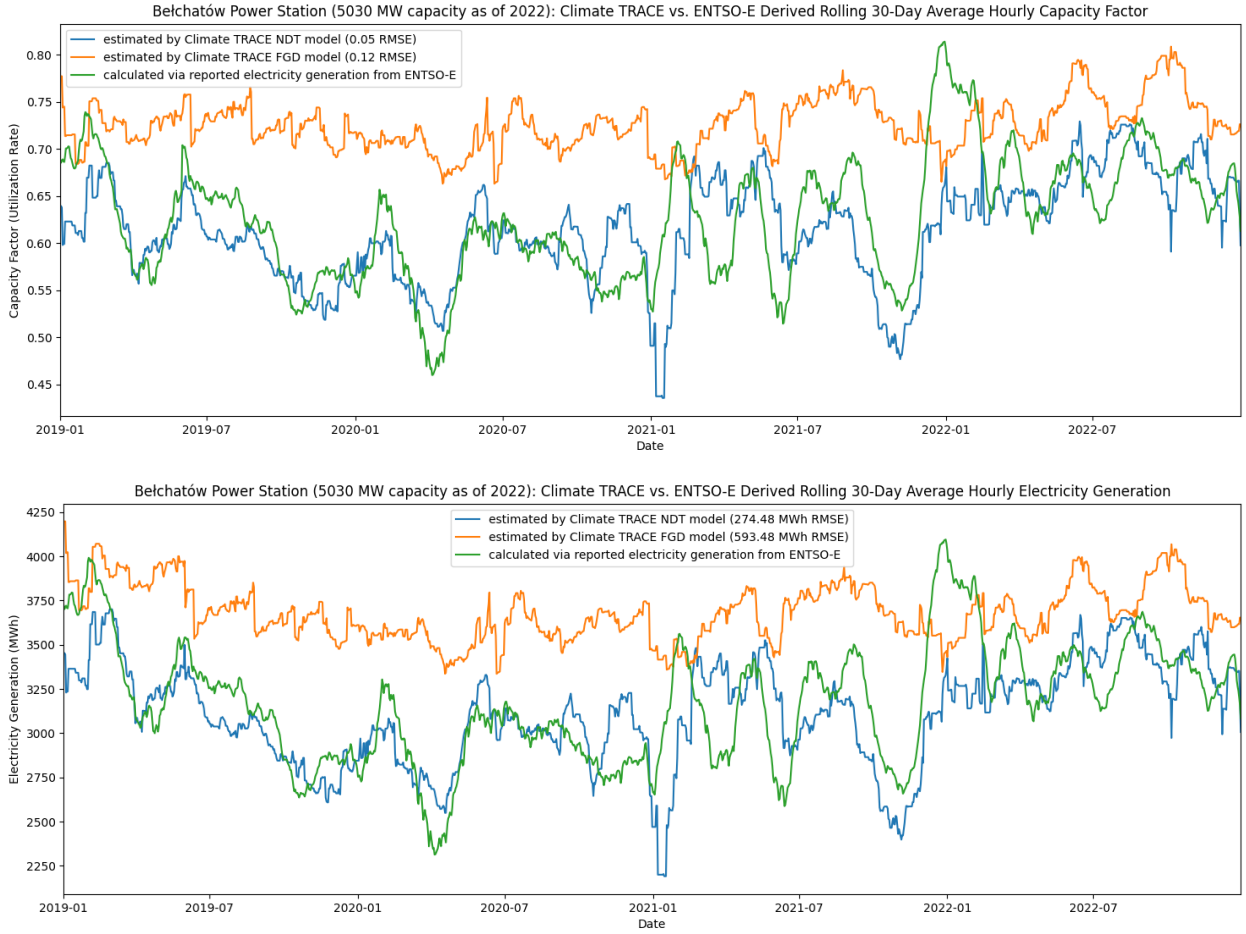


Figure 10 A sample of cross-validation predictions for a single plant from both our NDT and FGD generation models (which predict one value for each day, that value being the average hourly capacity factor over the preceding 30 days). The plant is Bełchatów in Poland, Europe's largest power plant, whose rolling average hourly 30-day capacity factor never fell below 45% capacity from 2019-2022 and was maintained above 60% for the entirety of 2022. The rolling-30-day-average hourly generation in MWh, inferred based on capacity factor, is displayed in the second plot. Note the NDT model performs over two times better than the FGD model for this plant, a representative characteristic of the NDT task given its clearer signal.

2.3.6 Cross-Validation

Since we aimed to estimate the emissions of power plants for which no reported generation data exists, it was essential to validate that our models generalize to plants that were not in the training set. To do this, we used four-fold cross-validation for model training, validation, and testing, while ensuring that all images of a single plant were contained within the same fold. We built a regular 1° x 1° grid in latitude-longitude space and placed plants in the same grid cell into the same fold.

For the sounding models, two folds of plants were used for training, one for validation, and one for testing. This process is illustrated in Figure 11. Since the generation models are trained on the outputs of other machine learning models - the sounding models - we must take extra care to avoid data leakage. Therefore, we adopted our cross-validation approach for the training and evaluation of the generation models. Each of the 4 generation model instances was trained with features based on soundings from the *validation* fold and evaluated with features based on the soundings from the *testing* fold. This way, both sets of sounding predictions come from the same sounding model and were not used for its training. While the sounding models might be overfit to their validation fold due to early stopping and hyperparameter tuning, the test fold was not used for any model optimization or selection. Thus, measuring performance on the test fold provides a reasonable representation of predictions on unseen plants.

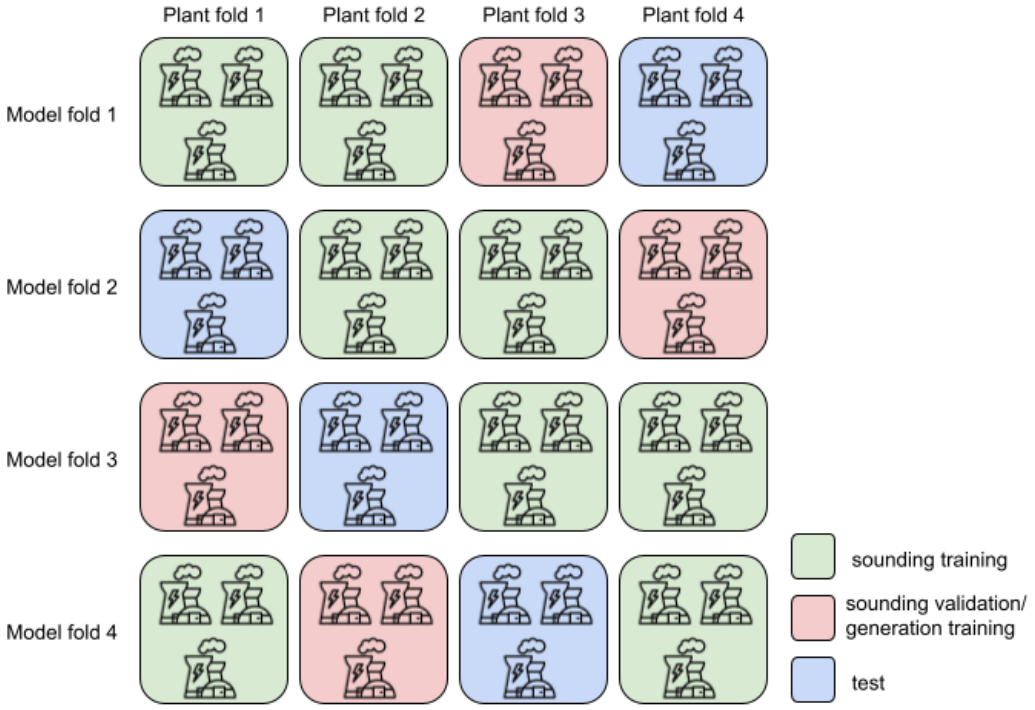


Figure 11 Diagram illustrating the construction and rotation of plant folds in our cross-validation loop.

2.3.7 ML Model Inference

To produce our final generation estimates, we followed the same data processing steps laid out in the previous sections, but with relaxed filters as described in Section 2.3.1.

Predictions for all images were generated over the period 2015-2022, using our sounding-level models described in Section 2.3.4. For inference, we averaged the sounding predictions from all

4 instances of a sounding model (one model for each fold). We then used these predictions to generate time series features and 30-day capacity factor predictions using the generation models as described in Section 2.3.5. Similar to the sounding models, we also averaged the predictions from all 4 instances of a generation model to obtain the final generation prediction. The result is a set of predictions, one per day per plant, of the average hourly capacity factor of that plant over the preceding 30 days. These are summed, weighted by how much the 30-day interval overlaps with the year, to produce estimates of the annual plant-level capacity factor.

We ran the sounding and generation models on NDT and FGD structures separately to estimate the activity of the entire plant. For plants that have only FGD or only NDT, the estimation process is straightforward (we simply use the prediction from the single applicable model), but, for plants with both NDT and FGD, we aggregated predictions from both model types by weighting the NDT model prediction two times more than the FGD model prediction, as the NDT models have lower error on average ((2*NDT + FGD)/3).

2.3.8 Country Level Baseline Capacity Factor Model

In addition to our machine learning models, we produced a second set of simpler baseline capacity factor estimates applicable to all of the plants in our dataset, regardless of type. We used EIA and EMBER Yearly Electricity Data country-level annual estimates of capacity and generation by fuel type to calculate the annual average fuel-specific capacity factor in each year reported for each country in the world. We then assumed the same capacity factors within each country for each part of the plant with the associated fuel-type.

These estimates were used for any plants where we do not have a capacity factor estimate from the ML models described above and are used in combination with our ML-based estimates where ML models are applicable.

2.3.9 Annual Emissions Factor Model

A power plant consists of one or more generating units, each of which may have a different fuel source and prime mover type, and therefore a different carbon intensity, from other units at the plant.

For each unit, an emissions factor was calculated through a combination of country-, fuel-, and prime-mover-specific average carbon intensities. Nominal carbon intensity values for combinations of energy source and prime mover technology were derived from a combination of US EPA Clean Air Markets Program Data (CAMPD), US EPA Emissions & Generation Resource Integrated Database (eGRID), and European Commission Joint Research Open Power Plants Database (JRC-PPDB-OPEN) data, and country specific calibration based on IEA data. The emissions factor of each unit was determined as follows:

1. A base value was gathered from the combination of the unit's energy source (e.g., coal, gas, oil, etc.) and prime mover technology (e.g., combined cycle, simple

cycle, etc.). This factor accounts for the typical efficiency differences between fuel and prime mover types.

2. If the combination of the energy source and prime mover did not have a value in the database, the average carbon intensity of the energy source was used.
3. The final emissions factor was calculated by applying a country calibration factor, a scalar that was multiplied to the base value to account for regional differences in power plant efficiency (due to age, technology level, size), fuel quality, and the impact of ambient conditions on carbon intensity that are not currently modeled.

2.3.10 Annual Capacity Factor, Generation, and Emissions Estimation

To create plant-level emissions estimates, first the estimated annual capacity factor of each plant is determined. For plants with no ML-based model, the country level baseline capacity factor (described in Section 2.3.8) was applied equally to all units in the dataset. For plants where generation model ensemble results (described in Section 2.3.5) are available and a plant was modeled by both NDT and FGD models, those results were combined with a weighted average with a weight of 2 for NDT and 1 for FGD. We then converted rolling 30-day average hourly capacity factor to annual hourly capacity factor by averaging, weighting each estimate by what fraction of the 30-day interval fell into the year in question. The annual ML-based result was then averaged with the country level baseline model and applied to each affected unit to predict the unit level annual hourly capacity factor.

The total unit level annual generation is the product of 1) the hourly capacity factor, 2) the unit capacity, and 3) the number of hours in the year. The unit level annual emissions are the product of 1) the generation and 2) the carbon intensity. Since CO_2 is the only gas currently tracked for the electricity sector, 20- and 100-year global warming potential (GWP) per plant per year were reported simply as equal to the CO_2 emissions, in metric tonnes.

Finally, the unit-level estimates were aggregated to the plant level to provide the annual source-level electric power sector emissions estimate and to the country level to provide the annual country-level total electric power sector emissions estimates.

2.3.11 Country Level Emissions

We published annual country level emissions predictions for 250 countries for the years 2015-2021. Country level estimates are aggregated from power plant unit-level emissions estimates.

2.3.12 Source Level Emissions

8333 unique power plants between 2019-2022 from across 169 countries, account for 12.34 GtCO2 in 2022 or 96% of total global fossil and waste fuel power emissions. This year, we report the estimates derived from our ML and/or country-/fuel-specific average models rather

than reported generation data, even if a plant has it, to ensure a uniform methodology is applied globally. This differs from our 2022 data release, as last year we reported emissions estimates derived from reported generation for the 48 plants that had reported generation. In addition to emissions estimates, we published a subset of inventory data relating to each plant, namely country, fuel type, latitude and longitude, capacity, and ownership. In the Climate TRACE source data reporting schema, sources that make use of the remote sensing based inference capacity factors are denoted as “i” and sources using only country-level capacity factors are denoted as “a" in the “other1” field.

2.4 Confidence and Uncertainty

2.4.1 Calibration Set

Since we lack high-resolution reported generation data globally (which is in fact the motivation for this project), we created a calibration set upon which to calculate uncertainties that we could propagate everywhere to estimate model uncertainties for every power plant in the world. This calibration set was composed of:

- Out-of-sample generation model predictions from our cross validation scheme described in Section 2.3.6 for training plants with reported generation data in CAMPD (U.S.), ENTSO-E (Europe), and NEM (Australia)
- Generation model predictions on plants excluded from training based on criteria described in Section 2.3.1 with reported generation data in CAMPD (U.S.), ENTSO-E (Europe), and NEM (Australia)
- Generation model predictions on plants with at least daily reported generation from EPİAŞ (Türkiye), NPP (India), and Taipower (Taiwan).

By binning all plants in the world according to the primary performance-discriminating factor outlined below in Section 2.4.2 (NDT vs. FGD), the generation model’s RMSE scores calculated on the calibration set within each bin were then propagated to all plants in the world that fit in that same bin. Our final bins and scores can be found in results Section 3.3.1.

2.4.2 Uncertainty and Confidence for Capacity Factor, Capacity, and Activity

With the awareness that our models (like all machine learning models) perform better under some circumstances than others, we set out to provide a quantitative assessment of model uncertainty for users. First, we identified the most prominent factors that contributed to differences in model performance by comparing model performance (mAP and misclassification rate for classification models, RMSE for regression models) across temperature, humidity, wind speed, NDT vs FGD, training vs. inference filter, number of FGD stacks and NDT towers, plant capacity, and sounding model disagreement. The primary factor that most differentiated ML model performance was whether the prediction was based on an FGD model alone or if it could pull from any model that used NDT, since our NDT machine learning models perform

substantially better than our FGD models, given the larger more obvious water vapor plume signal emitted by NDTs. The primary factor which differentiated performance in the non-ML country-/fuel-specific averages model was plant capacity; smallest power plants being hardest to predict. See results section 3.3.1 for the uncertainty and confidence breakdown for capacity factor.

Uncertainty for capacity was estimated to be 3% across all plants, receiving a confidence score of 4 (High, light green). Since activity (electricity generation) is calculated by multiplying capacity by capacity factor, the uncertainty was calculated as the square root of the sum of the squared fractional uncertainties for capacity and capacity factor, assuming independent random errors. This is because, for any x and y with independent random errors a and b, the error in $z = xy$ is $c = z\sqrt{(a/x)^2 + (b/y)^2}$ (Taylor, 1997). Thus, for electricity generation $g = cf$ for capacity c and capacity factor f with independent random errors ϵ_c and ϵ_f respectively, the electricity generation uncertainty is $\epsilon_g = g\sqrt{(\epsilon_c/c)^2 + (\epsilon_f/f)^2}$. Activity (electricity generation) confidence was taken to be the average between capacity and capacity factor confidences, resulting in some set to "high" (those modeled with ML+satellites and with some amount of NDT) and the remaining majority set to "medium."

2.4.3 Uncertainty and Confidence for CO_2 emissions

CO_2 emissions factor uncertainty was estimated to be 0.25 across all plants. Similar to activity (generation) above, because CO_2 emissions m is calculated as the product between capacity c, capacity factor f, and CO_2 emissions factor e, the CO_2 emissions uncertainty ϵ_m may be propagated as the square root of the sum of the squared fractional uncertainties of each of these three factors: $\epsilon_m = m\sqrt{(\epsilon_c/c)^2 + (\epsilon_f/f)^2 + (\epsilon_e/e)^2}$.

CO_2 emissions factor confidence was set at "low" (2 on a 1 to 5 scale) for all plants. Confidence for CO_2 emissions and 20- and 100-year global GWP was then taken as the average between the confidence scores for capacity, capacity factor, and CO_2 emissions factor, which ended up being 3 = "medium" for all plants.

For 20- and 100-year GWP, since the GWP factor for CO_2 is 1 for both time horizons, the uncertainty and confidence for CO_2 emissions were directly used for 20- and 100-year GWP uncertainty and confidence, taken to be the uncertainty for as well.

3. Results

3.1 Sounding Model Performance

Model performance was calculated on test predictions aggregated across all four cross-validation folds. We also calculated 90% (alpha = 0.1) confidence intervals by block bootstrapping the test set (blocking on plants, i.e. resampling plants rather than images, to ensure plants were not split up). Results for 2015-2022 are displayed in Table 2. The naïve baseline mAP is 0.5 for a model that predicts the most prevalent class, or always "on." The regression model's naïve baseline predicts the training set's mean capacity factor, resulting in an RMSE of 0.33 for NDT and 0.34 for FGD.

Table 2 Summary of sounding-level model cross validation test performance for each satellite and plant type. Bootstrapped 90% confidence intervals show that our models all significantly (alpha = 0.1) outperform a naïve mean baseline of 0.5 mAP for all classification models and 0.34 RMSE for regression. NDT = natural draft cooling towers and FGD = wet flue gas desulfurization.

Imagery	Model	NDT		FGD	
		Instantaneous on/off classification mAP (90% confidence interval)	**Instantaneous capacity factor regression RMSE (90% confidence interval)**	**Instantaneous on/off classification mAP (90% confidence interval)**	**Instantaneous capacity factor regression RMSE (90% confidence interval)**
PlanetScope	CNN (RESISC)	0.930 (0.899, 0.963)	0.196 (0.186, 0.206)	0.886 (0.860, 0.914)	0.262 (0.244, 0.276)
	XGBoost	0.956 (0.932, 0.986)	0.209 (0.194, 0.221)	0.890 (0.865, 0.914)	0.297 (0.284, 0.308)
Sentinel-2	CNN (RESISC)	0.958 (0.941, 0.974)	0.203 (0.193, 0.213)	0.906 (0.881, 0.927)	0.267 (0.249, 0.282)
	CNN (BEN)	0.934 (0.905, 0.966)	0.219 (0.207, 0.232)	0.840 (0.805, 0.875)	0.258 (0.246, 0.269)
	XGBoost	0.934 (0.904, 0.979)	0.210 (0.198, 0.220)	0.890 (0.862, 0.917)	0.269 (0.258, 0.279)
Landsat8	CNN (RESISC)	0.903 (0.867, 0.942)	0.267 (0.250, 0.280)	0.824 (0.793, 0.855)	0.299 (0.289, 0.309)
	CNN (BEN)	0.865 (0.825, 0.899)	0.264 (0.247, 0.279)	0.813 (0.788, 0.837)	0.300 (0.288, 0.312)
	XGBoost	0.899 (0.866, 0.938)	0.243 (0.221, 0.259)	0.880 (0.856, 0.906)	0.285 (0.271, 0.298)

All of our models outperformed these baselines – by an especially wide margin on the simpler task of classification. NDT models generally did better than FGD because NDT tends to produce larger plumes. RESISC models use 10 bands and so are only used for Sentinel-2 and Landsat 8. In general, RESISC models performed slightly better than BEN, while RESISC and XGBoost overall fared similarly with RESISC sometimes outperforming XGBoost and sometimes vice versa, emphasizing the utility of an ensemble approach. PlanetScope and Sentinel-2 models performed comparably, while the Landsat 8 models generally performed worse perhaps due to the coarser (30m) spatial resolution. However, the results show that Sentinel-2's 10m resolution is sufficient and PlanetScope's 3m resolution did not substantially improve model performance. Still, PlanetScope's daily revisits offer a major advantage in temporal resolution over Sentinel-2 and Landsat 8, resulting in a higher number of total observations over weeks, months, and years.

3.2 Generation Model Performance

Table 3 summarizes the results of our two generation models. Here we measured performance using the RMSE between the true and predicted 30-day capacity factors in the validation set. The baseline used is again the average value of the target in the relevant training set: 0.42 for NDT plants and 0.48 for FGD plants. The 30-day averaging of the target has a smoothing effect, leading to a lower baseline RMSE compared to the sounding-level regression tasks. Our generation models combine multiple models, satellites, and timescales to outperform this baseline significantly.

Table 3 Summary of generation model cross validation performance for each plant type. NDT = natural draft towers and FGD = wet flue gas desulfurization.

Model	30-day average capacity factor regression RMSE (90% confidence interval)	30-day average capacity factor mean baseline RMSE
NDT linear generation model cross-validation over training set	0.148 (0.130, 0.162)	0.272
FGD linear generation model cross-validation over training set	0.196 (0.186, 0.205)	0.272

Here, we again observe better performance on NDT plants than FGD plants, though both comfortably improve on the baseline (statistically significant, alpha = 0.1). This difference between NDT and FGD performance is explained both by the better performance of the NDT sounding-level models, as well as the reduced number of sounding-level predictions going into the FGD models due to the weather cuts described in section 2.3.2, which lead to periods with very little sounding-level information for some plants. These results also demonstrate that by

aggregating to a coarser temporal resolution (from hourly generation at the sounding-level to rolling 30-day average generation), our error rate reduces substantially.

3.3 Annual-Level Results

3.3.1 Source-Level Annual Capacity Factor Uncertainty, Confidence, and Bias

Table 4 Annual Plant Capacity Factor RMSE, non-ML country- and fuel-specific averages

Plant capacity in (0, 100] MW	Plant capacity in (100, 1000] MW	Plant capacity > 1000 MW
RMSE = 0.338 Confidence = 2 (Low, orange)	RMSE = 0.250 Confidence = 2 (Low, orange)	RMSE = 0.218 Confidence = 2 (Low, orange)

Table 5 Annual Plant Capacity Factor RMSE, non-ML country- and fuel-specific averages

Both NDT and FGD* or NDT only	FGD only
RMSE = 0.152 Confidence = 4 (High, light green)	RMSE = 0.191 Confidence = 3 (Medium, yellow)

*Certain plants have both NDT and FGD so both models can be applied and results combined by taking the weighted average (2*NDT + FGD)/3, weighting NDT twice as much as FGD since the NDT models perform better.

3.3.2 Sources of Bias and Error

We conducted an error and bias analysis by comparing our capacity factor estimates with a calculated capacity factor based on reported data available from the U.S. (1,279 unique plants), Europe/ENTSO-E (332 unique plants, India (166 unique plants), Türkiye (45 unique plants), Australia (14 unique plants), and Taiwan (11 unique plants, though none met the criteria for the ML/satellites-approach). We focused on the capacity factor estimates as that is the most uncertain and independently-estimated variable we publish, and we derive all generation and CO_2 emissions estimates from it. We calculated error as root mean squared error (RMSE) and bias as mean bias error (MBE). We calculated the capacity factor by matching reported generation data to the power plants in our dataset. We used that reported generation to calculate a plant's total generation, while we used its capacity information in our database for the denominator. We assume that all data sources report equally accurately.

The left-hand plot of Figure 12 shows a near-zero bias for the machine learning approach (-0.04 MBE in capacity factor). However, upon removing plants from the U.S. and Europe (representing 93% of plants used to train the machine learning models and the majority of plants in the left-hand plot), the right-hand figure reveals that the ML/satellites-based approach is

biased low (-0.13 MBE in capacity factor).

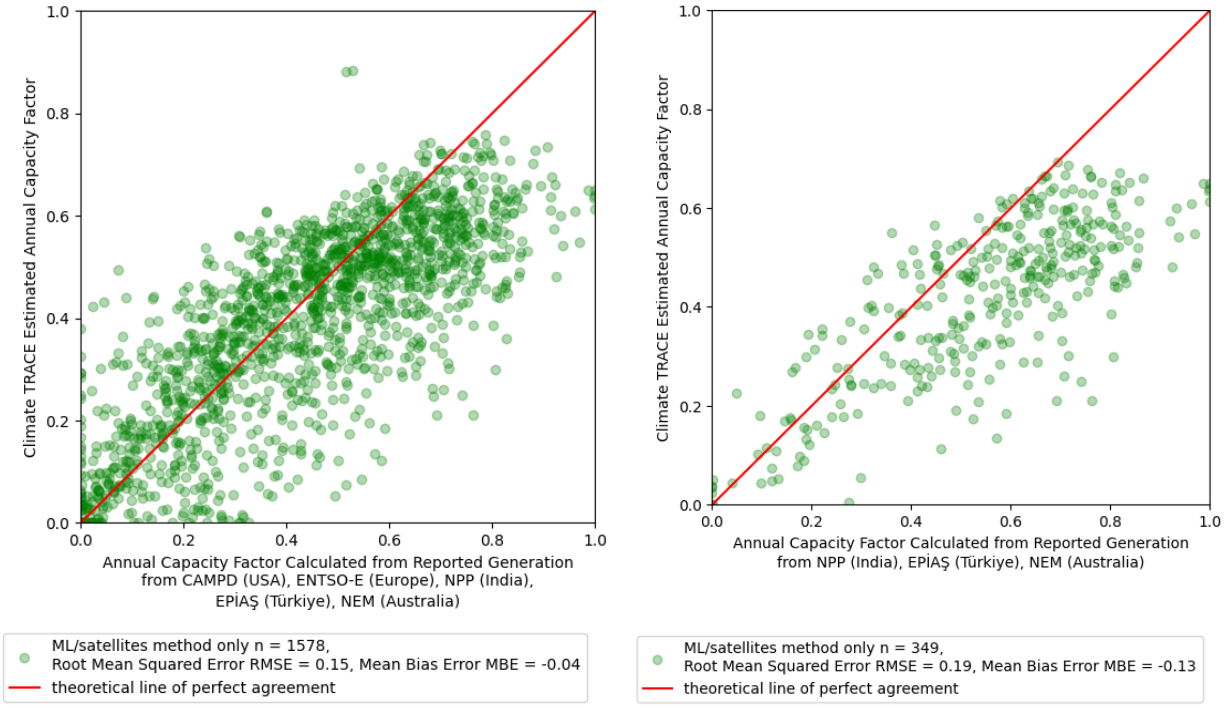


Figure 12. Each point represents the ML/satellites annual estimate for one power plant. The y-axis is Climate TRACE estimated capacity factor and the x-axis shows the calculated capacity factor using reported generation data. The two plots are identical except that the right plot removes plants from the U.S. and Europe, which represent the majority of training plants. It reveals that, assuming all countries report accurately, our ML/satellites-derived estimates may be biased low.

By comparison, the naïve estimation method using country-/fuel-specific averages, appears to overestimate slightly. Although the absolute bias is lower, note the higher RMSE and deviance from the line of perfect agreement, showing the weakness of this approach compared to the ML/satellites-based approach.

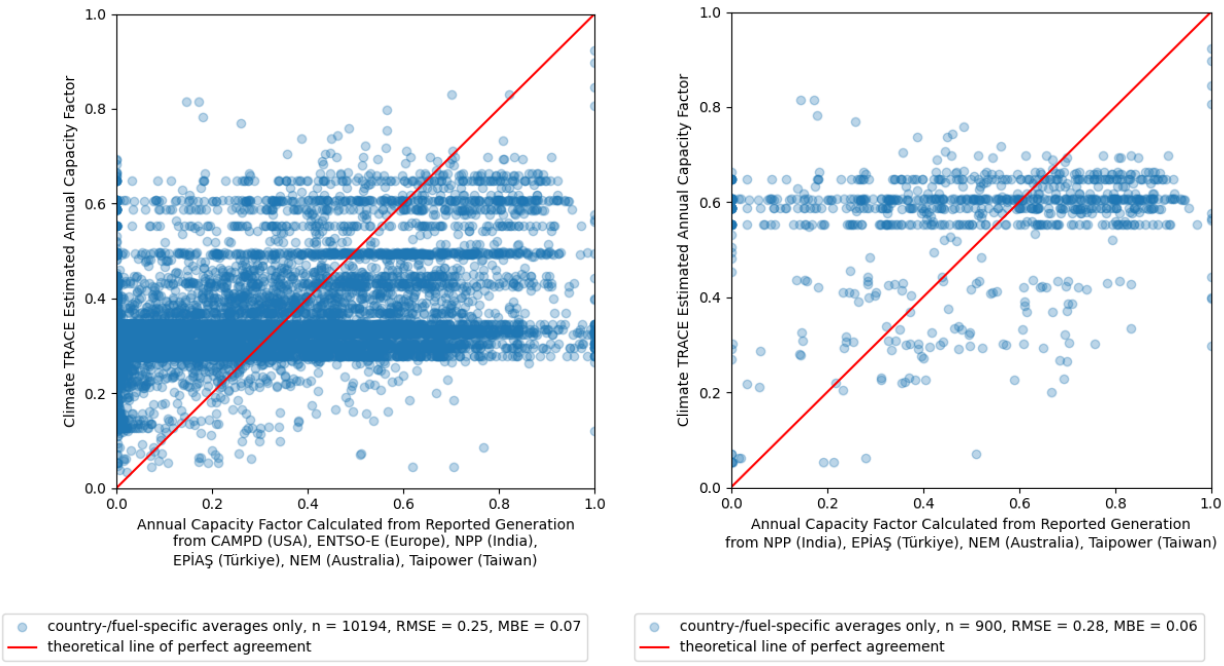


Figure 13. Each point represents the country-/fuel-specific averages method's annual estimate for one power plan. The y-axis shows Climate TRACE estimated capacity factor and the x-axis shows the calculated capacity factor using reported generation data. The two plots are identical except that the right plot removes plants from the U.S. and Europe.

We sought to mitigate both bias and error by combining the two approaches, where ML/satellite-based estimates were available. This reduces both bias and error, especially for outside the US and Europe where bias and error were worst, as shown in the plot below comparing all three approaches (compare green and orange points in plots below).

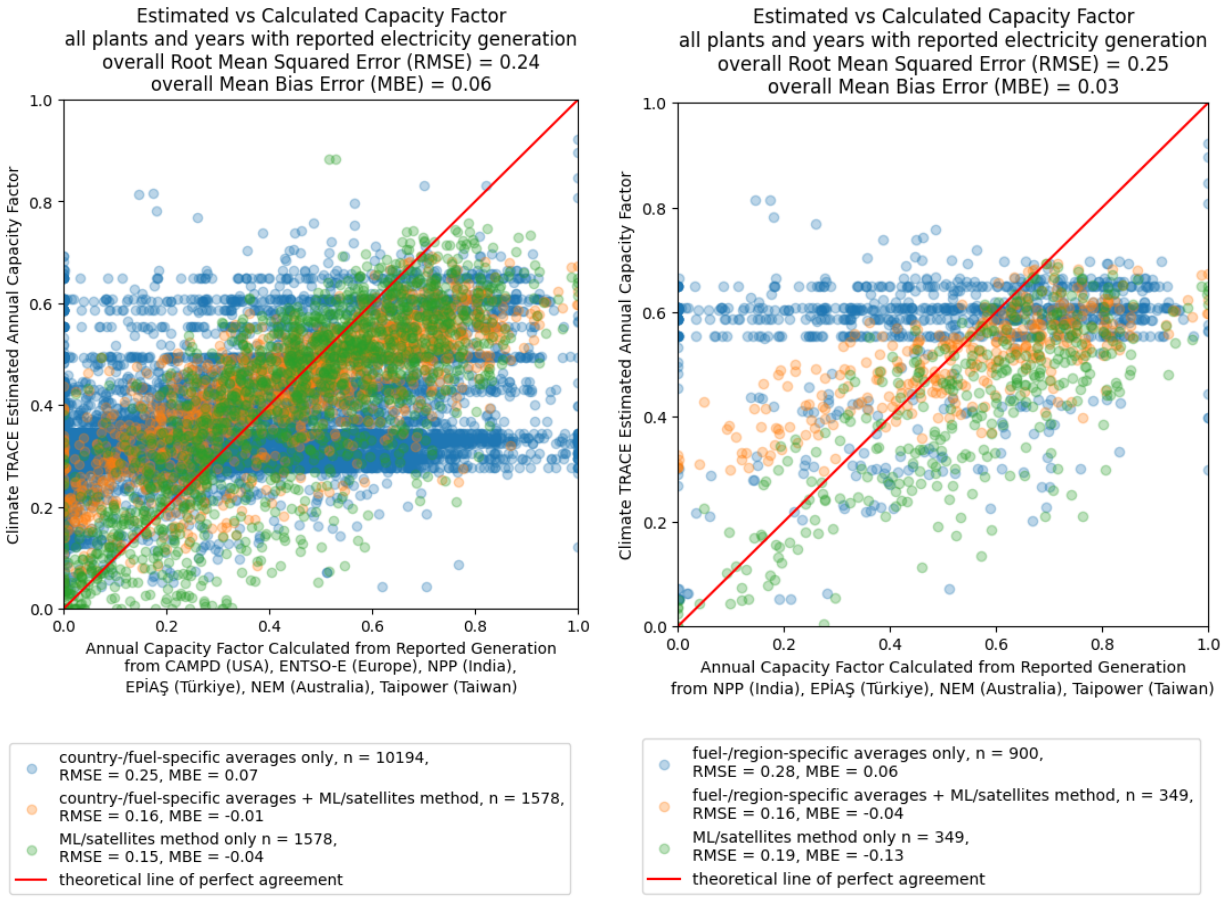


Figure 14. Each point represents the annual estimate for one power plant, using fuel-/country-specific averages (blue), ML/satellites (green) or a combination of the two (orange). The y-axis is Climate TRACE estimated capacity factor and the x-axis shows the calculated capacity factor using reported generation data. The two plots are identical except that the right plot removes plants from the U.S. and Europe, which represent the majority of training plants. This demonstrates that by combining the two approaches where possible, bias and error are both reduced.

We sought to mitigate bias as described above and elsewhere in our estimation pipeline. We also seek to be as transparent as possible about known sources of bias inherent in our machine learning and satellites modeling approach (which contributed to 38% of CO_2 emissions or 12% of plants reported at the source-level) and the country-/fuel-specific averages approach.

Machine Learning and Satellites, Sources of Error and Bias:

1. Because our models assume continuous wet FGD usage, if a plant is mislabeled and has dry FGD instead of wet or does not run its FGD continuously, our models will tend to underpredict emissions. We manually filtered out the most obvious examples of plants that showed no FGD signal but had other signs of activity from our FGD models and in those cases relied on the NDT model if possible or otherwise used the country + fuel type

imputation. However, there could be plants we missed and therefore underestimated their emissions.

2. For simplicity, our current filtering approach is not time-dependent and assumes fixed plant characteristics, yet plants may change over time and add or remove FGD or NDT units. A plant may meet inference criteria one year and not another, yet we treat all years the same. We will correct this in a future release to account for such subtleties. To address this issue for the time being, we filtered out plants that did not meet the inference criteria for 1 or more years during the period of analysis (2015-2022 inclusive). However, we did not exclude plants that only failed the filter for one year because they retired midway through that year.
3. The satellites' local overpass times are during the daytime and in the morning, averaging around 11am local time for all three sensors: Landsat 8, Sentinel-2, and PlanetScope (Figure 15). Therefore, all soundings, and thus all ML-based predictions, are based on images from this local time window. If a power plant generates more/less electricity at times when the satellites do not capture the plant, our generation estimates are biased low/high, respectively. Furthermore, since our generation model is currently trained on only the US, Australia, and Europe, it essentially learns how mid-morning power plant snapshots predict rolling 30-day average generation in those regions (especially the U.S. and Europe, as 93% of training plants are in these regions). Our method is therefore vulnerable to over- or under-prediction in regions of the world where dispatch patterns differ significantly from those in the training regions. We are working to remedy this source of bias by both expanding our training set to include more regions and by investigating additional proxy signals to augment our current NDT and FGD signals with a more complete view of each power plant. This may have been a contributing factor in our machine learning models' negative bias in India, Türkiye, and Australia.

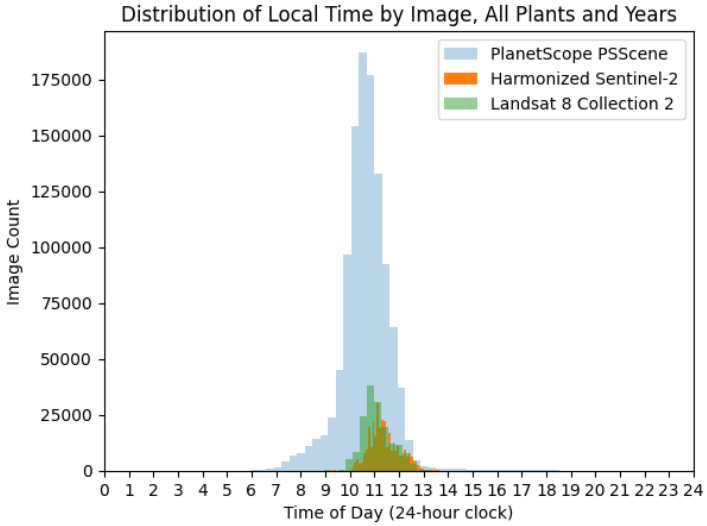


Figure 15. Distribution of local overpass time per image, by sensor.

4. The only proxy signal we currently use to estimate capacity factor is water vapor plume, whose size is sensitive to temperature and humidity: cold and wet conditions favor large plumes, dry and hot conditions result in smaller or fainter plumes. We currently mitigate this by having a temperature and humidity filter for both training and inference for FGD models, since FGD plumes are especially at risk of disappearing in hot and dry conditions compared to their larger NDT counterparts. Still, this means that regions that are too hot and dry to pass the filter will lack observations for models to ingest, such that we are forced to make predictions off of less data (this is one reason we widened the overly-conservative FGD temperature and humidity filters used last year). However, even if observations pass the filters, regions that are hotter and drier on average are at risk of underprediction. This may have been a contributing factor in our models' negative bias in India, Türkiye, and Australia. Adding additional proxy signals not as sensitive to local weather will be another way to mitigate this source of bias, and this is an area we are actively working on.
5. The majority of PlanetScope satellites were launched in 2017 or later and Sentinel-2B was launched in 2017, likely making satellite-derived estimates in years 2015 to 2017 less accurate due to the limited satellite coverage.

Fuel-/Country-Specific Averages, Sources of Error and Bias:

1. Assumes all plants run at the average country-level and fuel-specific capacity factor. This does not account for the typical variation in dispatch for plants serving baseload, intermediate load, and peaking load.

3.3.2 Annual Global and Country-Level CO_2 Emissions Estimates

Table 6 provides a sample of our country level emissions estimates by providing the top five highest-emitting countries, while Figure 16 below displays the top 10. This table includes the percent of emissions derived from satellite data. China is the largest emitter for 2022, about three times higher than the USA. However, factoring in population, the U.S. contributes the most per capita among nations with over 100 million people, contributing the most CO_2 from fossil and waste fuel power plants per capita in 2022: 4.5 metric tons of CO2 (tCO2) per capita compared to China's 3.3 tCO2 per capita in 2022. Emissions data for all 250 countries for 2015-2021 can be downloaded in full from the Climate TRACE website.

Table 6 The five largest electric power emissions estimates by country for the year 2022.

Country	2022 Climate TRACE estimated emissions in megatons CO_2 (Mt CO_2)	2022 Climate TRACE estimated emissions (tonnes CO_2 per capita)	% of Capacity ML + Satellite-Derived	% of emissions ML + Satellite-Derived
China	4,678	3.31	64%	64%
USA	1,496	4.49	16%	30%
India	1,249	0.88	28%	27%
Russia	559	3.90	25%	22%
Japan	440	3.52	12%	17%

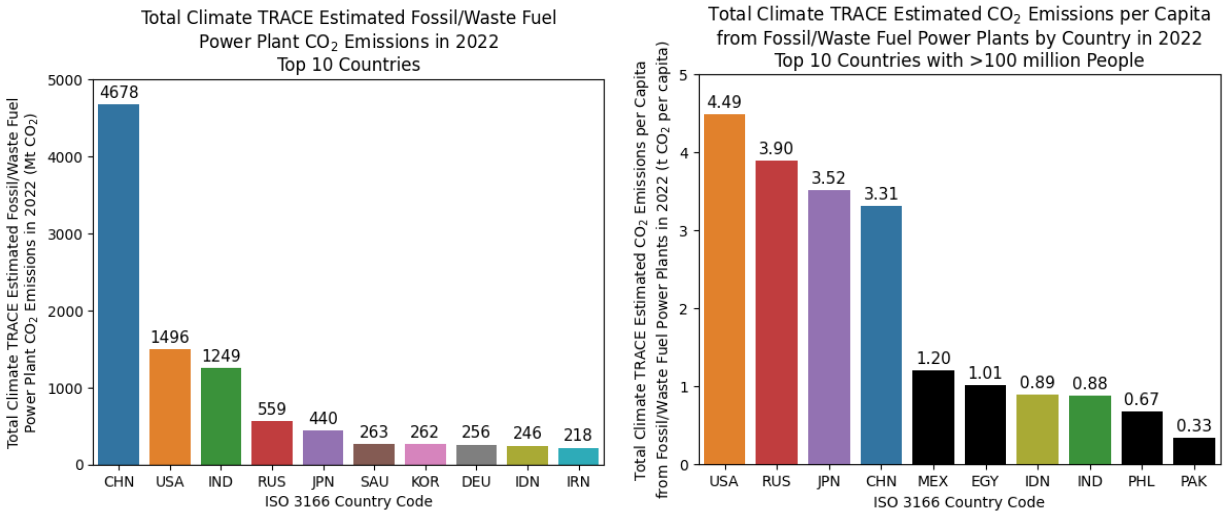


Figure 16. Total 2022 CO_2 emissions from fossil and waste fuel electricity generation by country for the top 10 and per capita for the top 10 countries with a population over 100 million.

3.3.3 Annual Source-Level CO_2 Emissions Results

Table 7 displays a sample of our source level emission estimates, showing the top five CO_2 emitters for 2022, while Figure 17 below shows the top 10. The highest-emitting power plant is Taichung in Taiwan. The top five emitting power plants are all located in Asia, specifically Taiwan, South Korea, Russia, and China. Coal-fired Bełchatów, the largest power plant in Europe, technically comes in at number 7 on the 2022 top 10 list, but roughly ties for 5th place with two Chinese coal-fired plants (Togtoh and Huadian Laizhou). Two coal-fired power plants in India (Vindhyachal and Mundra) round out the top 10 list alongside coal-fired Mailiao in

Taiwan. The complete facility-level estimates for power plants worldwide for 2019-2022 can be downloaded from the Climate TRACE website.

Table 7 The five power plants with the largest Climate TRACE estimated emissions in 2022.

Source Name	Capacity (MW)	Country	Emissions Estimate Method	Fuel type	Emissions ($MtCO_2$)
Taichung power station	5834	Taiwan	Country + Fuel Type Imputation	coal, oil	32.698
Taean power station	6480	South Korea	Country + Fuel Type Imputation	coal, other fossil fuel	32.237
Surgut GRES-2 power station	8898	Russia	Country + Fuel Type Imputation	gas	31.521
Dangjin power station	6040	South Korea	Country + Fuel Type Imputation	coal	30.915
Togtoh power station	6720	China	Satellite + Machine Learning and Country + Fuel Type Imputation	coal	27.976

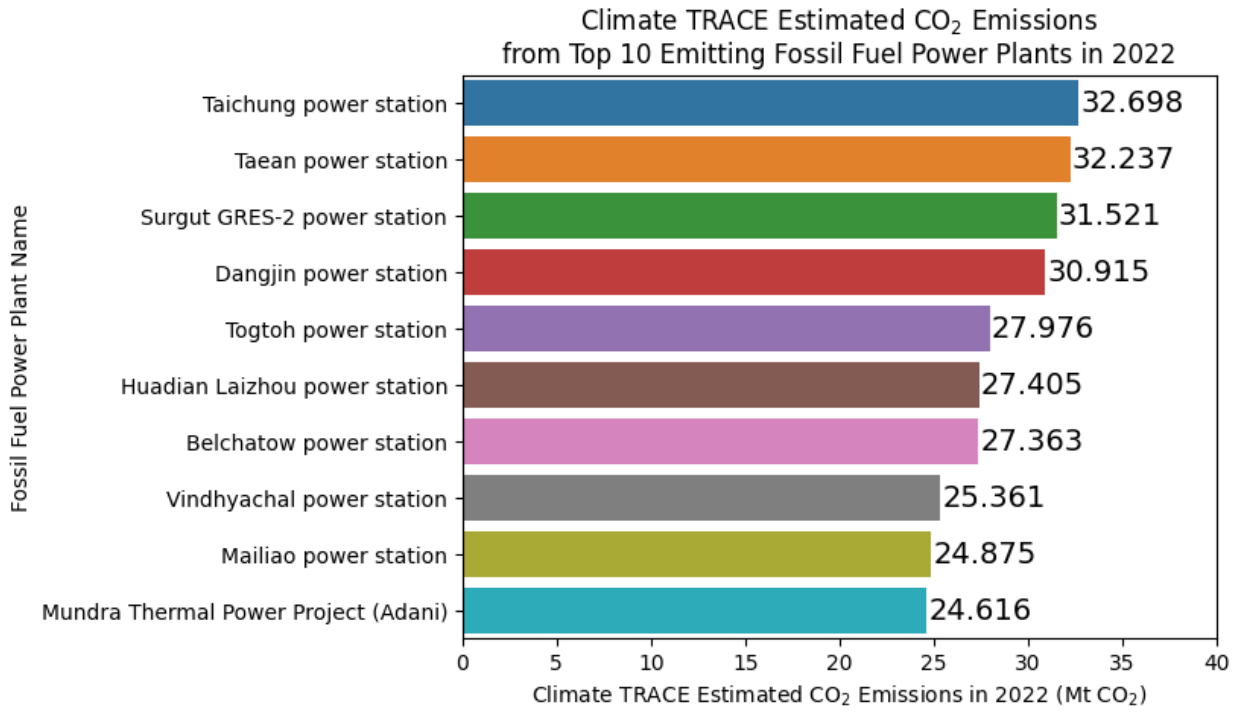


Figure 17. Total 2022 CO_2 emissions from the highest-emitting non-biomass power plants in 2022, all of which use fossil fuels.

4. Discussion

4.1 Global- and Country-Level Estimates and Comparisons to Existing Data

4.1.1 Existing Emissions Inventories

There are a variety of emissions inventories that include the energy sector that make both global and country level emissions estimates. Inventories available for energy include the Potsdam Institute for Climate Impact Research (PIK), United Nations Framework Convention on Climate Change (UNFCCC) National Submission, Emissions Database for Global Atmospheric Research (EDGAR) (Crippa et al., 2022), and CAIT. The Climate TRACE electricity generation sector currently covers emissions from combustion power plants with fuel sources including coal, natural gas, fuel oil, diesel, and waste for the period 2015-2022 for 250 countries. Biomass is not included in the country totals and is listed separately at the source level (see Table S2) to avoid double counting with the Climate TRACE forestry and land use sector.

There are some challenges in comparing Climate TRACE estimates to other inventories. The UNFCCC includes a breakdown of the different sub-sectors of energy, with the sectors and time period available varying by country. For Annex 1 countries, the category 1.A.1.a 'Public Electricity and Heat Production' is included in the UNFCCC estimates, which is more comparable to our data. The broader category of 1.A.1 Energy Industries used for non-Annex 1 countries, which includes electricity generation, as well as other energy industries such as the production of fossil fuels, makes comparison difficult. Because data for Non-Annex 1 countries is sparsely available for a similar time period of comparison, and also includes non-electricity related emissions sources, it was not used to generate world-level comparisons.

Some inventories only publish emissions estimates that represent aggregations across many IPCC sectors. For example, PIK's category '1.A' is equivalent in coverage to UNFCCC's '1.A.1 Energy Industries'. Both of these categories are equivalent to IPCC's sector 1.A. However, UNFCCC also publishes separate estimates for subsectors within 1.A.1 Energy Industries. Climate TRACE's estimates for electricity generation are the most equivalent in coverage to two of these subsectors: UNFCCC's 1.A.1.a.i Electricity Generation and 1.A.1.a.ii Combined Heat and Power Generation. Because PIK only publishes 1.A and not separate estimates for each of the IPCC subcategories within 1.A, this means that Climate TRACE's estimates for electricity generation cannot be compared *directly* to PIK's 1.A, as they are not equivalent in scope. However, to approximate a comparison, UNFCCC's estimates for the other subsectors in 1.A can be subtracted from PIK's estimate for 1.A. The result of this subtraction can then be used to make a rough comparison between Climate TRACE's estimate for electricity generation and PIK's estimate for the same. However, one caveat for interpreting these approximated comparisons is that the calculation assumes that PIK's estimates for the subtracted subsectors are identical to UNFCCC's – which is unlikely to be true. Thus, these calculated comparisons can

provide a general sense of whether there is a large difference between Climate TRACE's estimate and PIK's, but small differences may be attributable to the comparison being inexact.

Additionally, inventories can have different temporal coverages in terms of range and years sampled, which sometimes vary by country. The UNFCCC's best coverage is for Annex I countries, like Germany which has yearly data going back decades with the most recent data point covering 2020. PIK has more granular and recent coverage for non-Annex 1 countries, for example with yearly estimates from India most recently in 2021 and going back decades. However, as we require UNFCCC subcategory data to approximate a comparison with our electricity emissions estimates, the PIK data's comparability is limited by UNFCCC's temporal granularity and recency. CAIT has yearly estimates from 1990-2020 and EDGAR has yearly estimates from 1970-2022.

Lastly, some data sources rely on some of the same data sources for their analysis processed and combined in different ways. For example, our data (through EMBER), CAIT, and EDGAR rely on IEA data for parts of their analysis. On a country level, therefore, there is a large variation in which data sets are available and in which time range for comparison.

4.1.2 Global Emission Estimates Compared to Other Inventories

Figure 18 highlights global Climate TRACE and EDGAR totals and Annex I emissions. For both plots, we focus on comparing Climate TRACE to EDGAR since both have a full record that can be compared. The top chart, global, EDGAR reports higher emissions, an average of 13.95 Gt, whereas Climate TRACE reports a slightly lower average value, 12.23 Gt. This may be in part because the Climate TRACE power sector country totals do not include emissions from biomass fuels, to avoid double-counting with the Climate TRACE forestry and land use sector. EDGAR total global emissions for "Public Electricity and Heat Production" steadily increased from ~13.4 to ~14.6 Gt for 2015 to 2022, about an 8% increase. Similarly, Climate TRACE emission estimates report an increase from ~11.8 to ~13 Gt from 2015 to 2022, respectively, a 10% increase. Both EDGAR "Public Electricity and Heat Production" data and Climate TRACE emission estimates show the pandemic-related decrease in emissions in 2020, and the rebound in 2021 as demand for electricity increased as a result of re-openings. This is similarly reported in Liu et al. (2020) and Davis et al. (2022).

In contrast to the global emissions, 26 Annex I countries emissions display a decreasing trend for all inventories prior to 2021 (Figure 18, bottom chart). After 2021, both EDGAR and Climate TRACE both show emissions rebound as seen globally. Climate TRACE emissions are 4.5 Gt in 2015, then 3.6 Gt in 2020, then increase to 3.8 Gt in 2022. EDGAR emissions are 5.1 Gt in 2015, then 4.1 Gt in 2020, then increase to 4.4 Gt in 2022.

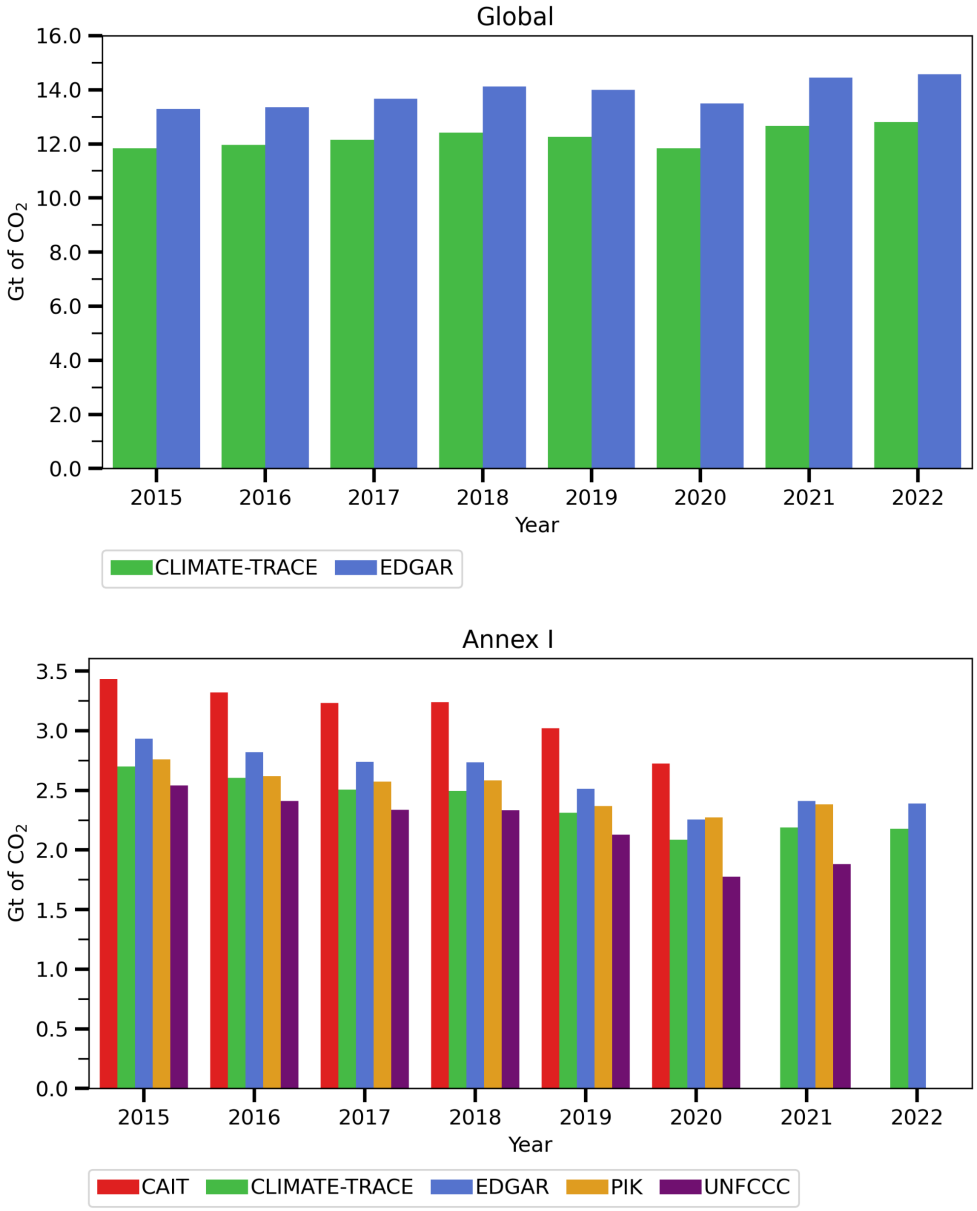


Figure 18. Top chart, total global emission estimates for electricity generation, reported in Gigatonnes (Gt; billions of metric tons) of CO_2. Climate TRACE = green bars, number of countries summed = 193; EDGAR = blue bars, number of countries summed = 180. Bottom chart emission estimates (Gt of CO_2) for 26 Annex I countries: BEL, BGR, CAN, HRV, CYP, CZE, FIN, GRC, HUN, ISL, LVA, LTU, MLT, NLD, NOR, POL, ROU, SVK, SVN, ESP, SWE, CHE, TUR, UKR, GBR, and USA. Note, UNFCCC did not report TUR emissions for 2020 and 2021. Inventories: CAIT (red bars), Climate TRACE (green bars), EDGAR (blue bars), PIK (orange bars), and UNFCCC (purple bars). CAIT, PIK, and UNFCCC emission estimates were only available up to 2020, 2021, and 2021, respectively.

Comparing Climate TRACE to both EDGAR 'Public Electricity and Heat Production' and UNFCCC 'Public Electricity and Heat Production' for Annex 1 show a similarity in trends and a lower estimate than EDGAR, also seen in the global totals. One possible explanation for these differences is that the Climate TRACE estimates do not include the emissions from heat production that is not associated with the generation of electricity. This is the case in China, Russia, and Europe where district heating infrastructure is prevalent (IEA Emissions Intensity Index 2021).

Since EDGAR and Climate TRACE have similar emissions values and encompass years 2015 to 2022, Figure 19 provides a log-scaled scatter plot comparing each individual country, color coded by geographic region. Select countries are labeled – the top ten highest emitting countries and the lowest emitting countries by geographic region. This provides a sense of the dispersion of emissions per country. China (CHN) is the largest emitter overall, followed by the United States (USA), India (IND), Russia (RUS), and Japan (JPN), matching Table 6. After Japan, the next five largest emitters include Saudi Arabia (SAU), South Korea (KOR), Germany (DEU), and Indonesia (IDN), and Iran (IRN). All these 10 countries fall close to the 1:1 line, showing agreement between EDGAR and Climate TRACE. As you go down the 1:1 line towards lower emissions, relatively smaller emitting countries scatter slightly further from the line, though this may simply be an artifact of the log scaling. Examples of these countries include Paraguay (PRY), Ethiopia (ETH), and Cook Islands (COK). African and South American countries - orange and red dots, respectively - tend to occupy the ~10^4 to ~10^8 metric tons of CO_2 (tCO_2) emissions range. Asian, North American, and Middle Eastern countries - blue, gray, and green dots, respectively - generally have emissions greater than 10^8 tCO_2.

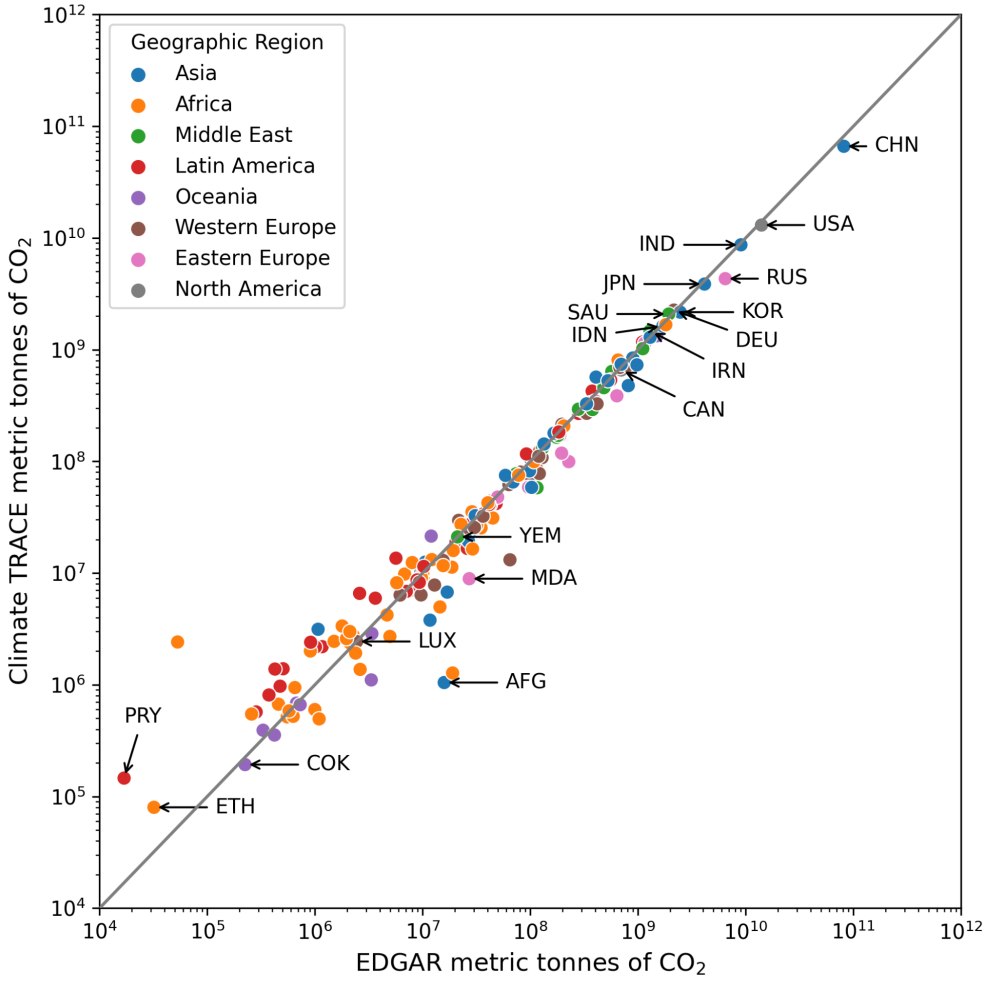


Figure 19. Scatter plot comparing overlapping countries between Climate TRACE and EDGAR emissions inventories. Each individual country's total metric tons of CO_2 (tCO_2) emissions from 2015 to 2022 is plotted on a log-scale. Each country's region is color coded, and the top 10 highest emitting countries and lowest emitting countries by geographic region are labeled. The black line represents the 1:1 line of perfect agreement between Climate TRACE and EDGAR.

4.1.3 Selected Country-Level Comparisons

Here Figure 20 presents country-level emissions estimates for the countries listed in Table 6 and include Laos and South Africa. These are compared to the existing emissions inventories described in section 4.1.1 where available, namely CAIT, UNFCCC, PIK and EDGAR. Comparisons are done using:

$$Percentage\ difference\ = 100\ x\ \frac{|\Sigma\ Climate\ Trace - \Sigma\ Other\ |}{(\Sigma\ Climate\ Trace + \Sigma\ Other)/2}$$

Where, the percentage difference between *Climate TRACE* emissions and *Other* emissions inventory (PIK, CAIT, UNFCCC, or EDGAR) emissions is the absolute value between the two emissions inventories and the average of the two emission inventories for overlapping years.

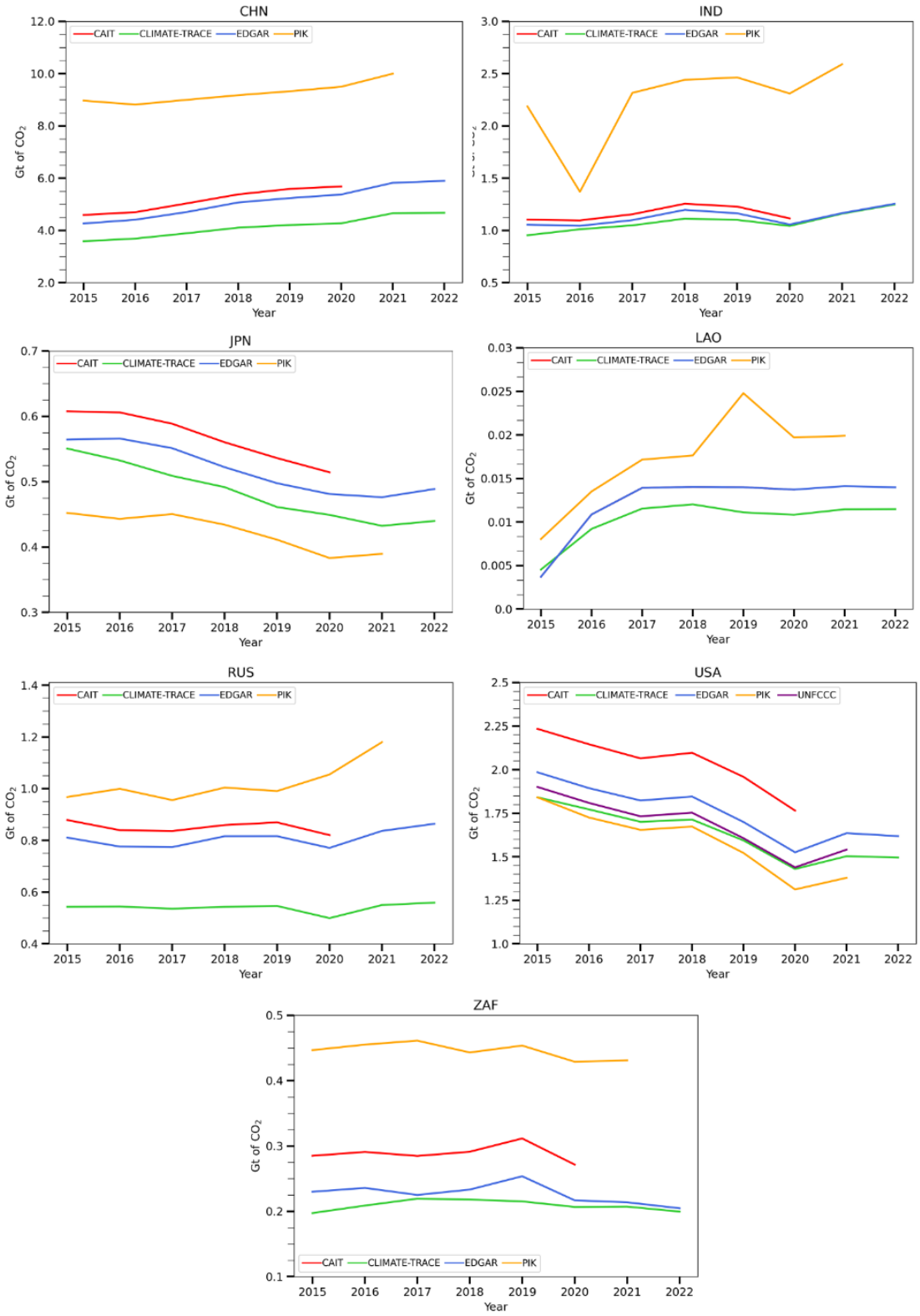


Figure 20. Inventories comparison for China (CHN), India (IND), Japan (JPN), Laos (LAO), Russia (RUS), United States (USA), and South Africa (ZAF). Each inventory has different date ranges based on the inventory's reporting year. Climate TRACE = green lines; CAIT = red lines; EDGAR = blue lines; PIK = orange lines; and UNFCCC = purple lines.

For China, for years where inventories overlap with Climate TRACE estimates, all show increasing emissions estimates from 2015 to 2020, with other inventories showing continued increase to 2022 (Figure 20). However, only Climate TRACE has the lowest emissions relative to the other inventories. Climate TRACE has a ~21%, ~26%, and ~78% overall emissions difference compared to EDGAR, CAIT, and PIK, respectively. For EDGAR, this dataset includes heat production, which will increase emissions estimates to be higher than ours. This is the case in all regions, such as China, Russia, and Europe where district heating infrastructure is prevalent (IEA Emissions Intensity Index 2021). We estimate that we include >95% of China's coal capacity in our inventory. Missing capacity can bias our estimates to be slightly lower given how unit capacities are core to both source- and country-level estimates. EIA and EMBER capacity and generation estimates show that plant-level capacity factors in China can be higher relative to the countries in our ML training set, which can lead to underestimation of capacity factors in plants covered by our ML models.

In India, Climate TRACE estimates broadly agree with EDGAR data, and both have similar emissions values for years 2015 to 2022 with a 4% overall difference (Figure 20). CAIT and Climate TRACE have an overall difference of ~10%. For the overlapping years, PIK has higher emissions estimates and a ~71% overall difference compared to Climate TRACE. CAIT, EDGAR, and Climate TRACE do not observe the emissions decrease in 2016 that PIK displays. All four inventories observe a slight emission increase in 2020, with Climate TRACE and EDGAR having a less steep increase, relative to PIK, that continues to increase into 2022.

In Japan, there is a wide range in inventory emissions estimates (Figure 20). Our estimates fall in the lower middle of this range. Climate TRACE emissions estimates overall are 7%, 13%, and 15% difference from EDGAR, PIK, and CAIT, respectively. All emissions inventories capture the emissions decrease starting in 2015. PIK captures Japan's emissions increasing in 2020, with Climate TRACE and EDGAR showing an increase later in 2021 that continues into 2022. Satellite-derived estimates do not contribute significantly to Climate TRACE Japan emissions estimates, so the differences observed are likely explained by disagreements in emissions factors or country-level capacity factors employed by either emission inventory.

Laos, along with many other smaller countries, have less recent inventory data to compare than countries with larger capacities. In some cases, smaller countries are not covered by large data providers like UNFCCC. Our global, satellite-based approach has the ability to cover countries which are more challenging to estimate using other techniques. For Laos, Climate TRACE emissions estimates have an overall difference of ~18% and 52% compared to EDGAR and PIK, respectively (Figure 20). All three emissions inventories capture the observed increase from

2015 to 2017. From 2017 to 2022, Climate TRACE and EDGAR emissions estimates plateau, whereas PIK displays a rapid increase in emissions in 2019 followed by a rapid decrease in 2020.

Russia is a country where there are large discrepancies between Climate TRACE estimates and other inventories' estimates (Figure 20). Climate TRACE estimates have an overall difference of ~40%, ~45%, and ~62% compared to EDGAR, CAIT, and PIK, respectively. CAIT, EDGAR, and Climate TRACE are similar in terms where emissions change little, about 0.1 to 0.2 Gt CO_2, from year to year in 2015 to 2019. Additionally, all three inventories capture the decrease observed in 2019 to 2020. Only PIK, EDGAR, and Climate TRACE capture the increase after 2020, with PIK displaying higher and steeper emissions increase relative to Climate TRACE and EDGAR estimates. Given Climate TRACE's reliance on generation estimates from EIA and EMBER and that we calibrate country-specific carbon intensities with published data, the relatively large difference in the Russian emissions estimates has several potential sources. One possibility is that Climate TRACE estimates are underestimating combustion power generation in Russia, resulting in lower emission estimates. Another could be that Russia has higher carbon intensities than we estimate. Finally, our inventory of Russian power plants may be incomplete. This hypothesis has a larger likelihood for Russia data than in regions such as the USA, Europe and Australia.

Climate TRACE predictions for the USA align well with other inventories (Figure 20). Climate TRACE estimates align closely with UNFCCC estimates for years 2015 to 2021, having an overall ~2% difference between the two. PIK and EDGAR have an overall ~4% and ~7% difference to Climate TRACE. CAIT and Climate TRACE have the largest overall difference of ~20%. All inventories display decreasing emissions from 2015 to 2020. From 2020 to 2021, UNFCCC, Climate TRACE, EDGAR, and PIK report increasing emissions.

Lastly, in South Africa, we predict significantly lower emissions than CAIT in the periods that the data overlap, an overall difference of 72%. Next, Climate and CAIT have a 31% difference between the two. EDGAR follows closely to Climate TRACE, with 2017, and 2020 to 2022 overlapping, and both having an overall difference of ~8%. All inventories, except Climate TRACE observe a pike in emissions in 2019, followed by a rapid decrease. Only Climate TRACE and EDGAR observe slightly decreasing emissions from 2020 to 2022.

In the selected countries in Figure 20, Climate TRACE emissions estimates are similar to other inventory emissions, such as India, United States of America, and South Africa. However, Climate TRACE emissions estimates disagree with other inventories, as is the case with Russia and China, reporting lower emissions relative to other inventories. Comparing Climate TRACE estimates to other inventories is challenging due to how other inventories rely on national data sources. An examination of existing global GHG inventories by Andrew (2020) suggests that while many exist, nearly all of them rely in no small part on the same original national data

sources. For example, the most complete inventory, EDGAR, is largely dependent on extrapolating from UNFCCC data, which greatly reduces its utility in cross-validating the accuracy of UNFCCC data. On the other hand, where we make satellite derived activity estimates, Climate TRACE contributes a new, independent method of assessing power plant activity via remote sensing and machine learning. While the satellite-derived estimates themselves are independent, downstream analysis and emissions factors required for the emissions estimates are not independent from other inventories. Still, having a more independent GHG inventory like Climate TRACE can offer support and insights on how well countries report their emissions to the UNFCCC.

4.2 Source-Level Estimates

Of the total source level inventory of 8,333 plants across 2019-2022 inclusive, 1,043 plants (accounting for 37% of CO_2 emissions in the four-year period) have their estimates derived from our machine learning approach combined with country-/fuel-specific averages, while the remaining 7,290 have their estimates based on country-/fuel-specific averages alone (Figure 21). This represents near-global coverage at the source level, encompassing 96% of global fossil and waste fuel power plant emissions as estimated by Climate TRACE from 2019-2022. In Figure 21, it can be seen that a large concentration of power plants in Eastern Europe, India, and East Asia have their emissions estimates augmented by the machine learning approach.

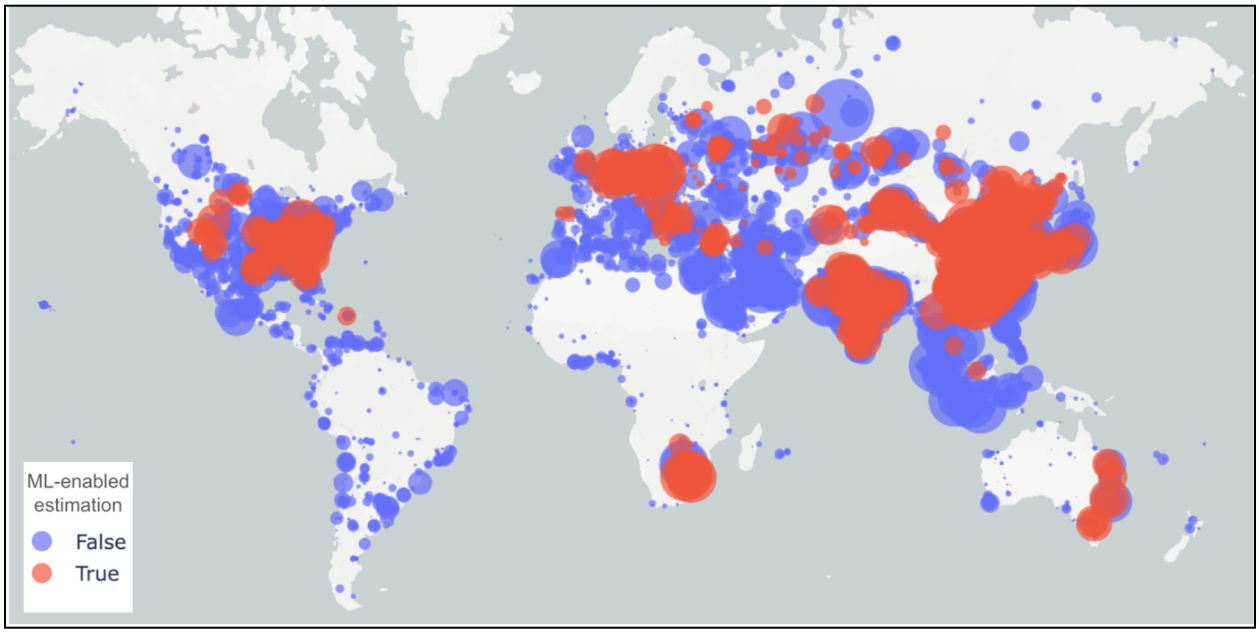


Figure 21. Map of all fossil and waste fuel power plants published in the Climate TRACE source-level data for 2022. The circle size is based on total 2022 CO_2 emissions as estimated by Climate TRACE and color-coded by estimation method: country-/fuel-specific averages (blue), country-/fuel-specific averages augmented by machine learning and satellite estimates (red).

The source-level data allows for more localized emissions analysis, such as that seen in Figure 22. The ten biggest CO_2 emitting fossil and waste fuel power plants have their CO_2 emissions plotted between the years of 2019 and 2022. As expected, all have a drop in emissions during 2020, given the COVID-19 pandemic, but only the Taean and Dangjin power stations in South Korea continued the emissions reduction path in 2021 and 2022 while the other power plants saw a post-pandemic rebound, with some (like those in India) increasing over their 2019 baseline emissions.

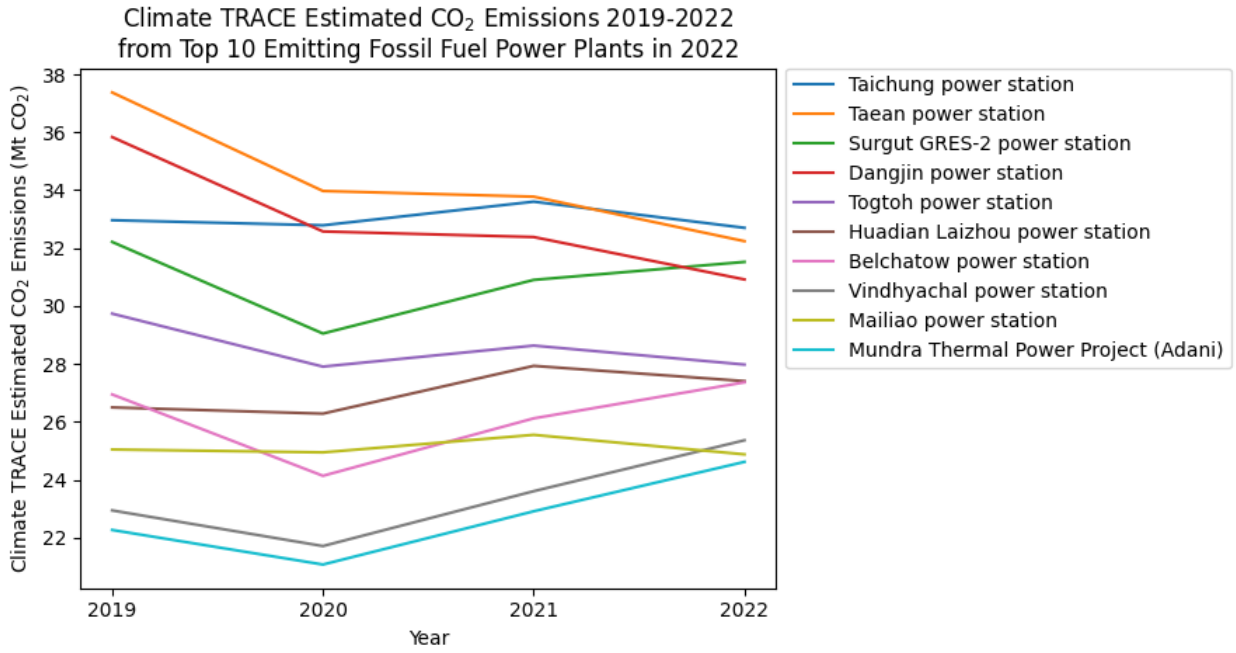


Figure 22. Estimated CO_2 emissions between the years of 2019 and 2022 inclusive for the ten highest-emitting non-biomass power plants in 2022, all of which use fossil fuels.

5. Conclusion

Monitoring individual power plants by applying ML models to satellite imagery makes it possible to build a picture of generation activity and emissions that is more uniform, recent, and detailed than other reported data. Our approach can be applied globally, including to regions where little to no generation or emissions data are publicly reported. By expanding publicly-available power plant emissions estimates on a frequent and plant-wise basis, this work can inform governments, corporations, and citizens to develop sustainability strategies for greater impact. Our approach is particularly impactful for sources for which no generation or emissions data publicly exists. The use of satellite imagery that is available in near real-time allows estimates to be prepared more quickly than other inventory methods and can track the emissions down to a source level.

Our CO_2 emissions estimates for power plants around the world are hosted by Climate TRACE. The inventory currently contains country-level annual emissions estimates for 2015-2022 and plant-level annual CO_2 estimates for 2019-2022 which covers 96% of Climate TRACE estimated fossil and waste fuel power plant emissions. Of these, 1043 unique plants from 2019-2022 were estimated using the remote sensing and machine learning approach described in Section 2, representing only 3% of power plants but responsible for roughly 37% of global fossil/waste fuel power plant CO_2 emissions 2019-2022.

We compared our Climate TRACE emissions estimates with other inventories at the global level, Annex I country aggregate level, and for selected countries. While there are some challenges in this comparison, the trends in our estimates are generally consistent with trends reported in other estimates. For both global and Annex I total emissions, Climate TRACE estimates were generally lower than other emissions inventories (Figure 18). One possible explanation for the lower estimates was that the Climate TRACE estimates do not include heat production emissions which are not associated with the generation of electricity and because our estimates do not include biomass fuels in the country totals to avoid double-counting with the Climate TRACE forestry and land use sector.

In the five largest emitting countries, Climate TRACE emissions estimates for USA, India, and Japan match similar trends and fall within the range of estimates compared to other inventories (Figures 19). However, Climate TRACE reports lower emissions for China and Russia than other estimates (Figure 19). These lower emissions may be due to incompleteness in our source inventory, an underestimation of total electricity generation from combustion power plants, and/or an underestimate of the carbon intensity of the plants in these countries.

We continue to refine and improve the accuracy and coverage of our predictions in an effort to provide plant-level emissions estimates for more power plants. This includes,

1. Improving our regression models by better understanding the relationship between plume size, generation, and weather conditions
2. Creating mechanisms to estimate model bias
3. Including new satellite observations
4. Sourcing additional reported data from regions outside the current training set to both validate and mitigate model bias
5. Investigating new proxy signals at plants that do not use NDT or FGD as well as signals widely-applicable to other fuel sources
6. Increasing the precision of the carbon intensity modeling of individual power plants.

This project is ongoing. We hope to share results from some of these avenues of research in a future update.

6. Supplementary materials metadata

Country-level emissions estimates for electricity generation are available for download at ClimateTRACE.org, and the following table summarizes this data.

Table S1 General dataset information for "country-climate-trace_electricity-generation_103123.csv"**.**

General Description	Definition
Sector definition	***Electricity Generation***
UNFCCC sector equivalent	***1.A.1.a.i Electricity Generation, 1.A.1.a.ii Combined Heat and Power Generation***
Temporal Coverage	*2015 – 2022 (inclusive)_*
Temporal Resolution	*Annual*
Data format(s)	*CSV*
Coordinate Reference System	*EPSG:4326, decimal degrees*
Number of sources available for download and percent of global emissions (as of 2022)	*8333 unique sources across 2019-2022, totaling 12.34 $GtCO_2$ in 2022, representing 96% of all fossil/waste fuel power plant emissions in 2022.*
Total emissions for 2022	*12.87 Gt CO_2*
Ownership	*Ownership data was obtained from Global Energy Monitor (GEM) and the U.S. Energy Information Administration (EIA) for the U.S.*
What emission factors were used?	*Carbon intensity values for combinations of energy source and prime mover technology were modeled from USA EPA, JRC data and IEA data.*
What is the difference between a "NULL / none / nan" versus "0" data field?	*"0" values are for true non-existent emissions. If we know that the sector has emissions for that specific gas, but the gas was not modeled, this is represented by "NULL/none/nan"*
total_CO2e_100yrGWP and total_CO2e_20yrGWP conversions	*Climate TRACE uses IPCC AR6 CO_2e global warming potentials (GWPs). CO_2e conversion guidelines are here: https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC_AR6_WGI_FullReport_small.pdf.*
Do the estimates include emissions from biomass fuels?	*No, due to concerns of overcounting since biomass is accounted for by the Climate TRACE forestry and land use sector. Electricity CO_2 emissions reported at the country level are from fossil/waste fuel electricity generation only. Emissions resulting from biomass are not included in the country totals but are available at the source-level for 2019-2022 (see Table S2 below) and upon request at the country level for 2015-2022. For plants that burn both fossil/waste and biomass, emissions are appropriately broken out based on the power plant's capacity of each fuel type.*

Source-level emissions estimates for electricity generation are available for download at ClimateTRACE.org, and the following two tables summarize this data.

Table S2 Source level metadata description confidence and uncertainty for "confidence-climate-trace_electricity-generation_103123.csv" and "uncertainty-climate-trace_electricity-generation_103123.csv"**.**

Confidence is defined on a 5-point scale from very low to very high:

- Very low (1); Purely assumption-driven or engineering estimates with details not verified by anything
- Low (2); Purely assumption-driven or engineering estimates with few details calibrated
- Medium (3); Estimated from a machine learning model applied outside the training data set on data with fairly similar physical characteristics
- High (4); Estimated from a machine learning model applied to the training data set or estimated
- Very High (5); Estimated by multiple independent sources with agreement

Data attribute	Confidence Definition	Uncertainty Definition
type	Confidence in fuel type reported. Set to 4 = High for all plants because data is sourced from the harmonized power plant inventory as described in Table 2	N/A
capacity_description	Confidence in capacity reported. Set to 4 = High for all plants because data is sourced from the harmonized power plant inventory as described in Table 2	Uncertainty in capacity reported. Estimated at 3% for all plants
capacity_units	megawatts (MW)	megawatts (MW)
capacity_factor_description	Capacity factor is electricity generation divided by power plant capacity. Capacity factor confidence was set to 4 = High for plants which used the NDT model and country-/fuel-specific averages, 3 = Medium for plants that did not use the NDT model but did use the FGD model and country-/fuel-specific averages, and 2= Low for plants that had access to only the country-/fuel-specific averages method.	Reported as the RMSE between predicted and reported capacity factor calculated over regions with reported generation data (India, Türkiye, US, Europe and Australia). For the ML approach, this was broken down by whether the plant was modelable with the NDT model or only by the FGD model. For the country-/fuel-specific averages approach, this was broken down by plant capacity.
capacity_factor_units	on a scale from 0 to 1 (proportion)	on a scale from 0 to 1 (proportion)
activity_description	uncertainty was calculated as the square root of the sum of the squared fractional uncertainties for capacity and capacity factor, assuming independent random errors. Thus, for electricity generation $g = cf$ for capacity c and capacity factor f with independent	average between capacity and capacity factor confidences, resulting in some set to 4="high" (those modeled with ML+satellites and with some amount of NDT) and

	random errors ϵ_c and ϵ_f respectively, the electricity generation uncertainty $\epsilon_g = g\sqrt{(\epsilon_c/c)^2 + (\epsilon_f/f)^2}$.	the remaining majority set to 3="medium."
activity_units	megawatts (MW)	megawatts (MW)
CO2_emissions_factor	2 = "low" for all plants	0.25 across all plants
CH4_emissions_factor	N/A	N/A
N2O_emissions_factor	N/A	N/A
other_gas_emissions_factor	N/A	N/A
CO2_emissions	taken as the average between the confidence scores for capacity, capacity factor, and CO_2 emissions factor, which ended up being 3 = "medium" for all plants	For capacity c, capacity factor f, and CO_2 emissions factor e, with uncertainties ϵ_c, ϵ_f, ϵ_e respectively, CO_2 emissions uncertainty ϵ_m for CO_2 emissions m may be propagated as the square root of the sum of the squared fractional uncertainties of each of these three factors: $\epsilon_m = m\sqrt{(\epsilon_c/c)^2 + (\epsilon_f/f)^2 + (\epsilon_e/e)^2}$
CH4_emissions	N/A	N/A
N2O_emissions	N/A	N/A
other_gas_emissions	N/A	N/A
total_CO2e_100yrGWP	same as CO_2 emissions, since the GWP factor for CO_2 is 1	same as CO_2 emissions, since the GWP factor for CO_2 is 1
total_CO2e_20yrGWP	same as CO_2 emissions, since the GWP factor for CO_2 is 1	same as CO_2 emissions, since the GWP factor for CO_2 is 1

Data citation format:

Freeman, J., Rouzbeh Kargar, A., Couture, H., Jeyaratnam, J., Lewis, J., Hobbs, M., Koenig, H., McCormick, C., Nakano, T., Dalisay, C., Davitt, A., Gans, L., Lewis, C., Volpato, G., and McCormick, G. (2023). *Electricity Generation Emissions Methodology.* WattTime, USA Transition Zero, UK, Pixel Scientia Labs, USA and Global Energy Monitor, USA, Climate TRACE Emissions Inventory. https://climatetrace.org [Accessed date]

Geographic boundaries and names (iso3_country data attribute): The depiction and use of boundaries, geographic names and related data shown on maps and included in lists, tables, documents, and databases on Climate TRACE are generated from the Global Administrative Areas (GADM) project (Version 4.1 released on 16 July 2022) along with their corresponding ISO3 codes, and with the following adaptations:

- HKG (China, Hong Kong Special Administrative Region) and MAC (China, Macao Special Administrative Region) are reported at GADM level 0 (country/national);
- Kosovo has been assigned the ISO3 code 'XKX';
- XCA (Caspian Sea) has been removed from GADM level 0 and the area assigned to countries based on the extent of their territorial waters;
- XAD (Akrotiri and Dhekelia), XCL (Clipperton Island), XPI (Paracel Islands) and XSP (Spratly Islands) are not included in the Climate TRACE dataset;
- ZNC name changed to 'Turkish Republic of Northern Cyprus' at GADM level 0;
- The borders between India, Pakistan and China have been assigned to these countries based on GADM codes Z01 to Z09.

The above usage is not warranted to be error free and does not imply the expression of any opinion whatsoever on the part of Climate TRACE Coalition and its partners concerning the legal status of any country, area or territory or of its authorities, or concerning the delimitation of its borders.

Disclaimer: The emissions provided for this sector are our current best estimates of emissions, and we are committed to continually increasing the accuracy of the models on all levels. Please review our terms of use and the sector-specific methodology documentation before using the data. If you identify an error or would like to participate in our data validation process, please contact us.

7. References

1. Andrew, R.M., 2020. A comparison of estimates of global carbon dioxide emissions from fossil carbon sources. *Earth System Science Data*, *12*(2), pp.1437-1465.
2. Bruckner, T., Bashmakov, I., Mulugetta, Y., Chum, H., de la Vega Navarro, A., Edmonds, J. , Faaij, A., Fungtammasan, B., Garg, A., Hertwich, E., Honnery, D., Infield, D., Kainuma, M., Khennas, S., Kim, S., Nimir, H.B., Riahi, K., Strachan, N., Wiser, R. and Zhang, X., 2014. Energy Systems. In: *Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* [Edenhofer, O., R. Pichs-Madruga, Y. Sokona, E. Farahani, S. Kadner, K. Seyboth, A. Adler, I. Baum, S. Brunner, P. Eickemeier, B. Kriemann, J. Savolainen, S. Schlömer, C. von Stechow, T. Zwickel and J.C. Minx (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.
3. Cheng, G., Han, J., Lu, X. Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE 2017, 105, 1865–1883.

4. Climate Watch Historical GHG Emissions. 2022. Washington, DC: World Resources Institute. Available online at: https://www.climatewatchdata.org/ghg-emissions
5. Couture, H., O'Connor, J., Mitchell, G., Söldner-Rembold, I., D'souza, D., Karra, K., Zhang, K., Kargar, A.R., Kassel, T., Goldman, B. and Tyrrell, D., 2020a. Towards tracking the emissions of every power plant on the planet. In *NeurIPS Workshop* (Vol. 3). Available: https://www.pixelscientia.com/pub/Couture-CCAI-NeurIPS2020.pdf
6. Crippa, M., Guizzardi, D., Banja, M., Solazzo, E., Muntean, M., Schaaf, E., Pagani, F., Monforti-Ferrario, F., Olivier, J., Quadrelli, R., Risquez Martin, A., Taghavi-Moharamli, P., Grassi, G., Rossi, S., Jacome Felix Oom, D., Branco, A., San-Miguel-Ayanz, J. and Vignati, E., CO2 emissions of all world countries - 2022 Report, EUR 31182 EN, Publications Office of the European Union, Luxembourg, 2022, doi:10.2760/730164, JRC130363
7. Cusworth, D.H., Duren, R.M., Thorpe, A.K., Eastwood, M.L., Green, R.O., Dennison, P.E., Frankenberg, C., Heckler, J.W., Asner, G.P. and Miller, C.E., 2021. Quantifying global power plant carbon dioxide emissions with imaging spectroscopy. *AGU Advances*, *2*(2), p.e2020AV000350.
8. Davis, S.J., Liu, Z., Deng, Z., Zhu, B., Ke, P., Sun, T., Guo, R., Hong, C., Zheng, B., Wang, Y. and Boucher, O., 2022. Emissions rebound from the COVID-19 pandemic. *Nature Climate Change*, *12*(5), pp.412-414.
9. Dos Reis, A.A., Werner, J.P., Silva, B.C., Figueiredo, G.K., Antunes, J.F., Esquerdo, J.C., Coutinho, A.C., Lamparelli, R.A., Rocha, J.V. and Magalhães, P.S., 2020. Monitoring pasture aboveground biomass and canopy height in an integrated crop–livestock system using textural information from PlanetScope imagery. *Remote Sensing*, *12*(16), p.2534.
10. Elavarasan, R.M., Shafiullah, G.M., Padmanaban, S., Kumar, N.M., Annam, A., Vetrichelvan, A.M., Mihet-Popa, L. and Holm-Nielsen, J.B., 2020. A comprehensive review on renewable energy development, challenges, and policies of leading Indian states with an international perspective. *IEEE Access*, *8*, pp.74432-74457.
11. EMBER 2022, *Ember Data Explorer, 2022*, Available on: https://ember-climate.org/data/data-explorer/ (Accessed 3 October 2022)
12. Ge, M.; Friedrich, J. 4 Charts Explain Greenhouse Gas Emissions by Countries and Sectors. World Resources Institute 2020. Available at: https://www.wri.org/blog/2020/02/greenhouse-gas-emissions-by-country-sector.
13. He, K., Zhang, X., Ren, S., Sun, J. Deep residual learning for image recognition. In Proc. CVPR, 2016.
14. Hobbs, M.; Kargar, A.R.; Couture, H.; Freeman, J.; Söldner-Rembold, I.; Ferreira, A.; Jeyaratnam, J.; O'Connor, J.; Lewis, J.; Koenig, H.; et al. Inferring carbon dioxide emissions from power plants using satellite imagery and machine learning. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, 2023.
15. IEA 2022. *Global energy-related CO2 emissions, 1990-2020***.** Available at: https://www.iea.org/data-and-statistics/charts/global-energy-related-co2-emissions-1990-2020 (Accessed 25 September 2022)

16. IEA Electricity Information 2022. *Electricity Information, July 2020***.** Available at: https://www.iea.org/data-and-statistics/data-product/electricity-information#documentation (Accessed 25 September 2022)
17. IEA Emissions Intensity Index 2021. *CO2 emissions intensity index for district heat production and heat production by country and region, 2021***.** Available at: https://www.iea.org/data-and-statistics/charts/co2-emissions-intensity-index-for-district-heat-production-and-heat-production-by-country-and-region-2021 (Accessed 3 October 2022)
18. Ilse, M., Tomczak, J., Welling, M. Attention-based Deep Multiple Instance Learning. In Proceedings of the International Conference on Machine Learning, 2018.
19. Kuhlmann, G., Broquet, G., Marshall, J., Clément, V., Löscher, A., Meijer, Y. and Brunner, D., 2019. Detectability of CO2 emission plumes of cities and power plants with the Copernicus Anthropogenic CO2 Monitoring (CO2M) mission. *Atmospheric Measurement Techniques*, *12*(12), pp.6695-6719.
20. Liu, F., Duncan, B.N., Krotkov, N.A., Lamsal, L.N., Beirle, S., Griffin, D., McLinden, C.A., Goldberg, D.L. and Lu, Z., 2020. A methodology to constrain carbon dioxide emissions from coal-fired power plants using satellite observations of co-emitted nitrogen dioxide. *Atmospheric Chemistry and Physics*, *20*(1), pp.99-116.
21. Liu, Z., Ciais, P., Deng, Z., Lei, R., Davis, S. J., Feng, S., Zheng, B., Cui, D., Dou, X., He, P., Zhu, B., Lu, C., Ke, P., Sun, T., Wang, Y., Yue, X., Wang, Y., Lei, Y., Zhou, H., . . . Schellnhuber, H. J. (2020). COVID-19 causes record decline in global CO2 emissions. *arXiv*. https://doi.org/10.1038/s41467-020-18922-7
22. Main-Knorn, M., Pflug, B., Louis, J., Debaecker, V., Müller-Wilm, U. and Gascon, F., 2017, October. Sen2Cor for sentinel-2. In *Image and Signal Processing for Remote Sensing XXIII* (Vol. 10427, pp. 37-48). SPIE.
23. Marchese, F., Genzano, N., Neri, M., Falconieri, A., Mazzeo, G. and Pergola, N., 2019. A multi-channel algorithm for mapping volcanic thermal anomalies by means of Sentinel-2 MSI and Landsat-8 OLI data. *Remote Sensing*, *11*(23), p.2876
24. Mia, M.B., Fujimitsu, Y. and Nishijima, J., 2017. Thermal activity monitoring of an active volcano using Landsat 8/OLI-TIRS sensor images: A case study at the Aso volcanic area in southwest Japan. *Geosciences*, *7*(4), p.118.
25. Moon, M., Richardson, A.D. and Friedl, M.A., 2021. Multiscale assessment of land surface phenology from harmonized Landsat 8 and Sentinel-2, PlanetScope, and PhenoCam imagery. *Remote Sensing of Environment*, *266*, p.112716.
26. Nassar, R., Hill, T.G., McLinden, C.A., Wunch, D., Jones, D.B. and Crisp, D., 2017. Quantifying CO2 emissions from individual power plants from space. *Geophysical Research Letters*, *44*(19), pp.10-045.
27. Paris Agreement to the United Nations Framework Convention on Climate Change; Number 16-1104, T.I.A.S., 2015.
28. Planet, 2022. Planet imagery product specifications. https://assets.planet.com/docs/Planet_Combined_Imagery_Product_Specs_letter_screen.pdf

29. Shikwambana, L., Ncipha, X., Malahlela, O.E., Mbatha, N. and Sivakumar, V., 2019. Characterisation of aerosol constituents from wildfires using satellites and model data: A case study in Knysna, South Africa. *International Journal of Remote Sensing*, *40*(12), pp.4743-4761.
30. Simonyan, K., Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 2014.
31. Sumbul, G., Charfuelan, M., Demir, B., Markl, V. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, 2019.
32. UNFCCC 2022. G*HG data from UNFCCC*. 1990-2020, Available at: https://unfccc.int/topics/mitigation/resources/registry-and-data/ghg-data-from-unfccc
33. Vaughn, T.L., Bell, C.S., Pickering, C.K., Schwietzke, S., Heath, G.A., Pétron, G., Zimmerle, D.J., Schnell, R.C. and Nummedal, D., 2018. Temporal variability largely explains top-down/bottom-up difference in methane emission estimates from a natural gas production region. *Proceedings of the National Academy of Sciences*, *115*(46), pp.11712-11717.
34. Taylor, J. Chapter 3. *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*, University Science Books, New York, 1997.