
Estimating Carbon Dioxide Emissions from Power Plant Water Vapor Plumes Using Satellite Imagery and Machine Learning

[Heather D. Couture](#)*, [Madison Alvara](#), [Jeremy Freeman](#)*, Aaron Davitt, Hannes Koenig, Ali Rouzbeh Kargar, Joseph O'Connor, Isabella Söldner-Rembold, André Ferreira, Jeyavinoth Jeyaratnam, Jordan Lewis, Colin McCormick, Tiffany Nakano, Charmaine Dalisay, Christy Lewis, Gabriela Volpato, Matthew Gray, Gavin McCormick

Posted Date: 24 February 2024

doi: 10.20944/preprints202402.1351.v1

Keywords: CO₂ emissions inventories; greenhouse gases; satellite observations; machine learning



Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Estimating Carbon Dioxide Emissions from Power Plant Water Vapor Plumes Using Satellite Imagery and Machine Learning

Heather D. Couture^{1,2,3,*†}, Madison Alvara^{1,3,†}, Jeremy Freeman^{1,3*}, Aaron Davitt^{1,3}, Hannes Koenig^{1,3}, Ali Rouzbeh Kargar^{1,3}, Joseph O'Connor^{3,4}, Isabella Söldner-Rembold^{3,4}, André Ferreira^{3,4}, Jeyavinoth Jeyaratnam^{1,3}, Jordan Lewis^{1,3}, Colin McCormick^{1,3,5}, Tiffany Nakano^{3,6}, Charmaine Dalisay^{3,6}, Christy Lewis^{1,3}, Gabriela Volpato^{1,3}, Matthew Gray^{3,4} and Gavin McCormick^{1,3}

¹ WattTime

² Pixel Scientia Labs

³ Climate TRACE

⁴ TransitionZero

⁵ Georgetown University

⁶ Global Energy Monitor

* Correspondence: heather@pixelscientia.com, jeremy@watttime.org

† These authors contributed equally to this work.

Abstract: Combustion power plants emit carbon dioxide (CO₂), a major contributor to climate change. Direct emissions measurement is cost-prohibitive globally, while reporting varies in detail, recency, and granularity. To fill this gap and greatly increase the number of power plants worldwide with independent emissions monitoring, we developed and applied machine learning (ML) models using power plant plumes as proxy signals to estimate electric power generation and CO₂ emissions using Landsat 8, Sentinel-2, and PlanetScope imagery. Our ML models estimated power plant activity on each image snapshot, and then an aggregation model predicted plant utilization over a 30-day period. Lastly, emissions factors specific to region, fuel, and plant technology were used to convert estimated electricity generation to CO₂ emissions. Models were trained with reported hourly generation data in the US, Europe, and Australia and were validated with additional generation and emissions data from the US, Europe, Australia, Türkiye, and India. All results with sufficiently large sample sizes indicate that our models outperformed baseline approaches. In conducting external validation to compare modeled versus reported annual generation and emissions in the countries where available, we calculated the root mean square error overall for modeled countries with validation data as respectively 1.75 TWh (236 plants across 17 countries over 4 years) and 2.04 Mt CO₂ (207 plants across 17 countries over 4 years). Ultimately, we applied our ML method to plants that comprise 32% of global power plant CO₂ emissions averaged over 2015-2022. This dataset is the most comprehensive independent and free-of-cost global power plant point-source emissions monitoring system currently known to the authors and is publicly available at climatetrace.org to support global emissions reduction.

Keywords: CO₂ emissions inventories; greenhouse gases; satellite observations; machine learning

1. Introduction

Responding to climate change requires timely and accurate measurement of greenhouse gas (GHG) emissions, especially CO₂. 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 [1]. The energy sector contributes the majority of GHG emissions globally. Depending on the data source, from 2015 to 2020, the energy sector contributed, on average, ~76% of emissions globally, which translates to between 33 and 37 GtCO₂ per year [2–4]. In total, the energy sector emitted

over 200,000 GtCO₂ during the six-year period [3]. Within the energy sector, the majority of GHG emissions originate from electricity generation of combustion power plants. This sub-sector accounts for an average of $\approx 44\%$ or ≈ 15 GtCO₂, representing $\approx 31\%$ of total global GHG emissions from 2015 to 2020 [2]. As temperatures continue to rise, a feedback cycle has been created between power plant emissions and temperature. For example, according to the World Meteorological Organization (WMO), the hottest years on record have been the recent nine years, 2015 to 2023, and heatwaves are expected to increase and become more intense [5,6]. When heat waves occur, cooling demands increase, and, as cooling demands increase, energy consumption increases, which leads to more power plant emissions that further accelerate the temperature rise [7,8]. To curb power plant emissions requires monitoring their activity.

Monitoring of power plant emissions varies globally and involves self-reporting. This includes the continuous emissions monitoring systems (CEMS) that are employed in countries with strict emissions laws, such as Japan, the US, the European Union (EU), and South Korea [9,10]. CEMS equipment is installed at individual power plants and measures emissions, providing reliable measurements to ensure the plant is complying with country emissions regulations. However, CEMS are costly, require specialized teams to maintain and calibrate, come in different forms that impact the measurement quality, and have limited deployment globally with some countries only deploying them at the largest emitting power plants [9–11]. A more common approach to monitoring emissions is bottom-up self-reporting. This approach quantifies power plant emissions using fuel consumption, fuel quality, and emission factors [12,13]. To use the bottom-up approach accurately and effectively requires quality data and, if not known, can have high uncertainties in fuel properties, which translates into high uncertainty in estimating power plant emissions [11,12,14]. Lastly, self-reporting varies by region in terms of recency and granularity. This creates challenges for policymakers in designing strategies to mitigate and reduce a country's emissions and, consequently, in meeting sustainability goals [11,15].

Within the last decade, efforts have been made to use GHG monitoring satellites and aerial surveys to infer and improve power plant emissions monitoring [11,14,15]. This includes studies that have derived CO₂ emissions from the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) and Global Airborne Observatory (GAO) aerial surveys, the Ozone Monitoring Instrument (OMI) on Aura, Orbiting Carbon Observatory-2 and -3 (OCO-2 and -3), and the PRecursoRE IperSpettrale della Missione Applicativa (PRISMA) satellites [10,11,14,15]. These studies have provided robust emissions estimates in a consistent way – not derived from the varying approaches described above. However, these studies have limited deployment, and some GHG monitoring satellites work best when observing some of the highest emitting power plants based on their sensor's CO₂ sensitivity and detection threshold [11,15]. Alternatively, to provide higher spatial and temporal resolutions for power plants, multi-spectral imaging satellites that observe Earth's land and ocean provide an opportunity to measure power plant emissions. Numerous multi-spectral satellites are in orbit, and one can combine the imagery from many of these satellites to produce robust and continuous global coverage, identifying more detailed power plant activity related to emissions.

Climate TRACE (Tracking Real-time Atmospheric Carbon Emissions) is a coalition of organizations working towards improved emissions monitoring (climatetrace.org). Climate TRACE members WattTime (watttime.org) and TransitionZero (transitionzero.org) developed the methods in this article that use multi-spectral satellite imagery combined with in situ (reported) generation data to train machine learning (ML) models to infer a power plant's generation via its activity. These models were trained using satellite imagery on plants in countries with hourly or sub-hourly generation data. Subsequently, they were applied globally to estimate power plant activity, which was combined with regional-, plant-type-, and fuel-specific carbon intensities to estimate plant-level CO₂ emissions.

2. Background

Several satellites measure CO₂ concentrations globally, including OCO-2, the Greenhouse Gases Observing Satellite (GOSAT), and PRISMA. These satellites use spectroscopic methods based on the

absorption of reflected sunlight to estimate the column-averaged dry-air mole fraction of CO₂, known as XCO₂.

Nassar et al. used XCO₂ retrievals from OCO-2 to quantify CO₂ emissions from large power plants, seven in the US and seven in other countries [16]. The narrow swath width of OCO-2 and cloud cover significantly limit the effective revisit time and applicability of this technique to snapshots of emission estimates - a total of twenty snapshot estimates across all fourteen of the study's power plants between the years 2014 to 2018. Other studies have also used OCO-2 and/or OCO-3 but with similar coverage limitations: one observation each of two plants [17], fourteen plumes from six plants [18], or fifty observations from twenty plants [19]. Cusworth et al. used XCO₂ retrievals from the AVIRIS Next-Generation (AVIRIS-NG) and GAO, as well as PRISMA, to quantify CO₂ emissions from seventeen US coal- and gas-fired power plants [11]. Several other researchers have also reported methods to use XCO₂ to estimate emissions from individual power plants [20–25]. These studies show that it is possible to estimate the emissions of large power plants with remote sensing and can provide an independent way to verify reported estimates. However, the small number of observations cannot create a comprehensive global power plant emissions monitoring system, and the detection limits of the approaches restrict their applicability to only the largest of power plants.

Future planned satellites, such as the Copernicus CO₂ Monitoring (CO2M) mission, aim to provide new opportunities for emissions monitoring at individual power plants [26,27]. However, such approaches may be years out and will lack the ability to track progress prior to their deployment, for instance quantifying the effectiveness of policy changes since the Paris Agreement.

Furthermore, despite the sensor- and orbit-specific differences of these studies, all methods that use satellite-retrieved XCO₂ face a similar set of challenges. First, CO₂ concentrations are affected by natural sources and sinks, particularly the oceans, soils, and forests. This background noise limits the sensitivity of XCO₂-based techniques to detect relatively low rates of anthropogenic emissions, such as those from lower-capacity power plants. Second, resolving emissions from closely adjacent sources is problematic without high spatial resolution. Third, the variability of emissions from some industrial sources can be quite high, requiring high revisit rates to observe patterns.

High-spatial-resolution (30m or less) multi-spectral imagery is another potential route to monitoring GHG emissions from individual sources. It has the important advantage of being available today at relatively high temporal resolution from many different satellites. Our project mainly uses this type of remote sensing data to build a "good enough, right now" emissions-monitoring system that does not need to wait for future satellites. Since multi-spectral imagery cannot directly measure CO₂ concentrations, we developed a set of proxies, or activity measurements, that are directly tied to emissions, such as visible vapor plumes from power plant cooling towers.

Prior work on this task includes a proof of concept by Carbon Tracker to estimate emissions of coal plants in the EU, US, and China using Planet Labs' satellite images [28]. Other efforts have been made to estimate power plant emissions using Sentinel-2 imagery and a multitask deep learning approach [29–31]. However, these studies were based on a heterogeneous dataset of European plants that included a variety of cooling types with inconsistent signals. Our own prior work expands to additional satellite constellations and power plant technologies but was focused on models to predict plant operational status (on/off and capacity factor) at a single point in time [32,33]. This article extends on our previous work by developing an ensemble model to aggregate data from across all satellite imagery sources and predict generation information for rolling 30-day intervals to produce annual CO₂ emissions estimates. We also validated our models on globally distributed plants from outside our training set.

3. Materials and Methods

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 vary, impacting

the visibility of flue gas plumes. For these reasons, directly inspecting smoke only provides a weak indicator of emissions. A better indicator of emissions is the vapor plumes from two primary sources:

- Natural draft cooling towers (NDT): Plants using NDT have a large hyperbolic structure that allows vapor plumes to form during evaporative cooling.
- Wet flue gas desulfurization (FGD): After desulfurization, flue gases become saturated with water, increasing the visibility of plumes from the flue stack.

In terms of size, NDT plumes are generally larger and wider than FGD plumes, making them easier to see in multi-spectral satellite imagery, as shown in Figure 1. A power plant may have one, both, or neither of these technologies. Due to the differing plume sizes and shapes from their sources, we created separate NDT and FGD models.

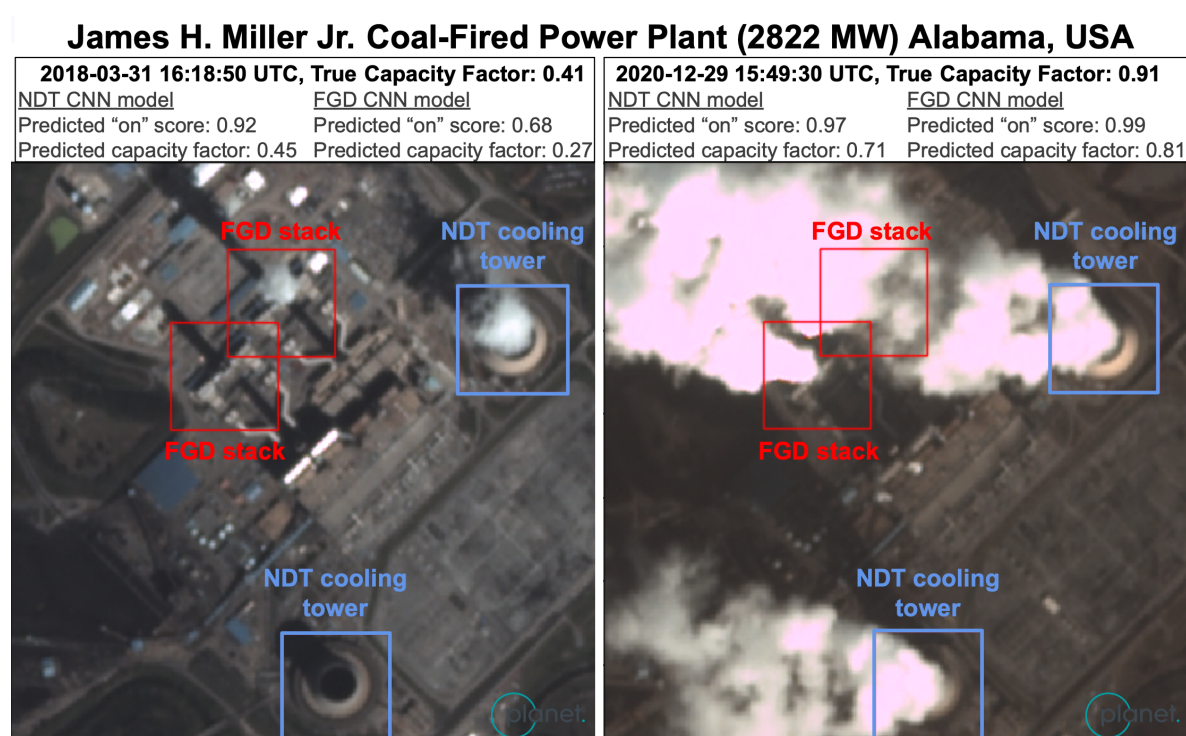


Figure 1. PlanetScope CNN predictions on the James H. Miller Jr. power plant at low vs. high generation on two observation dates. Separate NDT and FGD models predicted on NDT cooling tower (blue) and FGD stack (red) patches, respectively. These predictions were ingested by subsequent models to estimate generation, then CO₂, for the plant. © 2023 IEEE. Reproduced with permission from [33].

Observing the visible emitted water vapor plumes in multi-spectral satellite imagery, we built an ML modeling pipeline using gradient-boosted decision trees and convolutional neural networks (CNNs), then trained these ML models to infer a power plant's operational status. Specially, we designed models to perform two tasks:

1. Sounding-level models: consist of **A**) a classification model to classify if a plant was running or not (on/off), and **B**) a regression model to predict capacity factor (i.e. what proportion of the power plant's generation capacity was being used to generate power, generally between 0 and 1), for a given a satellite image of that plant at a certain point in time.
2. Generation models: aggregate the predictions from the sounding-level models into estimates of power plant capacity factor over the preceding 30 days.

Each sounding model was trained on satellite images paired with the reported generation status. After predicting the capacity factor, it was multiplied by capacity (maximum electric power output)

to infer generation. Our models are trained on power plants in countries for which we have hourly or sub-hourly generation data, which can be closely matched to the satellite image timestamp. Then the models were applied globally using country-, fuel-, and prime-mover-specific average carbon intensities to convert modeled plant-level generation to emissions estimates. Figure 2 provides an overview of how these different models and data sources are integrated to estimate emissions.

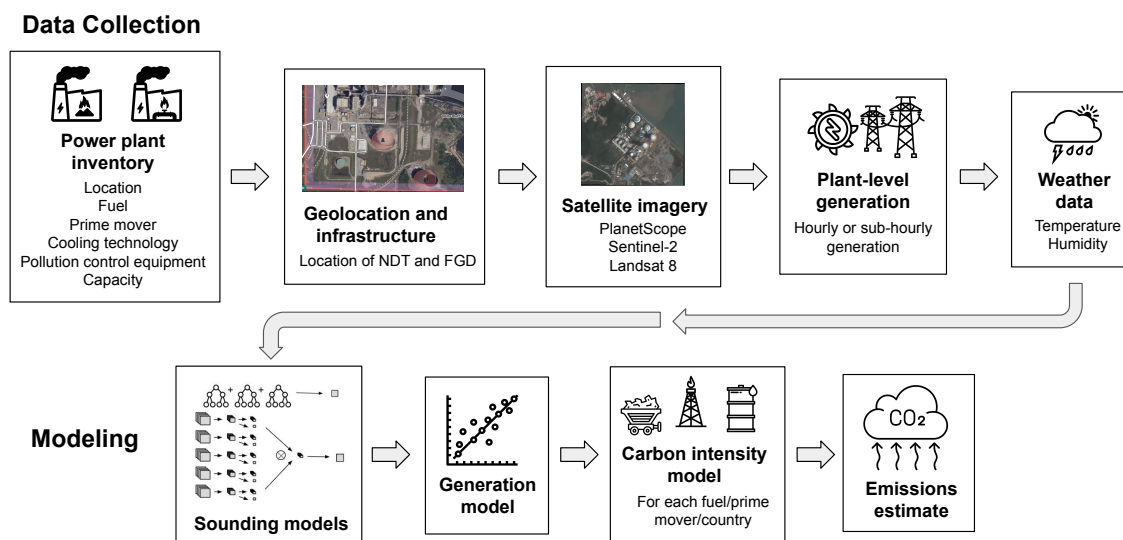


Figure 2. Overview of the data and models required to estimate CO₂ emissions from fossil-fuel power plants using satellite imagery.

3.1. Power Plant Datasets

In order to form a set of power plants for training our model and running inference, we partnered with Global Energy Monitor¹ to create a complete-as-possible, harmonized inventory of global combustion power plants that are currently operational. This was necessary because no single existing power plant dataset contains all the required information for this work. To use a plant in our ML modeling, we required the following inventory and auxiliary data:

- An accurate plant location for our satellite imagery.
- The location of FGD flue stacks and NDT cooling towers to focus our models on the relevant signals.
- Attributes of the power plant, including type, fuel, cooling technology, and air pollution control equipment to identify if the plant is suitable for our models.
- Local weather data to decide if temperature and humidity are conducive to vapor plume visibility.
- 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.
- Fuel and prime mover (i.e., steam or gas turbine) type to estimate the emissions factor.

In addition, plant-level electricity generation data is required for model training. Plants without generation data were used for inference only. We elaborate on each of these data requirements and our sources in Appendix A.

3.2. 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

¹ <https://globalenergymonitor.org/>

through the identification of emitted visible water vapor plumes. A sample image of a power plant from each satellite is shown in Figure 3. A description of each satellite and imagery processing steps is provided below.



Figure 3. 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).

PlanetScope. Planet Labs' PlanetScope satellite constellation consists of approximately 130 individual satellites, called "Doves," with the first launch of this constellation in 2014 [34]. The PlanetScope constellation provides daily revisits with an equatorial 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 [35,36]. PlanetScope PSScenes were downloaded via the Planet Labs API, providing a spatial resolution of ≈ 3 m, 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 [37,38]. Each Sentinel-2 satellite has a 10-day revisit time with a 5-day combined revisit and an equatorial crossing time of $\approx 10:30$ am [39]. 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 [38,39]. We downloaded Harmonized Sentinel-2A/B Level-1C Top of Atmosphere (TOA) products from Google Earth Engine (GEE). **Landsat 8.** The Landsat 8 mission is jointly managed by NASA and US Geological Survey [40,41]. 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 [42]. We downloaded the Landsat 8 Collection 2, Tier 1 TOA from GEE.

Each satellite has multiple bands with different spatial resolutions. Lower-resolution bands were upsampled to match the highest-resolution band for that satellite. Each band also has a different distribution of pixel values, which can cause instability during model training. For this reason, 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.

For some of our models, we created additional bands through post-processing:

- haze-optimized transformation (HOT), a linear combination of the blue and red bands:

$$\text{HOT} = \text{blue} - 0.5 * \text{red}$$

- whiteness [43], which consists of:

$$\text{whiteness} = \frac{|\text{blue} - \text{visible}| + |\text{green} - \text{visible}| + |\text{red} - \text{visible}|}{\text{visible}}$$

where $\text{visible} = \frac{1}{3}(\text{blue} + \text{green} + \text{red})$

- normalized difference vegetation index (NDVI), a ratio between the Red and Near-infrared (NIR) bands,

$$\text{NDVI} = \frac{\text{NIR} - \text{red}}{\text{NIR} + \text{red}}$$

- normalized difference shortwave infrared, a ratio between the shortwave infrared (SWIR) bands,

$$\text{normalized difference SWIR} = \frac{\text{SWIR2} - \text{SWIR1}}{\text{SWIR2} + \text{SWIR1}}$$

- normalized difference thermal infrared, a ratio between the shortwave infrared (TIR) bands,

$$\text{normalized difference TIR} = \frac{\text{TIR2} - \text{TIR1}}{\text{TIR2} + \text{TIR1}}$$

These were generated for the satellites that have these available bands, i.e., normalized difference SWIR for Landsat 8 and Sentinel-2 but not for Planet. The band combinations provided additional beneficial features for some of our models. For example, HOT and whiteness acted as a basic plume mask. Our gradient-boosted tree models used all these bands, while the neural networks were more limited because of transfer learning, as described in Section 3.5.

3.3. Ground Truth Labels for Model Training

To train our ML models, our satellite images needed 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 Appendix A.4. Our training plants are located in the US, Europe, and Australia. For regression models, each satellite image was labeled with the capacity factor at that timestamp: the generation of the plant divided by its capacity at the given timestamp (see Figure 1). 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."

3.4. Plant and Image Selection

For all satellite imagery, 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. Based on this ROI, we restricted the training data to optimize for both modelability and impact:

1. **Plant selection.** The first set of filters was based on the capacity of the plant and whether it was using mostly NDT or FGD technology at the time:
 - Coal must account for at least 50% of the plant's operating capacity.
 - For NDT models, $\geq 70\%$ of the plant's cooling was NDT.
 - For FGD models, $\geq 90\%$ of total generation and 100% of coal capacity was associated with wet FGD.
 - At least one NDT tower or FGD-enabled flue stack has been annotated in OpenStreetMap or in our in-house annotations database.
 - Capacity must be at least 500 MW.
 - 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, or insufficient or inconsistently reported generation data. This criterion is applicable to training only.

- Because our modeling approach assumes plants with wet FGD always run their pollution controls, we removed from training any operational plants that repeatedly exhibited sporadic or no visible FGD usage when the plant was reported "on." For inference, this is harder to assess in the absence of reported generation data. We flagged operational plants with FGD where FGD plumes were never detected under the expected plume-favorable weather conditions (detailed later in this section) by our classification models, nor observed upon manual inspection of 100 random images. We also flagged plants that exhibited other signals of operating (i.e., NDT plumes) with no FGD signal. The two primary reasons an operational plant may fail to exhibit an FGD vapor plume signal when generating electricity under the appropriate temperature and humidity conditions are:
 - Our power plant database has incorrect information suggesting the power plant has a "wet" process when it is actually "dry." This is possible for both NDT and FGD as either can be a "wet" or "dry" process; generally "dry" is more common in arid climates to conserve water.
 - The power plant fails to run its SO_x pollution control equipment (the flue gas desulfurization, FGD), so there is no FGD plume. Note that this is relevant only for FGD plumes, not NDT, because some type of cooling is necessary to operate a power plant, whereas pollution controls are not strictly necessary to operate (rather, they represent additional requirements set by clean air regulations).

For inference on NDT plants, these criteria were relaxed by applying our models to all plants with any positive amount of NDT capacity and a total capacity of at least 50 MW. Plants must meet these criteria in every year, 2015-2022, to be included.

2. **Image selection.** We also filtered based on the characteristics of each satellite image:

- Our FGD and NDT annotations are fully contained within the satellite image.
- 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.
- For PlanetScope, we used the post-2018 UDM2 bands to keep only images with <80% heavy haze.
- 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.
- Images with known data quality issues were discarded, e.g., exhibiting plumes when generation has been zero for at least an hour. Appendix A.4 details the scenarios in which we excluded images due to quality issues.
- When there were images of the same location with the same timestamp, we kept a single copy, breaking ties with the following: 1) largest area surrounding the plant contained, 2) least cloudy, 3) latest download timestamp, 4) random selection.

3. **Weather filters.** Images were excluded from FGD models when ambient weather conditions were unfavorable for plume visibility. At high temperatures 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:

- Exclude images in which the ambient temperature is $\geq 14^{\circ}\text{C}$ and relative humidity is $\leq 26\%$.
- Exclude images in which the ambient temperature is $\geq 24^{\circ}\text{C}$ and relative humidity is $\leq 36\%$.

- Exclude images in which the ambient temperature is $\geq 32^{\circ}\text{C}$.

After applying the above plant, image, and weather filters, our training dataset consisted of 74 plants for NDT and 99 for FGD for the years 2015-2022. Table 1 lists the number of training images for each satellite before and after filtering.

Table 1. Training image count for the years 2015-2022 for each satellite before and after filtering. For PlanetScope, we applied cloud filtering prior to downloading images, so the "Before filtering" column doesn't include cloudy images. Note that the same image may be used by both the NDT and FGD models since a power plant can have both types of technology.

| Satellite | NDT image count | | FGD image count | |
|-------------|------------------|-----------------|------------------|-----------------|
| | Before filtering | After filtering | Before filtering | After filtering |
| PlanetScope | 69,533 | 54,136 | 108,038 | 69,462 |
| Sentinel-2 | 18,964 | 15,064 | 30,541 | 18,390 |
| Landsat 8 | 9,040 | 7,176 | 12,885 | 8,235 |

3.5. Sounding Models

To estimate power plant generation and CO_2 emissions from satellite imagery at a specific timestamp, our sounding-level models predicted if a power plant was running or not (on/off) and estimated the capacity factor (generation divided by capacity). We included both sets of models because the on/off task was simpler and could be predicted more accurately, while the regression task was essential for differentiating low from high generation. We multiplied the predicted capacity factor by the capacity to infer generation.

As we estimate power plant activity by identifying vapor plumes, focusing our models on the structures that emit these plumes can improve performance. Therefore, we used the annotated NDT cooling tower and FGD stack patches as model inputs, as shown by the red and blue boxes in Figure 1. This helped produce more accurate classification models than a single image of the entire plant, which can have features of non-power plant activity impacting the accuracy [32]. Two different types of models were trained: gradient-boosted decision trees and convolutional neural networks (CNNs). For each model type, separate models for NDT and FGD were built, as well as for each satellite dataset, for a total of 16 sounding models. Figure 4 illustrates the structure of both model types.

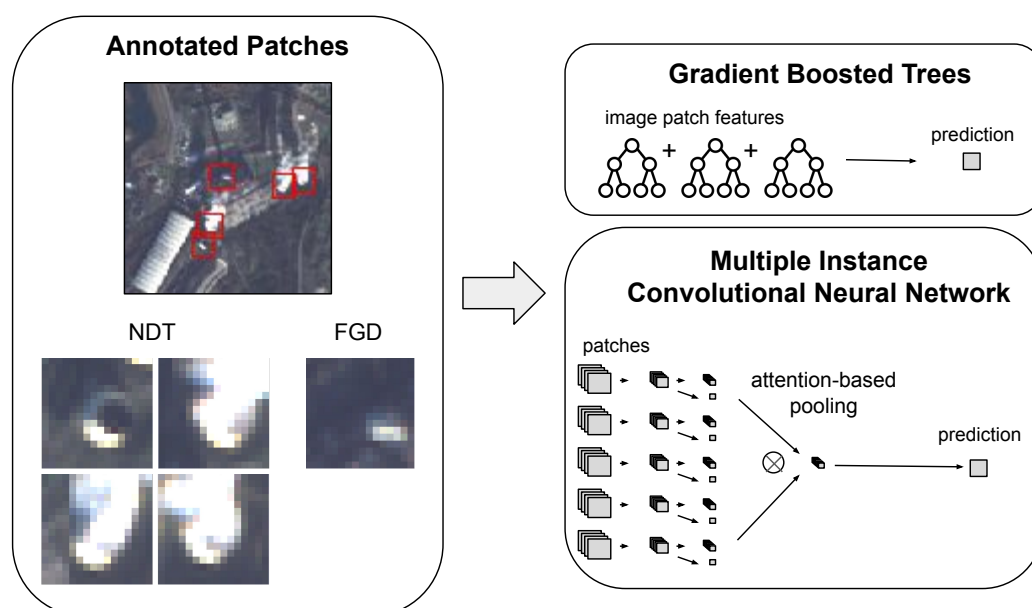


Figure 4. Overview of modeling approaches: NDT cooling towers and FGD stacks were cropped from satellite imagery and fed into ML models to classify the power plant's on/off status and predict the capacity factor. Image from [32].

3.5.1. Gradient-Boosted Decision Trees

The gradient-boosted decision tree models used XGBoost. For each image fed into the XGBoost models, we extracted patches of varying sizes centered on the FGD or NDT structures. A set of statistics was derived for the pixel values in each patch: mean, standard deviation, and 90th percentile. These vectors of statistics were then aggregated to the image level using mean, min, and max operations to accommodate plants with multiple FGD or NDT structures. We used multiple patch sizes around the FGD or NDT structures (4 to 32 pixels for Landsat 8, 4 to 32 pixels for Sentinel-2, and 8 to 64 pixels for PlanetScope) in order to capture features at different scales.

Visible plumes at power plants tend to be white. However, there can be other white features, such as buildings, near the cooling tower or flue stacks. To handle this, 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 as described above.

3.5.2. Convolutional Neural Networks

To enable the CNN models to handle multiple patches, we used a multiple instance framework to combine patches via an attention layer. The attention mechanism aggregates features from all the different patches for a particular image of a plant [32,33]. Each patch was first encoded with a CNN encoder truncated after a subset of the convolutional blocks. Transfer learning was used to initialize model weights in one of two ways:

- RESISC: with a ResNet50 CNN [44] pre-trained on the RESISC dataset, which consists of aerial RGB images labeled by land cover and scene classes [45]. 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.
- BigEarthNet: with a VGG16 CNN [46] pre-trained on the BigEarthNet dataset, which consists of Sentinel-2 images labeled by land cover class [47]. This model uses 10 bands from Sentinel-2, excluding the lowest resolution bands 1, 9, and 10. We were not able to apply this model to PlanetScope but did 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 enabled the model to learn a more diverse set of spectral characteristics, it contained only cloud-free images; our model must learn plume features during fine-tuning.

After the shared convolutional encoder, we used an attention layer to combine patch features as a weighted sum, with the weights determined by the model itself [48]. A dense layer then made the final prediction. More details about model configuration and training are provided in Appendix B.

3.6. Generation Ensemble Models

The sounding-level models described in Section 3.5 give us an instantaneous estimate of the power plant activity in a single satellite image. Collecting these instantaneous estimates creates an irregular time series of on/off classification and capacity factor regression estimates for each plant. 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 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. Afterward, the estimated generation was multiplied by an emissions factor, as described later in Section 3.7, to estimate emissions.

Separate generation models were trained for NDT and FGD plants and only for the years 2018 to 2022 inclusive. This is because there is very limited PlanetScope data available prior to 2018, making rolling 30-day predictions difficult due to the sparsity of soundings. Each generation model is an L1-regularized linear regression model with features formed from aggregating the sounding model predictions (described in Appendix C). To calculate features for each plant at each point in time, we aggregated the predictions from the sounding-level models over multiple lookback windows. In total, the NDT model had access to 162 features, while the FGD model had access to 324 features (see Appendix C for the list of 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 over which we aggregated predictions 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 5 shows an example of FGD and NDT model predictions compared with the capacity factor calculated from reported electricity generation.

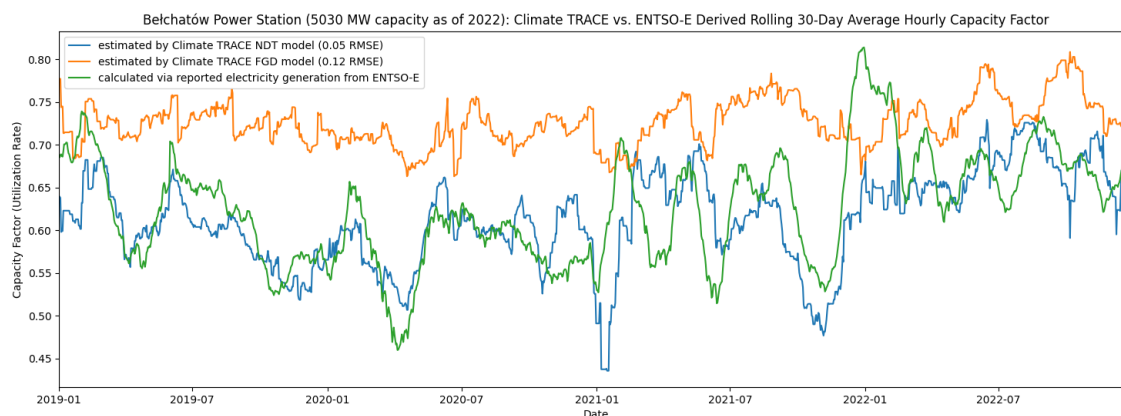


Figure 5. A sample of model predictions for Bełchatów in Poland, Europe’s largest power plant, for the years 2019 to 2022. Included are predictions from the NDT (blue line) and FGD (orange line) generation models to compare to the reported electricity generation (green line). The rolling average 30-day capacity factor never fell below 45% capacity from 2019-2022 and was maintained above 60% for the entirety of 2022. The NDT model performs over two times better than the FGD model for this power plant, a representative characteristic of the NDT task given its clearer signal.

3.7. Emissions Estimation

To generate plant-level emissions estimates, we first estimated the annual capacity factor of each plant. Then, the rolling 30-day average capacity factor estimates were converted to annual capacity factor by averaging and weighting each estimate by what fraction of the 30-day interval fell into the year in question. 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 (see Appendix A.6). Finally, the unit-level estimates were aggregated to the plant level to provide the annual facility-level emissions estimate.

3.8. Model Training with Cross-Validation

As our goal was to estimate power plant emissions when no reported generation data exists, it is essential to validate that our models generalize well. 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 $1^\circ \times 1^\circ$ 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. We independently tuned the hyperparameters of both XGBoost and CNN models 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; plants were “on” in about $\approx 80\%$ of the training images.

Since the generation models are trained on the outputs of other ML models – the sounding models – we must take extra care to avoid data leakage. Therefore, a cross-validation approach was adopted for the training and evaluation of the generation models. Each of the four generation model instances was trained with features based on soundings from the validation fold and evaluated with features based on the sounding predictions 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 overfitted 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.

3.9. 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 3.4.

Predictions for all images were generated over the period 2015-2022, using our sounding-level models described in Section 3.5. For inference, the sounding predictions were averaged from all four instances of a sounding model (one model for each fold). Then these predictions were converted to generate time series features and rolling 30-day average capacity factor predictions using the generation models as described in Section 3.6. Similar to the sounding models, the predictions were averaged from all four 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 capacity factor of that plant over the preceding 30 days. These were then summed and weighted by how much the 30-day interval overlaps with the year to produce estimates of the annual plant-level capacity factor.

The sounding and generation models were run 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 was straightforward: simply use the prediction from the single applicable model. For plants with both NDT and FGD, predictions were aggregated from both model types by weighting the NDT model prediction two times more than the FGD model prediction ($\frac{2*NDT+FGD}{3}$), reflecting the lower error and increased confidence in the NDT models.

4. Results

We used two mechanisms to validate our models. First, through cross-validation on our training set, performance was measured on the held-out fold and aggregated across the four folds (details in Section 3.8). Second, additional generation and emissions data was gathered from plants with NDT and/or FGD in Türkiye and India (Appendices A.4 and A.5), plus plants with NDT in the US, Europe, and Australia that did not meet our strict filters for training but did qualify with a looser filter due to a lower capacity or a more heterogeneous set of cooling types (Section 3.4). We refer to the first mechanism on training plants as "cross-validation" and the second on non-training plants as "external validation." Table 2 below summarizes the fraction of total power sector emissions and the number of observations, plants, and countries to which we applied our ML methods, as compared to prior work discussed in Section 2. We validated our methodology on nearly twice as many plants, two to four orders of magnitude more observations, and a broader selection of countries than any prior study. Furthermore, our inference predictions encompass 1,042 power plants across 41 countries, including several plants in each of the five countries that produce the most carbon emissions from power generation (in order): China, US, India, Russia, and Japan [49]. Figure 6 displays the coverage of our emissions estimates for 2022 on a map.

Table 2. Coverage summary of Climate TRACE’s ML approach, as compared with prior work discussed in Section 2; all studies that included ten or more plants are listed. Percent power emissions are calculated per year, then averaged to produce the value in the table, using as denominator the total power sector emissions as estimated by Climate TRACE for each year (note that the ML technique presented in this paper represents just one of the ways that Climate TRACE estimates power sector emissions [50]).

| Approach | % Power Emissions | # Observations | # Plants | # Countries | Years |
|---------------------------------|-------------------|----------------|----------|-------------|-----------|
| Climate TRACE ML | | | | | |
| Cross-val | 7% | 157,831 | 139 | 12 | 2015-2022 |
| External val (generation) | 2% | 107,440 | 101 | 16 | 2018-2022 |
| External val (CO ₂) | 7% | 162,588 | 207 | 17 | 2019-2022 |
| All inference | 32% | 1,198,167 | 1042 | 41 | 2015-2022 |
| Jain [31] | <1% | 2,131 | 146 | 18 | 2019-2021 |
| Hanna et al. [30] | <1% | 1,639 | 146 | 11 | 2020 |
| Lin et al. [19] | <1% | 50 | 22 | 1 | 2018-2021 |
| Cusworth et al. [11] | <1% | 28 | 21 | 4 | 2014-2020 |
| Nassar et al. [16] | <1% | 20 | 14 | 6 | 2014-2018 |

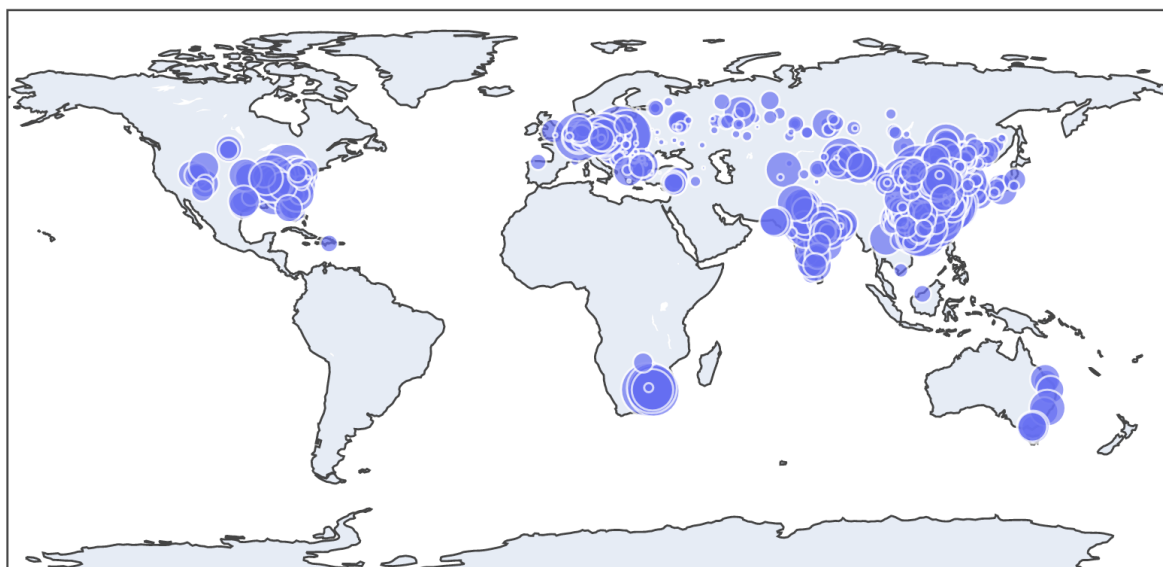


Figure 6. A map displaying global CO₂ emissions estimates produced by our ML models (equivalent to "All inference" in Table 2). Each dot represents a single power plant, with the size of the dot corresponding to the amount of estimated CO₂ emissions for that plant in 2022.

Using these two mechanisms, the performance was evaluated for our sounding, generation, and emissions models. The results are detailed in this section. We also calculated 90% 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).

4.1. Sounding Model Validation

Classification and regression results for 2015-2022 are displayed in Tables 3 and 4, respectively. A naive baseline mAP is 0.5 for a model that predicts the most prevalent class, or always "on." For the regression model, a naive baseline that predicts the training set’s mean capacity factor would produce an RMSE of 0.34 and a mean bias error (MBE) of 0 across all imagery sources for both NDT and FGD. Models were trained and cross-validated on 99 plants with FGD (25 in Europe and 74 in the US) and 74 plants with NDT (45 in Europe, 22 in the US, and 7 in Australia). All of our models outperformed

their respective baselines – and 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.

Table 3. On/off classification cross-validated mean average precision (mAP) and block-bootstrapped 90% confidence intervals for 2015-2022.

| Satellite | Model | NDT mAP [CI] | FGD mAP [CI] |
|-------------|--------------|----------------------|----------------------|
| PlanetScope | XGBoost | 0.956 [0.932, 0.987] | 0.886 [0.857, 0.911] |
| PlanetScope | CNN - RESISC | 0.930 [0.901, 0.963] | 0.885 [0.861, 0.916] |
| Sentinel-2 | XGBoost | 0.932 [0.904, 0.980] | 0.889 [0.858, 0.917] |
| Sentinel-2 | CNN - RESISC | 0.959 [0.941, 0.974] | 0.903 [0.880, 0.925] |
| Sentinel-2 | CNN - BEN | 0.933 [0.906, 0.966] | 0.839 [0.801, 0.873] |
| Landsat 8 | XGBoost | 0.899 [0.866, 0.942] | 0.878 [0.855, 0.904] |
| Landsat 8 | CNN - RESISC | 0.901 [0.866, 0.942] | 0.824 [0.793, 0.854] |
| Landsat 8 | CNN - BEN | 0.865 [0.826, 0.899] | 0.811 [0.784, 0.837] |

Table 4. Capacity factor regression cross-validated root mean squared error (RMSE) and mean bias error (MBE), alongside respective block-bootstrapped 90% confidence intervals, for 2015-2022.

| Satellite | Model | NDT | | FGD | |
|-------------|--------------|----------------------|-------------------------|----------------------|-------------------------|
| | | RMSE [CI] | MBE [CI] | RMSE [CI] | MBE [CI] |
| PlanetScope | XGBoost | 0.209 [0.194, 0.221] | -0.021 [-0.034, -0.006] | 0.297 [0.285, 0.308] | -0.056 [-0.076, -0.037] |
| PlanetScope | CNN - RESISC | 0.196 [0.186, 0.206] | -0.024 [-0.037, -0.010] | 0.263 [0.244, 0.277] | -0.065 [-0.084, -0.045] |
| Sentinel-2 | XGBoost | 0.210 [0.198, 0.220] | -0.020 [-0.034, -0.006] | 0.270 [0.258, 0.281] | -0.029 [-0.049, -0.011] |
| Sentinel-2 | CNN - RESISC | 0.203 [0.192, 0.213] | -0.019 [-0.034, -0.004] | 0.267 [0.248, 0.282] | -0.048 [-0.066, -0.029] |
| Sentinel-2 | CNN - BEN | 0.220 [0.206, 0.233] | -0.025 [-0.041, -0.008] | 0.259 [0.246, 0.269] | 0.005 [-0.014, 0.025] |
| Landsat 8 | XGBoost | 0.243 [0.220, 0.260] | -0.011 [-0.032, 0.013] | 0.285 [0.271, 0.298] | -0.063 [-0.085, -0.042] |
| Landsat 8 | CNN - RESISC | 0.266 [0.249, 0.280] | -0.089 [-0.110, -0.068] | 0.299 [0.289, 0.309] | -0.036 [-0.054, -0.015] |
| Landsat 8 | CNN - BEN | 0.264 [0.247, 0.278] | -0.031 [-0.054, -0.006] | 0.301 [0.288, 0.313] | 0.025 [0.002, 0.050] |

Comparing the different model types drew out a few more insights. The BEN models require 10 bands and so can only be used for Sentinel-2 and Landsat 8. In general, RESISC models performed slightly better than BEN. Overall, RESISC and XGBoost fared similarly, with RESISC sometimes outperforming XGBoost and sometimes vice versa, emphasizing the utility of an ensemble approach. Both PlanetScope and Sentinel-2 models performed similarly; the on/off classification mAP for both varied from 0.93 to 0.96 for NDT and from 0.84 to 0.90 for FGD across all models, and the capacity factor regression RMSE for both ranged from 0.20 to 0.22 for NDT and 0.26 to 0.30 for FGD. The Landsat 8 models generally were less accurate, with on/off classification mAP values of 0.81 to 0.90, and capacity factor regression RMSE values between 0.24 to 0.30. The Landsat 8 models' lower performance is most likely due to the coarser 30m spatial resolution and revisit rate of 16 days, which results in fewer training images. Additionally, the capacity factor regression RMSE values for Sentinel-2 and PlanetScope overlap, which suggests that Sentinel-2's 10m resolution is sufficient to identify on/off classification and the capacity factor. Even though PlanetScope's 3m resolution did not substantially improve the model's performance, the daily revisits improve the temporal resolution, allowing for more frequent capturing of the on/off and capacity factor identification at power plants, which produces a greater number of total observations in each year.

Appendix D includes some additional results validating our sounding models on non-training plants in the US, Europe, Australia, and Türkiye. On these plants, our NDT classification models using Sentinel-2 and PlanetScope imagery performed similarly with mAP between 0.74 and 0.78, while the Landsat 8 models received lower average classification mAP scores of ≈ 0.72 . These NDT classification results are worse than the cross-validation performance but are still significantly ($\alpha = 0.1$) better than a naïve baseline of 0.5 mAP. The NDT regression models' overall average performance was 0.22, 0.23, and 0.27 for PlanetScope, Sentinel-2, and Landsat 8 models, respectively, which is comparable to

the cross-validation performance above. For the FGD model validation, only three plants with FGD had reported data available to compare against, but a similar trend was seen as the Sentinel-2 and PlanetScope models performed better than the Landsat 8 models.

4.2. Generation Model Validation

To validate our generation models, we measured performance using the RMSE between the reported and predicted 30-day rolling average capacity factors in the validation set. Table 5 summarizes the results of our two generation models for both cross-validation (training plants in the US, Europe, and Australia) and external validation (non-training plants in the US, Europe, Australia, Türkiye, and India). While we included imagery from as early as 2015 for the sounding model, we focused on 2018-2022 for the generation model. This is because PlanetScope has limited imagery before 2018, and this higher-temporal-resolution is crucial for predicting the 30-day rolling capacity factor. The baseline used is again the average value of the target in the relevant training set: 0.44 for NDT plants and 0.47 for FGD plants. Compared to the sounding-level regression tasks described previously, the NDT ensemble model lead to a lower baseline RMSE of 0.27. This is due to the smoothing effect from 30-day averaging. Our generation ensemble models combine multiple sounding model types, satellite imagery sources, and timescales to outperform this baseline significantly ($\alpha = 0.1$) with an RMSE of 0.149 to 0.199 in almost all cases for both cross-validation and external validation. The one exception is FGD external validation, which was based on only four plants (due to limited reported data) that produced an RMSE of 0.323. These four plants include the same three in Türkiye used for sounding validation and one additional plant in India.

Table 5. Generation ensemble model performance for each plant type, 2018-2022, for training plants using cross-validation and on additional plants as external validation. Baseline results and a 90% confidence interval are also included.

| Plant Type | Validation Type | Plant Count | RMSE | | MBE | |
|------------|-----------------|-------------|----------------------|----------|-------------------------|----------|
| | | | ML [CI] | Baseline | ML [CI] | Baseline |
| NDT | cross | 73 | 0.149 [0.129, 0.162] | 0.272 | -0.014 [-0.028, 0.002] | 0.000 |
| NDT | external | 104 | 0.199 [0.188, 0.210] | 0.273 | -0.075 [-0.091, -0.058] | 0.048 |
| FGD | cross | 97 | 0.196 [0.187, 0.205] | 0.270 | 0.012 [-0.007, 0.031] | 0.000 |
| FGD | external | 4 | 0.323 [0.216, 0.384] | 0.359 | -0.252 [-0.330, -0.127] | -0.272 |

Limited conclusions can be drawn from the external validation on FGD plants, as there are only four plants with reported data in this group. Sounding-level FGD results for Türkiye showed both a high RMSE and a large negative MBE (Table A3). These errors are once again reflected in the results for our generation model (Table 5, bottom row).

Our generation models performed better on NDT than FGD plants, which 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 filters described in Section 3.4 that 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.

4.3. Annual Validation

Facility-level annual average capacity factor, total electricity generation, and total CO₂ emissions were calculated from the sounding and generation model outputs for 2019 to 2022, inclusive. This year range was selected because generation model training was restricted to 2018 to 2022 (due to scarcity of PlanetScope data prior to 2018) and because of the overlapping nature of the 30-day windows and fiscal year differences necessitating lookbacks to the previous year.

In addition to our ML models, we produced a second set of simpler baseline capacity factor estimates. We used country-level annual estimates of capacity and generation by fuel type from the EIA and EMBER Yearly Electricity Data². From this, we calculated 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. This does not account for the typical variation in dispatch for plants serving base, intermediate, and peaking load. This is the baseline used for comparison in this section. Metrics were calculated by comparing estimates derived from our ML methods to reported generation data summarized in Appendix A.4 or emissions data summarized in Appendix A.6.

Table 6 displays performance metrics for the annual average capacity factor. For the US and Europe, cross-validation performance is comparable with an RMSE of 0.17 and 0.12, respectively, and external validation is 0.14 for both. The region with the highest (lowest performing) external validation RMSE was Türkiye with 0.24. The region with the largest MBE was Australia with -0.12. This may be due to Australian power plants generally running at higher capacity factors than the rest of the training set. The fifty plants in India, on the other hand, achieved results very similar to the US training plants, lending credibility to the generalizability of our ML emissions estimation approach.

Table 6. Annual average capacity factor estimation performance for 2019-2022 compared to capacity factor derived from reported generation data for the US, Europe, Australia, India, and Türkiye, as reported by their respective government mechanisms summarized in Appendix A.4. A comparison to the country- and fuel-specific average baseline is also included.

| Region | Validation Type | Plant Count | RMSE | | MBE | |
|-----------|-----------------|-------------|-------------------|----------|----------------------|----------|
| | | | ML [CI] | Baseline | ML [CI] | Baseline |
| US | cross | 78 | 0.17 [0.16, 0.18] | 0.22 | -0.02 [-0.05, 0.00] | -0.05 |
| US | external | 6 | 0.14 [0.12, 0.17] | 0.19 | 0.03 [-0.06, 0.12] | 0.04 |
| Europe | cross | 59 | 0.12 [0.11, 0.13] | 0.20 | 0.04 [0.02, 0.06] | 0.05 |
| Europe | external | 27 | 0.14 [0.12, 0.16] | 0.18 | 0.02 [-0.01, 0.05] | 0.10 |
| Australia | cross | 7 | 0.20 [0.08, 0.26] | 0.18 | -0.12 [-0.20, -0.04] | -0.10 |
| Australia | external | 1 | 0.09 [N/A] | 0.09 | -0.01 [N/A] | 0.04 |
| India | external | 50 | 0.14 [0.13, 0.15] | 0.20 | -0.03 [-0.05, -0.01] | 0.08 |
| Türkiye | external | 8 | 0.24 [0.21, 0.26] | 0.32 | 0.00 [-0.12, 0.10] | 0.13 |

Comparing modeled to reported annual generation and emissions for individual power plants, the overall RMSEs, aggregating across both cross- and external validation, were 1.75 TWh and 2.04 Mt CO₂, respectively. Tables 7 and 8 show the performance of our ML methods for annual total generation (in terawatt-hour or TWh) and annual CO₂ emissions (in megatonnes or Mt) estimation. For US cross- and external validation, Europe cross-validation only, and India (all external validation), the ML-based estimate had a significantly lower RMSE than the baseline approach using country averages. However, the RMSE for Europe external validation and all Australian plants was no better than the baseline approaches. Furthermore, for Australia, the errors were significantly worse than anywhere else with the ML-based method, underpredicting substantially. This may once again be because the ML models failed to capture the higher average utilization rate in Australia, which makes up a small percentage of training plants, relative to the lower average utilization rate of the rest of the training set. That being said, the overall plant-wise annual RMSEs of 1.75 TWh and 2.04 Mt CO₂ are small relative to the annual generation and CO₂ emissions of the world's largest power plants, as can be seen in Figure 7.

² <https://ember-climate.org/data-catalogue/yearly-electricity-data/>

Table 7. Annual total generation estimation performance compared to reported generation data (in terawatt-hours, TWh) for the US, Europe, Australia, India, and Türkiye for years 2019 to 2022.

| Region | Validation Type | Plant Count | RMSE (TWh) | | MBE (TWh) | |
|-----------|-----------------|-------------|-------------------|----------|----------------------|----------|
| | | | ML [CI] | Baseline | ML [CI] | Baseline |
| US | cross | 78 | 1.68 [1.51, 1.84] | 2.99 | 0.06 [-0.16, 2.90] | -0.65 |
| US | external | 6 | 2.92 [0.68, 3.92] | 3.90 | -0.66 [-2.17, 0.88] | 0.31 |
| Europe | cross | 59 | 1.38 [1.09, 1.58] | 2.43 | 0.35 [0.15, 0.53] | -0.16 |
| Europe | external | 27 | 1.39 [1.09, 1.61] | 1.27 | -0.55 [-0.82, -0.28] | 0.65 |
| Australia | cross | 7 | 2.05 [0.91, 2.60] | 2.19 | -1.46 [-2.08, -0.63] | -1.33 |
| Australia | external | 1 | 1.12 [N/A] | 0.94 | -0.49 [N/A] | 0.37 |
| India | external | 50 | 2.12 [1.78, 2.37] | 2.58 | -1.38 [-1.63, -1.12] | 0.98 |
| Türkiye | external | 8 | 2.08 [1.31, 2.57] | 2.80 | -0.94 [-1.91, 0.01] | 0.97 |

Table 8. Annual emissions estimation performance compared to reported data for the US, EU, Australia, and India for years 2019 to 2022

| Region | Validation Type | Plant Count | RMSE (Mt CO ₂) | | MBE (Mt CO ₂) | |
|-----------|-----------------|-------------|----------------------------|----------|---------------------------|----------|
| | | | ML [CI] | Baseline | ML [CI] | Baseline |
| US | cross | 77 | 2.11 [1.73, 2.36] | 2.68 | 0.37 [0.08, 0.67] | -1.13 |
| US | external | 5 | 1.12 [0.37, 1.43] | 2.25 | -0.61 [-1.08, -0.10] | -0.56 |
| EU | cross | 58 | 2.27 [1.70, 2.68] | 3.04 | 0.57 [0.19, 0.97] | -0.18 |
| EU | external | 36 | 1.00 [0.76, 1.16] | 1.05 | -0.41 [-0.58, -0.23] | 0.25 |
| Australia | cross | 6 | 4.37 [2.42, 5.67] | 3.70 | -2.69 [-4.49, -0.73] | -2.01 |
| Australia | external | 1 | 3.49 [N/A] | 2.62 | -3.39 [N/A] | -2.41 |
| India | external | 24 | 1.72 [1.49, 1.92] | 2.65 | -0.81 [-1.15, -0.49] | 1.63 |

Figure 7 plots our estimates for average annual capacity factor, total annual generation, and annual CO₂ emissions at individual plants against reported electricity and emissions data. Although our models have an RMSE of 0.12 to 0.24 for capacity factor, they are reliable in differentiating low versus high utilization. Validating against reported data, instead of remote sensing-derived GHG estimates discussed in Section 2, provides robust and more plentiful verification of specific plants. One area for improvement is that the model struggles to predict high-capacity factors, favoring mid-range predictions instead, which contributes to a negative bias in this range. However, once total generation and emissions are calculated from our capacity factor estimates, and compared to reported values, the bias of our models is much more evenly distributed across plants with different total generation and emissions. Looking at our external validation results, our models underestimate reported annual emissions across the board, and most substantially in Australia (n=7) where plants generally run at a higher capacity factor relative to the training regions according to reported generation data. On the other hand, the negative bias of emissions estimates in India (n=24) is more in line with the degree of underestimation seen for the US and Europe external validation, lending credibility to the generalizability of our ML emissions estimation approach beyond training regions. Furthermore, the error and bias are small enough relative to the total emissions of major power plants that, although it may not measure emissions perfectly, the ML technique presented here is immediately useful in helping the world track relative comparisons between power plants and thus quantify marginal progress toward urgent emissions reduction goals.

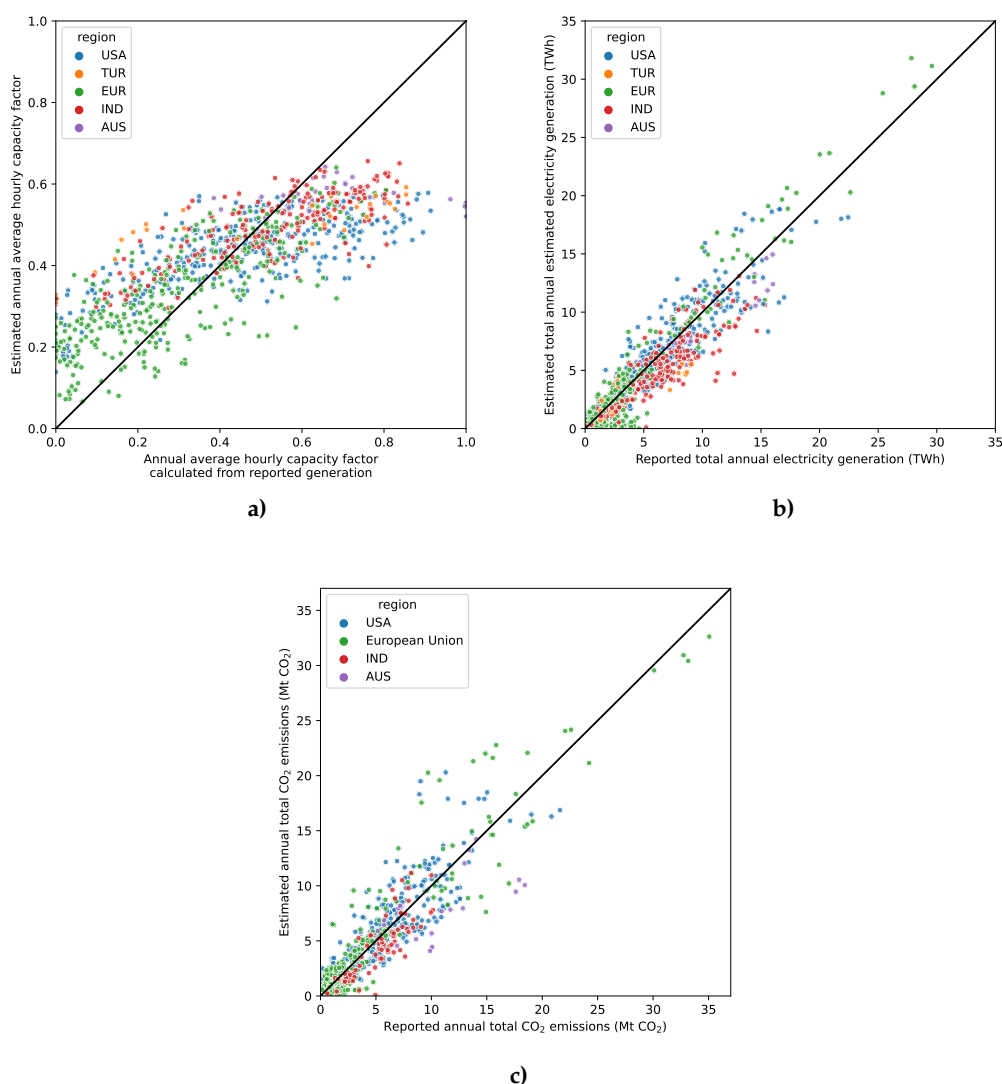


Figure 7. A comparison of ML estimated vs. reported a) capacity factor, b) electricity generation (terawatt hour, TWh), and c) CO₂ emissions (megatonnes, Mt) for 2019-2022 [49]. Each dot represents an individual plant and year matched to a reported electricity generation or emissions source from the respective region's reporting agency. Note: reported CO₂ emissions were not plotted for Türkiye as no emissions data is available (Appendix A.5).

5. Discussion

Our results in Section 4 reveal where our models perform well, as well as a few regions and plant types with larger errors and, in some cases, a negative bias in estimates. We have identified the following sources of error that may have contributed:

1. **FGD Usage.** 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. However, there could be plants we missed and therefore underestimated their emissions.
2. **Changing Plant Characteristics.** 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 FGD or NDT criteria one year and not

another, yet we treat all years the same. To address this issue for the time being, we filtered out plants that did not meet the filtering criteria for one 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. **Satellite Overpass Time.** The satellites' local overpass times are during the daytime and in the morning, averaging around mid-morning or ≈ 11 am local time for all three sensors: Landsat 8, Sentinel-2, and PlanetScope. Therefore, all soundings, and thus all predictions, are based on images from this 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 restricted to training on the US, Australia, and Europe, it may simply learn how mid-morning power plant snapshots predict rolling 30-day average generation in those regions (especially the US and Europe, as 93% of training plants are in these regions). Our method can, therefore, over- or under-predict 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 ML models' negative bias in India, Türkiye, and Australia.
4. **Weather Influencing Signal Visibility.** The primary proxy signal we currently use to estimate capacity factor is the water vapor plume, whose size is sensitive to temperature and humidity: cold and wet conditions favor large plumes, and dry and hot conditions result in smaller or fainter plumes. We currently mitigate this with a temperature and humidity filter 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. 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 that are not as sensitive to local weather will reduce this bias, and this is an area we are actively working on.
5. **More Satellite Images in Recent Years.** The majority of PlanetScope satellites were launched in 2017 or later, with more satellites added through 2022. Further, Sentinel-2B was launched in 2017. This makes satellite-derived estimates in the years 2015 to 2017 less accurate due to the limited satellite coverage and observations. Because there is less confidence on ML predictions prior to 2019, we restricted asset-level reporting on the Climate TRACE website to 2019-2022, while ML predictions on 2015-2018 are used only for aggregating into country totals.

6. Conclusions

Applying ML models to multi-spectral satellite imagery enables the identification of power plant generation activity and emissions with more comprehensiveness, consistency, recency, and greater detail than current reported data that tends to vary in quality and coverage. Additionally, combining imagery from multiple satellites provides more observations, and thereby more activity estimates, over current GHG-concentration remote sensing observations. The approach developed here can be applied to countries where generation or emissions data is not publicly available or is not possible due to technical limitations. Our ML and satellite monitoring approach creates the ability to provide publicly available power plant emissions estimates on a frequent and plant-wise basis. Currently, our estimated power plant CO₂ emissions are available on the Climate TRACE website (climatetrace.org). As of fall 2023, the website contains country-level annual CO₂ emissions estimates for 2015-2022 and power plant, source-level, annual CO₂ emissions estimates for 2019-2022. The source-level estimates include power plants for the US, Europe, Australia, Türkiye, and India discussed in Section 4.3. Additionally,

the models were run on FGD and NDT plants in other regions of the world meeting the criteria described in Section 3.4. In total, we used our ML method to estimate the emissions of 1,042 power plants, representing 3% of power plants globally but responsible for roughly 32% of global combustion power plant CO₂ emissions averaged over 2015-2022, representing a major step forward in the world's ability to monitor power plant emissions.

Our models' predictions were found to be more accurate than the alternative baseline calculations using country- and fuel-specific capacity factor averages. We identified a negative bias in Australia and Türkiye and hypothesized some sources for this error (e.g., plants running at higher average capacity factors than the bulk of the training data, see Appendix D) that we will work to mitigate in the future. 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,

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

The use of satellite imagery that is available at low latency allows for estimates to be derived at a higher recency than other GHG inventory methods and can track and identify the emissions down to the source level. This approach is a promising step forward in providing more up-to-date estimates and can complement current approaches to estimate emissions. This work can provide useful information to governments, corporations, and citizens that seek to reduce their GHG emissions to meet The Paris Agreement and sustainability goals. As this project is an ongoing effort, the Climate TRACE website (climatetrace.org) will continue to be updated with the best-known available methods to provide global coverage for power plant emissions estimation, and contributions from the community are welcome and encouraged.

Author Contributions: Conceptualization, H.D.C., M.A., J.F., H.K., A.R.K., J.O., I.S.-R., A.F., J.J., J.L., and C.M.; methodology, H.D.C., M.A., J.F., H.K., A.R.K., J.O., and A.F.; software, H.D.C., M.A., J.F., H.K., A.R.K., J.O., I.S.-R., A.F., J.J., and J.L.; validation, M.A., J.F., A.D., H.K., C.L., and G.V.; data curation, M.A., H.K., A.R.K., J.O., I.S.-R., A.F., J.J., J.L., T.N., C.D., C.L., and G.V.; writing—original draft preparation, H.D.C. and M.A.; writing—review and editing, all; visualization, H.D.C. and M.A.; supervision, J.F., J.O., C.M., M.G., and G.M.; funding acquisition, M.G. and G.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by Al Gore, Generation Investment Management, Google.org, Beneficus Foundation, and Patrick J. McGovern Foundation.

Data Availability Statement: Links to our numerous data sources are provided in Appendix A. Emissions estimates resulting from our work are published at climatetrace.org.

Acknowledgments: The authors would like to thank Nick Amuchastegui, Grace Mitchell, Thomas Kassel, Keto Zhang, Krishna Karra, Julia Wang, Lee Gans, and the Climate TRACE coalition.

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

| | |
|-----------------|--|
| AVIRIS-NG | Next-Generation Airborne Visible/Infrared Imaging Spectrometer |
| BEN | BigEarthNet |
| CNN | convolutional neural network |
| CO ₂ | carbon dioxide |
| CO2M | Copernicus CO2 Monitoring |
| CEMS | continuous emissions monitoring systems |
| EIA | Energy Information Administration |
| EU | European Union |
| FGD | flue gas desulfurization |
| GAO | Global Airborne Observatory |
| GEE | Google Earth Engine |
| GOSAT | Greenhouse Gases Observing Satellite |
| HOT | haze-optimized transformation |
| mAP | mean average precision |
| MBE | mean bias error |
| ML | machine learning |
| NDT | natural draft wet cooling towers |
| NDVI | normalized difference vegetation index |
| NIR | near-infrared |
| OCO-2 | Orbiting Carbon Observatory |
| OLI | Operational Land Imager |
| PRISMA | PRecursore IperSpettrale della Missione Applicativa |
| RMSE | root mean squared error |
| ROI | region of interest |
| SWIR | shortwave infrared |
| TIR | thermal infrared |
| TIRS | Thermal Infrared Sensor |
| TOA | Top of Atmosphere |
| UDM2 | Usable Data Mask |
| US | United States |

Appendix A. Data Sources

Appendix A.1. Global Fossil Power Plant Inventory

We needed a power plant inventory containing information on location, fuel type, capacity, operating dates, cooling type, and pollution control technology at both the plant and unit levels. 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 A1 below describes how we use each of the datasets, including which data we republish.

Table A1. Datasets employed to create a harmonized Global Fossil Power Plant Inventory.

| Dataset | Plant Metadata Used |
|--|---|
| US Energy Information Administration EIA-860, EIA-860m https://www.eia.gov/electricity/data/eia860/ | Plant Name Unit Fuel Type Location Unit Capacity Unit Operating Dates Unit Cooling type Unit Pollution Control Tech SO2 |
| World Resources Institute (WRI) Global Power Plant Database (GPPD) https://datasets.wri.org/dataset/globalpowerplantdatabase | Plant Name Plant Fuel Type Location Plant Capacity Plant Operating Dates |
| S&P Global/Platts World Electric Power Plant (WEPP) https://www.spglobal.com/marketintelligence/en/solutions/market-intelligence-platform Note: The source level dataset is proprietary and is used internally only. | Unit Fuel Type Unit Capacity Unit Operating Dates Unit Cooling type Unit Pollution Control Tech SO2 |
| Global Energy Monitor (GEM) Global Coal Plant Tracker (GCPT) and Global Gas Plant Tracker (GGPT) https://globalenergymonitor.org/ | Plant Name Unit Fuel Type Location Unit Capacity Unit Operating Dates |
| Other Sources (e.g., press releases, newspaper articles, company websites) | All |

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 as much as possible 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.

Appendix A.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 (Figure A1). We used tags to label parts of the plant from which we expect to see vapor plumes: NDT cooling towers and FGD flue stacks. These annotations were used to focus our ML models on the most pertinent parts of the plant, improving their performance.



Figure A1. 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).

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 page³.

Appendix A.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 3.4. We obtained historical weather data from 2015-2022 for all of our plants from World Weather Online⁴.

³ https://wiki.openstreetmap.org/wiki/Organised_Editing/Activities/Climate_TRACE

⁴ <https://www.worldweatheronline.com>

Appendix A.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)⁵, European Network of Transmission System Operators for Electricity (ENTSO-E)⁶, and Australia National Electricity Market (NEM)⁷. 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 asset'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 that 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 inconsistently 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 ML 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).

For model validation, we also gathered reported electricity generation data from three additional countries for as many NDT or FGD power plants as possible:

- Türkiye – Enerji Piyasaları İşletme A.Ş. (EPIAŞ), also known as Energy Exchange Istanbul (EXIST)⁸. Reports hourly electricity generation data.
- India – National Power Portal (NPP)⁹. 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

⁵ <https://campd.epa.gov/>

⁶ <https://transparency.entsoe.eu/generation/r2/actualGenerationPerGenerationUnit/show>

⁷ <https://aemo.com.au/energy-systems/electricity/national-electricity-market-nem/about-the-national-electricity-market-nem>

⁸ <https://www.epias.com.tr/>

⁹ <https://npp.gov.in/>

database for multiple years). Although we did not use them for model training, all three datasets were included in our validation described in Sections 4.1 to 4.3, which we used to estimate model error.

Appendix A.5. Plant-Level Emissions Data

We sourced facility-level power plant CO₂ emissions data reported by the following agencies to compare against estimates derived from our machine learning approach:

- United States – US EPA Clean Air Markets Program Data (CAMPD)¹⁰
- European Union – European Union Emissions Trading System (EU ETS)¹¹
- Australia – Clean Energy Regulator (CER) National Greenhouse and Energy Reporting (NGER) Scheme¹²
- India – Central Electricity Authority (CEA)¹³

Note that, for Australia and India, emissions data is reported on the local fiscal year (July 1 to June 30 in Australia, April 1 to March 31 in India). For comparability to annual-level ML-based predictions, emissions are converted to the Gregorian calendar year by assuming equal emissions in each month and summing over the months within each calendar year.

Appendix A.6. Annual Emissions Factors

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)¹⁴, and European Commission Joint Research Open Power Plants Database (JRC-PPDB-OPEN)¹⁵ data, and country-specific calibration based on IEA¹⁶ 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 by the base value to account for regional differences in power plant efficiency (due to age, technology level, and size), fuel quality, and the impact of ambient conditions on carbon intensity that are not currently modeled.

Appendix B. CNN Model and Training Details

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.

¹⁰ <https://campd.epa.gov/>

¹¹ <https://www.eea.europa.eu/data-and-maps/data/european-union-emissions-trading-scheme-17/eu-ets-data-download-latest-version>

¹² <https://www.cleanenergyregulator.gov.au/NGER>

¹³ <https://cea.nic.in/>

¹⁴ <https://www.epa.gov/egrid>

¹⁵ <https://data.jrc.ec.europa.eu/dataset/9810feeb-f062-49cd-8e76-8d8cfd488a05>

¹⁶ <https://www.iea.org/>

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.

In summary, we optimized the following hyperparameters:

- Patch size [8, 16, 32, 48, 64]
- Backbone truncation layer [block2_pool, block3_pool, block4_pool]
- Attention heads [1, 8]
- Augmentation magnitude [0.5, 1, 1.5, 2, 2.5, 3]
- Weight decay [10^{-6} , 10^{-5} , 10^{-4} , 10^{-3}]
- Early stopping patience [10, 20, 30]
- Batch size [32, 54, 128]
- Learning rate [10^{-5} , 10^{-2}]
- Number of epochs [40, 50, 60, 70, 80, 90, 100]
- Loss (for regression only) [mean squared error, huber]

Appendix C. Generation Model Features

This section details the features used in our generation model (Section 3.6). The 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:

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

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

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:

$$\bar{y}_s = \frac{1}{|P||M_s|} \sum_{p \in P} \sum_{m_s \in M_s} y_{pm_s}$$

$$\bar{z}_s = \frac{1}{|P||M_s|} \sum_{p \in P} \sum_{m_s \in M_s} z_{pm_s}$$

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:

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

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:

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

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

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:

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

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

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.

Appendix D. External Validation for Sounding Model

We acquired additional electricity generation data for external validation from reporting agencies in Türkiye and India, as described in Appendix A.4. A comparison to our models is shown in Tables A2 A3. Since India reports only at the daily level, not hourly, it could only be used to validate the generation and annual-level models, not the sounding models that require hourly generation data. Some plants in the US, Europe, and Australia were excluded from training due to the strict filters set

in Section 3.4; however, they do meet the looser also described in that Section. There are very few external validation plants for FGD as compared to NDT because the plant filters were relaxed for NDT but not FGD at inference time.

Table A2. On/off classification and regression external validation on 42 NDT plants across the US, Europe, Australia, and Türkiye, 2015-2022.

| Satellite | Model | Classification | Regression | |
|-------------|--------------|----------------------|----------------------|-------------------------|
| | | mAP [CI] | RMSE [CI] | MBE [CI] |
| PlanetScope | XGBoost | 0.779 [0.733, 0.837] | 0.221 [0.206, 0.233] | 0.004 [-0.017, 0.028] |
| PlanetScope | CNN - RESISC | 0.782 [0.721, 0.851] | 0.215 [0.198, 0.231] | -0.009 [-0.034, 0.020] |
| Sentinel-2 | XGBoost | 0.749 [0.708, 0.802] | 0.240 [0.223, 0.255] | -0.008 [-0.037, 0.024] |
| Sentinel-2 | CNN - RESISC | 0.752 [0.713, 0.808] | 0.231 [0.210, 0.250] | -0.040 [-0.068, -0.008] |
| Sentinel-2 | CNN - BEN | 0.738 [0.704, 0.795] | 0.232 [0.214, 0.248] | -0.014 [-0.041, 0.018] |
| Landsat 8 | XGBoost | 0.730 [0.689, 0.803] | 0.257 [0.230, 0.278] | 0.033 [0.000, 0.066] |
| Landsat 8 | CNN - RESISC | 0.717 [0.669, 0.772] | 0.271 [0.244, 0.294] | -0.017 [-0.053, 0.019] |
| Landsat 8 | CNN - BEN | 0.721 [0.683, 0.791] | 0.272 [0.243, 0.295] | -0.003 [-0.039, 0.039] |

Table A3. On/off classification and regression external validation on 3 FGD plants in Türkiye, 2015-2022.

| Satellite | Model | Classification | Regression | |
|-------------|--------------|----------------------|----------------------|-------------------------|
| | | mAP [CI] | RMSE [CI] | MBE [CI] |
| PlanetScope | XGBoost | 0.664 [0.531, 0.909] | 0.487 [0.463, 0.504] | -0.373 [-0.416, -0.323] |
| PlanetScope | CNN - RESISC | 0.778 [0.623, 0.943] | 0.394 [0.392, 0.395] | -0.299 [-0.303, -0.295] |
| Sentinel-2 | XGBoost | 0.615 [0.519, 0.766] | 0.455 [0.411, 0.492] | -0.350 [-0.407, -0.277] |
| Sentinel-2 | CNN - RESISC | 0.604 [0.523, 0.871] | 0.461 [0.416, 0.500] | -0.356 [-0.410, -0.292] |
| Sentinel-2 | CNN - BEN | 0.682 [0.529, 0.940] | 0.474 [0.444, 0.496] | -0.390 [-0.419, -0.346] |
| Landsat 8 | XGBoost | 0.586 [0.517, 0.765] | 0.519 [0.444, 0.584] | -0.406 [-0.516, -0.296] |
| Landsat 8 | CNN - RESISC | 0.550 [0.517, 0.714] | 0.509 [0.479, 0.543] | -0.400 [-0.477, -0.329] |
| Landsat 8 | CNN - BEN | 0.569 [0.568, 0.829] | 0.430 [0.408, 0.447] | -0.318 [-0.372, -0.260] |

The same naive baselines used in Tables 3 and 4 were applied for the external validation set as a comparison to the ML approach: simply predicting the mean off/on value or mean capacity factor over the training set. For NDT, these baselines produced an RMSE of 0.326 to 0.343 and MBE of 0.119 to 0.180. For FGD, the baselines were 0.436 to 0.461 and -0.406 to -0.299 for RMSE and MBE, respectively.

In external validation, our NDT models achieved regression performance comparable to the cross-validation results but performed more poorly for classification with mAP ranging from 0.72 to 0.78.

For FGD, our external validation results were significantly poorer. Classification mAP ranged from 0.55 to 0.78, and regression RMSE did not always outperform the naive baseline. Even more concerning, there was a large negative bias ranging from -0.299 to -0.406. However, it must be noted that the set of FGD plants available for external validation was extremely limited: only 3 plants compared to 42 available for NDT. Therefore, additional external validation of FGD is needed. Furthermore, we suspect that the high error and large negative bias for those three Turkish FGD plants could be because the average annual capacity factor of those three plants is around 0.8, yet our FGD training set consists of only European and US plants with average annual capacity factors around 0.4 and 0.5, respectively.

References

1. Paris Agreement to the United Nations Framework Convention on Climate Change; Number 16-1104, T.I.A.S., 2015.
2. Ge, M.; Friedrich, J. 4 Charts Explain Greenhouse Gas Emissions by Countries and Sectors. *World Resources Institute* 2020. <https://www.wri.org/blog/2020/02/greenhouse-gas-emissions-by-country-sector>.

3. Watch, C. Climate watch historical GHG emissions. *World Resources Institute: Washington, DC, USA* **2022**. <https://www.climatewatchdata.org/ghg-emissions>.
4. IEA, I. Greenhouse gas emissions from energy data explorer. IEA Paris, 2023. <https://www.iea.org/data-and-statistics/data-tools/greenhouse-gas-emissions-from-energy-data-explorer>.
5. Ullah, S.; You, Q.; Chen, D.; Sachindra, D.; AghaKouchak, A.; Kang, S.; Li, M.; Zhai, P.; Ullah, W. Future population exposure to daytime and nighttime heat waves in South Asia. *Earth's Future* **2022**, *10*, e2021EF002511.
6. 2023 shatters climate records, with major impacts, 2023. <https://wmo.int/news/media-centre/2023-shatters-climate-records-major-impacts>.
7. Energy, G. Global Energy & CO2 Status Report. *International Energy Agency: Paris, France* **2019**. <https://www.iea.org/reports/global-energy-co2-status-report-2019>.
8. Energy, G. CO2 Emissions in 2022. *International Energy Agency: Paris, France* **2023**. <https://www.iea.org/reports/co2-emissions-in-2022>.
9. Sloss, L.L. Efficiency and emissions monitoring and reporting. *Paris: IEA Clean Coal Centre* **2011**.
10. Liu, F.; Duncan, B.N.; Krotkov, N.A.; Lamsal, L.N.; Beirle, S.; Griffin, D.; McLinden, C.A.; Goldberg, D.L.; Lu, Z. A methodology to constrain carbon dioxide emissions from coal-fired power plants using satellite observations of co-emitted nitrogen dioxide. *Atmospheric Chemistry and Physics* **2020**, *20*, 99–116. doi:10.5194/acp-20-99-2020.
11. 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.; Miller, C.E. Quantifying Global Power Plant Carbon Dioxide Emissions With Imaging Spectroscopy. *AGU Advances* **2021**, *2*.
12. Kuhlmann, G.; Broquet, G.; Marshall, J.; Clément, V.; Löscher, A.; Meijer, Y.; Brunner, D. Detectability of CO₂ emission plumes of cities and power plants with the Copernicus Anthropogenic CO₂ Monitoring (CO2M) mission. *Atmospheric Measurement Techniques* **2019**, *12*, 6695–6719.
13. Vaughn, T.L.; Bell, C.S.; Pickering, C.K.; Schwietzke, S.; Heath, G.A.; Pétron, G.; Zimmerle, D.J.; Schnell, R.C.; Nummedal, D. 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* **2018**, *115*, 11712–11717.
14. Cusworth, D.H.; Thorpe, A.K.; Miller, C.E.; Ayasse, A.K.; Jiorle, R.; Duren, R.M.; Nassar, R.; Mastrogiacomo, J.P.; Nelson, R.R. Two years of satellite-based carbon dioxide emission quantification at the world's largest coal-fired power plants. *Atmospheric Chemistry and Physics* **2023**, *23*, 14577–14591.
15. Nassar, R.; Hill, T.G.; McLinden, C.A.; Wunch, D.; Jones, D.B.; Crisp, D. Quantifying CO₂ emissions from individual power plants from space. *Geophysical Research Letters* **2017**, *44*, 10–045.
16. Nassar, R.; Mastrogiacomo, J.P.; Bateman-Hemphill, W.; Callum McCracken, J.P.; MacDonald, C.G.; Hill, T.; O'Dell, C.W.; Kiel, M.; Crisp, D. Advances in quantifying power plant CO₂ emissions with OCO-2. *Remote Sensing of Environment* **2021**, *264*, 112579.
17. Hu, Y.; Shi, Y. Estimating CO₂ emissions from large scale coal-fired power plants using OCO-2 observations and emission inventories. *Atmosphere* **2021**, *12*, 811.
18. Guo, W.; Shi, Y.; Liu, Y.; Su, M. CO₂ emissions retrieval from coal-fired power plants based on OCO-2/3 satellite observations and a Gaussian plume model. *Journal of Cleaner Production* **2023**, *397*, 136525.
19. Lin, X.; van der A, R.; de Laat, J.; Eskes, H.; Chevallier, F.; Ciais, P.; Deng, Z.; Geng, Y.; Song, X.; Ni, X.; others. Monitoring and quantifying CO₂ emissions of isolated power plants from space. *Atmospheric Chemistry and Physics* **2023**, *23*, 6599–6611.
20. Yang, D.; Zhang, H.; Liu, Y.; Chen, B.; Cai, Z.; Lü, D. Monitoring carbon dioxide from space: Retrieval algorithm and flux inversion based on GOSAT data and using CarbonTracker-China. *Advances in Atmospheric Sciences* **2017**, *34*, 965–976.
21. Shim, C.; Han, J.; Henze, D.K.; Yoon, T. Identifying local anthropogenic CO₂ emissions with satellite retrievals: a case study in South Korea. *International Journal of Remote Sensing* **2019**, *40*, 1011–1029.
22. Zheng, T.; Nassar, R.; Baxter, M. Estimating power plant CO₂ emission using OCO-2 XCO₂ and high resolution WRF-Chem simulations. *Environmental Research Letters* **2019**, *14*, 085001.
23. Yang, S.; Lei, L.; Zeng, Z.; He, Z.; Zhong, H. An Assessment of Anthropogenic CO₂ Emissions by Satellite-Based Observations in China. *Sensors* **2019**, *19*, 1118.

24. Reuter, M.; Buchwitz, M.; Schneising, O.; Krautwurst, S.; O'Dell, C.W.; Richter, A.; Bovensmann, H.; Burrows, J.P. Towards monitoring localized CO₂ emissions from space: co-located regional CO₂ and NO₂ enhancements observed by the OCO-2 and S5P satellites. *Atmospheric Chemistry and Physics* **2019**, *19*, 9371–9383.
25. Nassar, R.; Moeini, O.; Mastrogiacomo, J.P.; O'Dell, C.W.; Nelson, R.R.; Kiel, M.; Chatterjee, A.; Eldering, A.; Crisp, D. Tracking CO₂ emission reductions from space: A case study at Europe's largest fossil fuel power plant. *Frontiers in Remote Sensing* **2022**, *3*.
26. Sierk, B.; Fernandez, V.; Bézy, J.L.; Meijer, Y.; Durand, Y.; Courrèges-Lacoste, G.B.; Pachot, C.; Löscher, A.; Nett, H.; Minoglou, K.; others. The Copernicus CO₂M mission for monitoring anthropogenic carbon dioxide emissions from space. International Conference on Space Optics—ICSO 2020. International Society for Optics and Photonics, 2021, Vol. 118523M.
27. Kuhlmann, G.; Henne, S.; Meijer, Y.; Brunner, D. Quantifying CO₂ Emissions of Power Plants With CO₂ and NO₂ Imaging Satellites. *Frontiers in Remote Sensing* **2021**, *2*.
28. Gray, M.; Watson, L.; Ljungwaldh, S.; Morris, E. Nowhere to hide: Using satellite imagery to estimate the utilisation of fossil fuel power plants. *Carbon Tracker Initiative* **2018**. <https://carbontracker.org/reports/nowhere-to-hide/>.
29. Mommert, M.; Sigel, M.; Neuhausler, M.; Scheibenreif, L.M.; Borth, D. Characterization of Industrial Smoke Plumes from Remote Sensing Data. NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning, 2020.
30. Hanna, J.; Mommert, M.; Scheibenreif, L.M.; Borth, D. Multitask Learning for Estimating Power Plant Greenhouse Gas Emissions from Satellite Imagery. NeurIPS 2021 Workshop on Tackling Climate Change with Machine Learning, 2021.
31. Jain, A. Employing Deep Learning to Quantify Power Plant Greenhouse Gas Emissions via Remote Sensing Data. AAAI 2022 Fall Symposium: The Role of AI in Responding to Climate Challenges, 2022.
32. Couture, H.D.; O'Connor, J.; Mitchell, G.; Söldner-Rembold, I.; D'souza, D.; Karra, K.; Zhang, K.; Kargar, A.R.; Kassel, T.; Goldman, B. Towards tracking the emissions of every power plant on the planet. NeurIPS Workshop on Tackling Climate Change with Machine Learning, 2020, Vol. 3.
33. Hobbs, M.; Kargar, A.R.; Couture, H.; Freeman, J.; Söldner-Rembold, I.; Ferreira, A.; Jeyaratnam, J.; O'Connor, J.; Lewis, J.; Koenig, H.; McCormick, C.; Nakano, T.; Dalisay, C.; Davitt, A.; Gans, L.; Lewis, C.; Volpato, G.; Gray, M.; McCormick, G. Inferring carbon dioxide emissions from power plants using satellite imagery and machine learning. IEEE International Geoscience and Remote Sensing Symposium, 2023.
34. Planet imagery product specifications. *Planet Labs* **2022**. https://assets.planet.com/docs/Planet_Combined_Imagery_Product_Specs_letter_screen.pdf.
35. 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.; Magalhães, P.S. Monitoring pasture aboveground biomass and canopy height in an integrated crop–livestock system using textural information from PlanetScope imagery. *Remote Sensing* **2020**, *12*, 2534.
36. Moon, M.; Richardson, A.D.; Friedl, M.A. Multiscale assessment of land surface phenology from harmonized Landsat 8 and Sentinel-2, PlanetScope, and PhenoCam imagery. *Remote Sensing of Environment* **2021**, *266*, 112716.
37. Sentinel-2 - Missions - Sentinel Online.
38. Main-Knorn, M.; Pflug, B.; Louis, J.; Debaecker, V.; Müller-Wilm, U.; Gascon, F. Sen2Cor for sentinel-2. Image and Signal Processing for Remote Sensing XXIII. SPIE, 2017, Vol. 10427, pp. 37–48.
39. Shikwambana, L.; Ncipha, X.; Malahlela, O.E.; Mbatha, N.; Sivakumar, V. Characterisation of aerosol constituents from wildfires using satellites and model data: A case study in Knysna, South Africa. *International Journal of Remote Sensing* **2019**, *40*, 4743–4761.
40. Landsat 8 | U.S. Geological Survey.
41. Marchese, F.; Genzano, N.; Neri, M.; Falconieri, A.; Mazzeo, G.; Pergola, N. A multi-channel algorithm for mapping volcanic thermal anomalies by means of Sentinel-2 MSI and Landsat-8 OLI data. *Remote Sensing* **2019**, *11*, 2876.
42. Mia, M.B.; Fujimitsu, Y.; Nishijima, J. 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* **2017**, *7*, 118.

43. Xiong, Q.; Wang, Y.; Liu, D.; Ye, S.; Du, Z.; Liu, W.; Huang, J.; Su, W.; Zhu, D.; Yao, X.; others. A cloud detection approach based on hybrid multispectral features with dynamic thresholds for GF-1 remote sensing images. *Remote Sensing* **2020**, *12*, 450.
44. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. CVPR, 2016.
45. 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.
46. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* **2014**.
47. Sumbul, G.; Charfuelan, M.; Demir, B.; Markl, V. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. IEEE International Geoscience and Remote Sensing Symposium, 2019.
48. Ilse, M.; Tomczak, J.; Welling, M. Attention-based Deep Multiple Instance Learning. International Conference on Machine Learning, 2018.
49. Hobbs, M.; Rouzbeh, A.; Couture, H.; Freeman, J.; Jeyaratnam, J.; Lewis, J.; Koenig, H.; Nakano, T.; Dalisay, C.; McCormick, C. Estimating Fossil Fuel Power Plant Carbon Dioxide Emissions Globally with Remote Sensing and Machine Learning. *AGU23* **2023**.
50. Freeman, J.; others. Power sector: electricity generation. <https://github.com/climatetracecoalition/methodology-documents/tree/main/2023/Power>, Climate TRACE, 2023.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.