

---

# Towards Tracking the Emissions of Every Power Plant on the Planet

---

**Heather D. Couture<sup>1,5</sup>, Joseph O'Connor<sup>2</sup>, Grace Mitchell<sup>1</sup>, Isabella Söldner-Rembold<sup>2</sup>, Durand D'souza<sup>3</sup>, Krishna Karra<sup>1</sup>, Keto D. Zhang<sup>1</sup>, Ali Rouzbeh Kargar<sup>1</sup>, Thomas Kassel<sup>1</sup>, Brian W. Goldman<sup>4</sup>, Daniel Tyrrell<sup>4</sup>, Wanda Czerwinski<sup>4</sup>, Alok Talekar<sup>4</sup>, Colin McCormick<sup>1,6</sup>**  
<sup>1</sup>WattTime, <sup>2</sup>Energy & Clean Air Analytics, <sup>3</sup>Carbon Tracker,  
<sup>4</sup>Google.org, <sup>5</sup>Pixel Scientia Labs, <sup>6</sup>Georgetown University

## Abstract

Greenhouse gases emitted from fossil-fuel-burning power plants are a major contributor to climate change. Current methods to track emissions from individual sources are expensive and only used in a few countries. While carbon dioxide concentrations can be measured globally using remote sensing, background fluctuations and low spatial resolution make it difficult to attribute emissions to individual sources. We use machine learning to infer power generation and emissions from visible and thermal power plant signatures in satellite images. By training on a data set of power plants for which we know the generation or emissions, we are able to apply our models globally. This paper demonstrates initial progress on this project by predicting whether a power plant is on or off from a single satellite image.

## 1 Introduction

Greenhouse gases (GHG) produced by burning fossil fuels are a major contributor to the global climate crisis. Fossil fuel power plants account for 30% of GHG emissions [1] or an estimated 15 billion tons of  $CO_2$  per year [2]. These eighteen thousand or so fossil fuel burning plants [3] meet 85% of global electricity demand [4]. As renewable energy becomes more economical, fossil fuel power is decreasing in many parts of the world [5], but new plants are still being built in others, especially in China and India [6]. Coal-fired power plants are the single largest emitter of  $CO_2$  worldwide [2].

Averting the most severe impacts of climate change requires understanding the sources of emissions – which power plants and how much they emit. The United States is one of the few countries that publicly releases high-time-resolution data on emissions from individual power plants. Every major US plant has on-site Continuous Emissions Monitoring Systems (CEMS) and reports data to the Environmental Protection Agency. But the costs of installing and maintaining these systems make them impractical for use in many countries. In addition, monitoring systems can be tampered with or suffer technical failures. Other countries report annual emissions totals that may be rough estimates instead of actual measurements and lack verification [7].

Policymakers, scientists, and companies need granular and reliable data to design sustainability strategies. To that end, we are building a system to estimate the emissions of every power plant in the world using remote sensing data from existing satellites. We have developed machine learning models for predicting whether a plant is on or off from a single image and tested them on two different satellite constellations. This paper describes our models and analyzes their performance.

Emissions estimates from this project will be made public. While use cases are still being explored, we anticipate a number of possible opportunities. Renewable energy developers can use the data to pinpoint locations where new wind or solar farms will have the most impact. Regulatory agencies

can create and enforce new environmental policy. Individual citizens can see how much their local power plants are contributing to climate change. Finally, NGOs can track progress toward the Paris Climate Agreement, which will be renegotiated at COP26 in 2021.

Our project is part of a larger initiative called Climate TRACE (Tracking Real-time Atmospheric Carbon Emissions) that aims to track human-emitted GHGs from all sectors – from electricity to heavy industry to shipping [8].

## 2 Background

Concentrations of  $CO_2$  in the atmosphere are measured locally at observatories such as Mauna Loa (producing the well-known Keeling Curve [9]). They are also measured globally by two satellite missions: the Orbiting Carbon Observatory (OCO-2) and the Greenhouse Gases Observing Satellite (GOSAT). These satellites use spectroscopic methods based on the absorption of reflected sunlight to estimate the column-averaged dry-air mole fraction of  $CO_2$ , known as  $XCO_2$ .

Several researchers have used  $XCO_2$  to try to estimate emissions from individual sources such as fossil fuel power plants [10–15]. However, this is difficult because GHG concentrations are also affected by natural sources such as  $CO_2$  and methane releases from the ocean, volcanoes, biomass decomposition, and soil, plant, and animal respiration. This background noise makes it difficult to detect changes in concentration due to anthropogenic sources such as burning fossil fuels (coal, natural gas, and oil) and industrial processes (e.g. cement production and iron smelting). Anthropogenic sources also fluctuate depending on local demand and fuel prices, among other factors.

A second challenge is distinguishing emissions sources that are close together. For example, each measurement from OCO-2 represents a one square mile area on the ground, which is too coarse for differentiating many emissions sources. The dispersal of emitted  $CO_2$  by wind also complicates this situation. New satellite data with better spatial and temporal resolution of  $CO_2$  concentrations may eventually become available [16, 17].

High-spatial-resolution optical imagery is another potential route to monitoring GHG emissions from individual sources. It has the important advantage of being available today at relatively high time 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 optical imagery cannot directly measure  $CO_2$  concentrations, we develop a set of proxies that are directly tied to emissions, such as visible vapor plumes from power plant cooling towers. Our project also makes use of thermal infrared (TIR) data that can detect heat from fossil fuel combustion and serve as a proxy for the associated emissions.

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 [18]. Our project extends their work to additional regions, satellite constellations, and power plant cooling technologies. We tackle each new satellite and power plant cooling technology with a unique model and, in future work, will aggregate estimates across satellites to produce monthly or annual emissions totals.

## 3 Methodology

Power plants emit GHGs through a chimney called the flue stack, producing a small smoke plume. Plants that are more efficient or have air pollution control equipment generally have smoke plumes that are difficult to see. For this reason, directly inspecting this smoke only provides a weak indicator of emissions. A better indicator of emissions is the vapor plume from a power plant's cooling system. Fossil fuel power plants must continuously cool the process steam that is used to turn a turbine to generate electricity. We use visible and thermal signs of cooling as additional indicators of emissions.

Most plants use cooling towers or once-through cooling to cool the process steam. Cooling towers produce a large water vapor plume that is often easy to spot in visible imagery. Once-through cooled plants release heat by discharging warm water to a nearby source, which can be detected by instruments such as the Landsat 8 Thermal Infrared Sensor. We create models using both visible and thermal imaging for each cooling type.

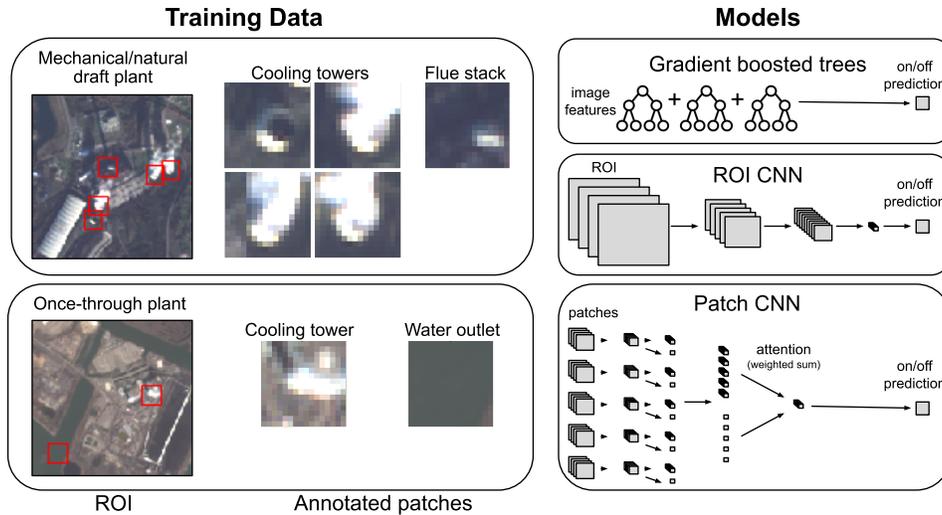


Figure 1: An overview of the training data and model types we apply to satellite images. We apply gradient boosted trees to a set of engineered features summarizing pixel values of the image and annotated patches. The ROI CNN model is applied to a  $1.5 \times 1.5 \text{ km}^2$  ROI around the plant. The patch CNN model applies a CNN to each patch and aggregates features with an attention model.

While the goal of our project is to estimate total emissions for a plant over some period of time, our initial efforts have focused on a simpler setup: can we predict whether a plant is on or off from a single satellite image? The remainder of this paper will explore models to answer this question.

Our models operate on Sentinel-2 [19] and Landsat 8 [20] images. Each has multiple bands (using light of different wavelengths) with different spatial resolutions. We upsampled all lower resolution bands to match the highest resolution band for each satellite. Each band also has a different distribution of pixel values. We standardized each band to a mean of 0.5 and a standard deviation of 0.25, placing most pixel values between 0 and 1.

We developed four different types of machine learning models to predict the on/off status of a plant from a satellite image. Each model is trained on satellite images paired with the ground truth on/off status (see Appendix A for details on the data). Figure 1 provides an overview of these methods. Our region of interest (ROI) models use a  $1.5 \times 1.5 \text{ km}^2$  ROI around the plant as input. Our patch models use a set of annotated  $16 \times 16$  pixel patches, including cooling towers, flue stacks, and water outlets. We created both gradient boosted tree (GBT) and convolutional neural net (CNN) models.

Each model also incorporates temperature and dew point data, as they may influence the visibility of the plume. They are included in the feature set for GBT models and concatenated to the top layer before the softmax for the CNN models.

We trained our models on plants in countries for which we have hourly or sub-hourly generation data by gathering and cleaning data from multiple sources. Appendix A provides more details on this process. We trained different models for each power plant cooling type.

We tested the accuracy of our models over historical data (2013-2020) from coal-burning power plants in the United States, Europe, and Australia with a capacity greater than 500 MW, for which we have ground truth generation data. After data cleaning, this resulted in 127 mechanical/natural draft plants and 36 once-through plants. The number of images for each satellite is shown in Table 2 in the appendix. We used four-fold cross-validation to run our experiments, where all images from a particular plant were placed in the same fold. Two folds were used for training, one for validation, and one for testing. All reported results are the average test performance over the four folds.

Our two classes, on and off, are very imbalanced with 82% of images labeled as on. Therefore, we selected the mean average precision (mAP) to measure model performance. mAP is calculated by computing the average precision (the area under a precision-recall curve) for each class (on and off) and averaging over the two classes. More details on each of these models is provided in Appendix B and the procedure we used for training is provided in Appendix C.

## 4 Results

To focus on the most suitable CNN architectures, we first experimented with different pre-trained CNNs to initialize the ROI model. Appendix D provides more details and concludes that models transferred from remote sensing data sets performed best. However, including all satellite bands did not always perform best. We suspect that this is due to either the extra bands not providing any new information or because the pre-trained CNN using all the bands was not tuned for clouds (or plumes).

Next, we compared the four models described in Section 3 on both Sentinel-2 and Landsat 8 for mechanical/natural draft and once-through cooling. Table 1 shows the results. Overall, model accuracy was greater for Sentinel-2, which is expected as it has a higher spatial resolution than Landsat 8. For mechanical/natural draft, the patch models all produced significantly better results than the non-patch ones, demonstrating the importance of localizing models to the most relevant regions in the image. The CNN models beat the GBT ones for mechanical/natural draft, whereas they performed similarly for once-through. The results on once-through plants show the greater challenges for this cooling type where there is no vapor plume, only a small smoke plume and thermal signatures. We also have 3.5 times fewer once-through plants in our training set, increasing the challenge.

Table 1: A comparison of model performance (mAP) on different satellites and cooling types.

Model type	Sentinel-2	Landsat 8
<b>Mechanical/natural draft</b>		
ROI GBT	0.647	0.616
ROI+Patch GBT	0.789	0.713
ROI CNN	0.681	0.651
Patch CNN	0.813	0.756
<b>Once-through</b>		
ROI GBT	0.616	0.627
ROI+Patch GBT	0.626	0.606
ROI CNN	0.612	0.598
Patch CNN	0.623	0.566

Further analysis has shown that the different models performed best for different categories of images; therefore, we plan to use an ensemble approach to incorporate all model predictions when estimating monthly generation and emissions. Our analysis has also shown some common modes of failure: 1) When temperature is high or humidity is low, plants are often predicted to be off when they are on because the water vapor plume is not visible. 2) Mechanical draft plants are problematic – especially in unfavorable weather conditions – as the vapor plume can be quickly dispersed. From these insights into our current models, we are now experimenting with different model types and gathering data from alternative satellite constellations in order to build better models for these situations.

## 5 Conclusions and Future Work

We have shown that it is possible to achieve high quality estimates of power plant behavior using existing satellites, an important first step toward global emissions monitoring. The next step is aggregating these predictions into a power generation estimate for each plant over some period of time such as one month. The generation predictions will then be converted into  $CO_2$  emissions estimates using fuel efficiency information for each country. We have implemented an initial prototype of this end-to-end model and are working to improve each component and apply it globally. Our initial focus is on  $CO_2$  emissions from coal-fueled plants that use mechanical/natural draft or once-through cooling; however, we plan to extend our methods to other GHGs, fuel types, and cooling technologies.

While we are careful to assess our models on a held-out set using cross-validation, we must also take steps to validate our models globally. Most countries do not release emissions data that is granular enough to validate estimations for individual power plants, but we plan to validate at the annual country level and against national fuel statistics that are available for all countries.

Two main challenges still exist related to the coverage of our model. The first is that the countries in our ground truth training data may not be representative of plants globally. Less wealthy countries are underrepresented. Without mitigation, this will bias our model to assume that all plants in our global test set are operating in wealthier countries. Power plants in other countries may have different equipment or operational patterns. As our training data is currently focused on the US and Europe, our models may also be biased towards northerly climates and biomes, which could be significant due to the role weather plays in the models. For these reasons, we continue our efforts to collect data from a more diverse set of countries for use in training or validation.

The second challenge is that we are limited by the time of day of satellite observations. Visible satellite imagery is generally only available during the day, not at night. We are further limited by the sun-synchronous orbit of Sentinel-2 and Landsat 8 that observe at roughly the same time each day. This is problematic because power generation varies through the day and night. We are experimenting with visible imagery from satellites that visit more frequently and at other times of day, including Planet Labs' PlanetScope and the Chinese National Space Administration's Gaofen, and with radar data from satellites that can observe at night. Ultimately, the broader our coverage, the more reliable our models will be.

## Acknowledgments and Disclosure of Funding

This work was supported by an AI Impact award from Google.org and funding from Bloomberg Philanthropies.

## References

- [1] Mengpin Ge and Johannes Friedrich. 4 charts explain greenhouse gas emissions by countries and sectors. *World Resources Institute*, Feb 2020. <https://www.wri.org/blog/2020/02/greenhouse-gas-emissions-by-country-sector>.
- [2] Hannah Ritchie and Max Roser.  $CO_2$  and greenhouse gas emissions. *Our World in Data*, Aug 2020. <https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions>.
- [3] Logan Byers, Johannes Friedrich, Roman Hennig, Aaron Kressig, Xinyue Li, Laura Malaguzzi Valeri, and Colin McCormick. A global database of power plants. *World Resources Institute*, April 2018.
- [4] Hannah Ritchie and Max Roser. Energy. *Our World in Data*, Jul 2018. <https://ourworldindata.org/energy>.
- [5] Mark Morey. U.s. coal-fired electricity generation in 2019 falls to 42-year low. *U.S. Energy Information Administration*, May 2020. <https://www.eia.gov/todayinenergy/detail.php?id=43675>.
- [6] Benjamin Storrow. Global  $CO_2$  emissions were flat in 2019—but don't cheer yet. *Scientific American*, Feb 2020. <https://www.scientificamerican.com/article/global-co2-emissions-were-flat-in-2019-but-dont-cheer-yet/>.
- [7] What is transparency and reporting? *United Nations Framework Convention on Climate Change*. <https://unfccc.int/process-and-meetings/transparency-and-reporting/the-big-picture/what-is-transparency-and-reporting>.
- [8] Climate TRACE. <https://www.climatetrace.org>.
- [9] The keeling curve. <https://sioweb.ucsd.edu/programs/keelingcurve/>, 2020.
- [10] Ray Nassar, Timothy G Hill, Chris A McLinden, Debra Wunch, Dylan BA Jones, and David Crisp. Quantifying  $CO_2$  emissions from individual power plants from space. *Geophysical Research Letters*, 44(19):10–045, 2017.
- [11] Dongxu Yang, Huifang Zhang, Yi Liu, Baozhang Chen, Zhaonan Cai, and Daren Lü. Monitoring carbon dioxide from space: Retrieval algorithm and flux inversion based on GOSAT data and using CarbonTracker-China. *Advances in Atmospheric Sciences*, 34(8):965–976, 2017.
- [12] Changsub Shim, Jihyun Han, Daven K Henze, and Taeyeon Yoon. Identifying local anthropogenic  $CO_2$  emissions with satellite retrievals: a case study in south korea. *International Journal of Remote Sensing*, 40(3):1011–1029, 2019.
- [13] Tao Zheng, Ray Nassar, and Martin Baxter. Estimating power plant  $CO_2$  emission using OCO-2  $XCO_2$  and high resolution WRF-Chem simulations. *Environmental Research Letters*, 14(8): 085001, 2019.

- [14] Shaoyuan Yang, Liping Lei, Zhaocheng Zeng, Zhonghua He, and Hui Zhong. An assessment of anthropogenic  $CO_2$  emissions by satellite-based observations in china. *Sensors*, 19(5):1118, 2019.
- [15] Maximilian Reuter, Michael Buchwitz, Oliver Schneising, Sven Krautwurst, Christopher W O’Dell, Andreas Richter, Heinrich Bovensmann, and John P Burrows. Towards monitoring localized co 2 emissions from space: co-located regional co 2 and no 2 enhancements observed by the oco-2 and s5p satellites. *Atmospheric Chemistry and Physics*, 19(14):9371–9383, 2019.
- [16] David Crisp, Yasjka Meijer, Rosemary Munro, Kevin Bowman, and Abhishek Chatterjee et al. A constellation architecture for monitoring carbon dioxide and methane from space. *Committee on Earth Observation Satellites*, 2018.
- [17] Bern Sierk, Jean-Loup Bézy, Armin Löscher, and Yasjka Meijer. The european  $CO_2$  monitoring mission: observing anthropogenic greenhouse gas emissions from space. In *International Conference on Space Optics—ICSO 2018*, volume 11180, page 111800M. International Society for Optics and Photonics, 2019.
- [18] Matt Gray, Laurence Watson, Sebastian Ljungwaldh, and Edward Morris. Nowhere to hide: Using satellite imagery to estimate the utilisation of fossil fuel power plants. *Carbon Tracker Initiative*, Oct 2018. <https://carbontracker.org/reports/nowhere-to-hide/>.
- [19] Copernicus sentinel data, processed by ESA. [https://www.esa.int/Our\\_Activities/Observing\\_the\\_Earth/Copernicus](https://www.esa.int/Our_Activities/Observing_the_Earth/Copernicus).
- [20] Landsat data, processed by NASA and USGS. <https://landsat.gsfc.nasa.gov/landsat-8/landsat-8-overview/>.
- [21] Global coal plant tracker. <https://endcoal.org/global-coal-plant-tracker/>.
- [22] World electric power plants database. <https://www.spglobal.com/platts/ko/products-services/electric-power/world-electric-power-plants-database>.
- [23] Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 2017.
- [24] Air markets program data. <https://ampd.epa.gov/ampd/>.
- [25] European network of transmission system operators for electricity. <https://www.entsoe.eu/>.
- [26] National electricity market. <https://aemo.com.au/en/energy-systems/electricity/national-electricity-market-nem>.
- [27] Dark sky. <https://darksky.net>.
- [28] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International Conference on Machine Learning*, 2018.
- [29] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proc. International Conference on Learning Representations*, 2015.
- [30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, 2016.
- [31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proc. CVPR*, 2009.
- [32] Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IEEE International Geoscience and Remote Sensing Symposium*, 2019.
- [33] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

Table 2: Ground truth data set sizes: number of training images available after filtering for each satellite and cooling type.

Cooling type	Sentinel-2	Landsat 8
Mechanical/natural draft	17,568	12,316
Once-through	4,484	3,469

## A Data Sources, Collating, and Cleaning

**Power plant data** We used the Global Power Plant Database (GPPD) [3], Global Coal Plant Tracker [21], and some manual work to obtain the geolocation of each plant. We obtained the capacity, fuel type, and cooling type from S&P Global Platts’ World Electric Power Plants Database [22].

**Satellite imagery** Our current sources of satellite imagery are Sentinel-2 [19] and Landsat 8 [20]. We used Google Earth Engine [23] to access these images and associated cloud masks. Sentinel-2 images range from 10 to 60 m/pixel resolution, depending on the band. Landsat 8 is 30 to 100 m/pixel.

**Ground truth labels** The US and Taiwan are some of the only countries reporting both plant emissions and power generation at hourly intervals. Some regions like Europe and Australia report generation only. Still others report daily aggregated generation. Due to the limited number of countries with high-time-resolution emissions data, we used more widely available generation data to train our models.

Hourly or sub-hourly power generation data was obtained from the Environmental Protection Agency’s Air Markets Program Data for the US (AMPD) [24], from the European Network of Transmission System Operators for Electricity for Europe (ENTSOE) [25], and from the Australian Energy Market Operator (AEMO) [26].

**Weather** We obtained temperature and dew point readings for each satellite image from DarkSky, which interpolates from nearby weather station readings [27].

**Data collating** We joined the GPPD and Platts databases to get the geolocation and power plant properties. We matched these records with the information provided in our ground truth generation sources (AMPD, ENTSOE, and NEM) to get information on how much each unit of each plant generated and, where available, emitted.

**Patches** We manually annotated a set of bounding boxes ("patches") around each cooling tower, flue stack, and water outlet at each power plant in our training set.

**Data cleaning** We filtered out plants with mislabeled images (e.g., no generation reported but plume visible) or errors in geolocation for further investigation. Plants for the mechanical/natural draft and once-through groups were selected those with a minimum of 70% of capacity attributed to the cooling type for the group, more than 50% of generation attributed to coal, and a minimum capacity of 500 MW. We also dropped individual images from our training set in the following situations: 1) more than 20% cloud cover, 2) more than 30% of image pixels lie outside tile, 3) more than 10% of generation attributed to a non-coal fuel, and 4) images with duplicate time stamps (we kept the most recent one). Table 2 lists the number of training images our collating and cleaning process produced for each cooling type and each satellite.

## B Model Architecture

In this section, we will describe each of the models mentioned in Section 3 in more detail.

**ROI GBT** We designed a set of features to capture how much "white stuff" is in the middle of the image over the plant compared to the rest of the image, to help control for clouds. These features include the mean, standard deviation, and 90<sup>th</sup> percentile pixel values in central image crops of increasing scale (50 × 50 pixels, 100 × 100 pixels, the whole image), producing nine features per band. We used these features to train a gradient boosted trees (GBT) model using XGBoost (1024 trees, a maximum depth of 3, and a learning rate of 0.01).

**ROI+Patch GBT** The patch variant of our GBT model includes all of the ROI GBT features and a set of patch features. We compute the mean, standard deviation, and 90<sup>th</sup> percentile pixel values in each patch, then aggregate each across all patches in an image by computing the mean, minimum, and maximum. We used all of these features to train an XGBoost model.

**ROI CNN** We applied a pre-trained convolutional neural network (CNN) to the image, with the original aggregation and classification layers removed. We added a new aggregation layer using an attention mechanism that computes a weighted average of the CNN features where the weights are learned from the features themselves [28]. We then added a new softmax classification layer to predict the on/off status of the plant.

We experimented with both VGG16 [29] and ResNet50 [30] architectures and used transfer learning to initialize the weights. Appendix D describes the pre-trained CNNs we used in more detail.

**Patch CNN** As an alternative to the ROI model, these models focus on the annotated patches. We encoded each patch with a CNN (truncated after the second convolutional block for both VGG16 and ResNet50) and then aggregated the result as a weighted sum using an attention model [28]. We placed a softmax classification layer on top.

## C Model Training

We trained our CNN models using the AdamW optimizer that uses weight decay for regularization. We also regularized with image augmentation, including transformations for flipping, rotation, brightness, contrast, Gaussian noise, translation, and zooming. As we train 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.

As our on and off classes are imbalanced (82:18), we used over- and under-sampling during training so that each batch is randomly generated with roughly half of the images from each class. Additionally, we balanced by plant so that each plant is represented equally during training.

We used transfer learning to initialize the CNN weights from a pre-trained model (discussed more in the next section). The weights for the attention and classification layers were initialized randomly. We first kept the transferred weights fixed and trained for one epoch with five different random initializations. We then selected the model with the best validation loss and continued training all CNN layers.

We tuned all hyperparameters to maximize mAP on the validation set with cross-validation; all reported results are on the test set.

## D Transfer Learning

Training a CNN on a limited number of images is challenging because the model over-fits very easily – it performs well on the training set but does not generalize to a held out test set. Transfer learning has come to be the standard first step on small data sets. The CNN weights are initialized with weights from a model that was pre-trained on a different data set that is larger.

The most commonly used data set for transfer learning is ImageNet – a data set of 1.4 million photographs from 1,000 different classes of objects and scenes [31]. These photographs have quite a different appearance than remote sensing images. They also contain only three color channels (RGB), whereas Sentinel-2 and Landsat 8 images have 13 and 11 bands, respectively.

We experimented with three different data sets for pre-training: ImageNet, BigEarthNet (which consists of Sentinel-2 images labeled by land cover class) [32], and RESISC (which consists of aerial RGB images labeled by land cover and scene classes – notably, including cloud) [33]. We applied ImageNet and RESISC pre-trained models to the RGB bands of Sentinel-2 and Landsat 8. The BigEarthNet pre-trained model used 10 bands from Sentinel-2. In applying to Sentinel-2, we used the same bands; for Landsat 8 we matched the bands with a similar wavelength and substituted in the remaining Landsat 8 bands for those remaining unmatched bands. Fine-tuning enabled our models to overcome the initial band mismatch.

Table 3 shows a comparison of the different CNN architectures and data sets for pre-training. The RESISC pre-trained model performed best for mechanical/natural draft on Sentinel-2, while BigEarthNet was the best for all other setups. We suspect that RESISC performs so well even though it uses only 3 bands instead of 10 because it includes a cloud class. This makes it easier to fine-tune for spotting vapor plumes from mechanical/natural draft plants, as plumes look a lot like clouds. Even with the mismatch of bands in transferring BigEarthNet to Landsat8, this setup showed the greatest success for this satellite.

All further experiments (including Table 1) used a RESISC pre-trained CNN for Sentinel-2 on mechanical/natural draft plants and BigEarthNet for all other models.

Table 3: Mean average precision for both types of plants with the ROI CNN model.

CNN architecture	Pre-training data set	Bands	Sentinel-2	Landsat 8
<b>Mechanical/natural draft</b>				
VGG16	ImageNet	RGB	0.743	0.630
VGG16	BigEarthNet	10 bands	0.700	<b>0.651</b>
ResNet50	RESISC	RGB	<b>0.752</b>	0.614
<b>Once-through</b>				
VGG16	ImageNet	RGB	0.568	0.594
VGG16	BigEarthNet	10 bands	<b>0.612</b>	<b>0.598</b>
ResNet50	RESISC	RGB	0.563	0.525