



RESEARCH ARTICLE

# Pollution Tracker: Finding industrial sources of aerosol emission in satellite imagery

Peter Manshausen<sup>1\*</sup>, Duncan Watson-Parris<sup>1,2</sup>, Lena Wagner<sup>3</sup>, Pirmin Maier<sup>3</sup>, Sybrand J. Muller<sup>3</sup>, Gernot Ramminger<sup>3</sup> and Philip Stier<sup>1</sup>

<sup>1</sup>Atmospheric, Oceanic, and Planetary Physics, Department of Physics, University of Oxford, Oxford, UK

<sup>2</sup>Scripps Institution of Oceanography and Halıcıoğlu Data Science Institute, University of California San Diego, La Jolla, CA, USA

<sup>3</sup>GAF AG, Munich, Germany

\*Corresponding author. Email: [peter.manshausen@physics.ox.ac.uk](mailto:peter.manshausen@physics.ox.ac.uk)

**Received:** 26 January 2023

**Keywords:** Computer vision, remote sensing, classification, air pollution, aerosols, clouds

## Abstract

The effects of anthropogenic aerosol, solid or liquid particles suspended in the air, are the biggest contributor to uncertainty in current climate perturbations. Heavy industry sites, such as coal power plants and steel manufacturers, large sources of greenhouse gases, also emit large amounts of aerosol in a small area. This makes them ideal places to study aerosol interactions with radiation and clouds. However, existing data sets of heavy industry locations are either not public, or suffer from reporting gaps. Here, we develop a supervised deep learning algorithm to detect unreported industry sites in high-resolution satellite data, using the existing data sets for training. For the pipeline to be viable at global scale, we employ a two-step approach. The first step uses 10 m resolution data, which is scanned for potential industry sites, before using 1.2 m resolution images to confirm or reject detections. On held out test data, the models perform well, with the lower resolution one reaching up to 94% accuracy. Deployed to a large test region, the first stage model yields many false positive detections. The second stage, higher resolution model shows promising results at filtering these out, while keeping the true positives, improving the precision to 42% overall, so that human review becomes feasible. In the deployment area, we find five new heavy industry sites which were not in the training data. This demonstrates that the approach can be used to complement existing data sets of heavy industry sites.

## Impact Statement

Aerosols from industrial air pollution play an important role in both public health and climate. Complete and freely accessible data on large aerosol sources are of great interest, but not currently available. Here, we present a scoping study to fill the gaps in existing data sets by automatically detecting heavy industry in satellite data. This approach can be rolled out globally to complete the available data.

## 1. Climate change and the need for pollution source databases

Anthropogenic aerosol, such as black carbon or sulfate, plays an important role for the earth's energy budget. Aerosol can interact directly with solar radiation by absorbing or scattering light. More subtly, it can change the properties of clouds to make them more reflective by providing additional

condensation nuclei and increasing the number of cloud droplets, which in turn sets off a series of further adjustments. Together, aerosol-induced changes in cloud properties are the largest source of uncertainty in current estimates of man-made climate forcing [Masson-Delmotte et al., 2021]. This is mainly because the different effects are difficult to untangle and experiments are infeasible in nature on the large scales required.

Research has therefore relied heavily on so-called opportunistic experiments [Christensen et al., 2021], where an aerosol source is known and localized, such that polluted and unpolluted clouds can be directly compared. Examples of such sources include volcanoes [Malavelle et al., 2017], ships [Coakley and Walsh, 2002], and heavy industry [Toll et al., 2019]. In the past, studies of these sources have been limited to some hundreds of ‘events’, such as a ship or heavy industry site producing a visible change in clouds. Research is ongoing to enlarge such data sets, e.g. by automatically detecting such cloud changes caused by ships with machine learning [Watson-Parris et al., 2022]. Another approach is to use more information about the sources themselves: Manshausen et al. [2022] use millions of ship paths and reconstruct where their pollution has changed cloud properties, uncovering previously undetected changes in cloud water. A similar approach seems promising for large industrial sites.

However, existing data such as that collected by the International Energy Agency or by private companies is not publicly available. The databases that are available such as the EDGAR data do not include point sources outside Europe and North America [Janssens-Maenhout et al., 2015]. Here, we build on two openly accessible databases of point sources to fill in such gaps (see Data).

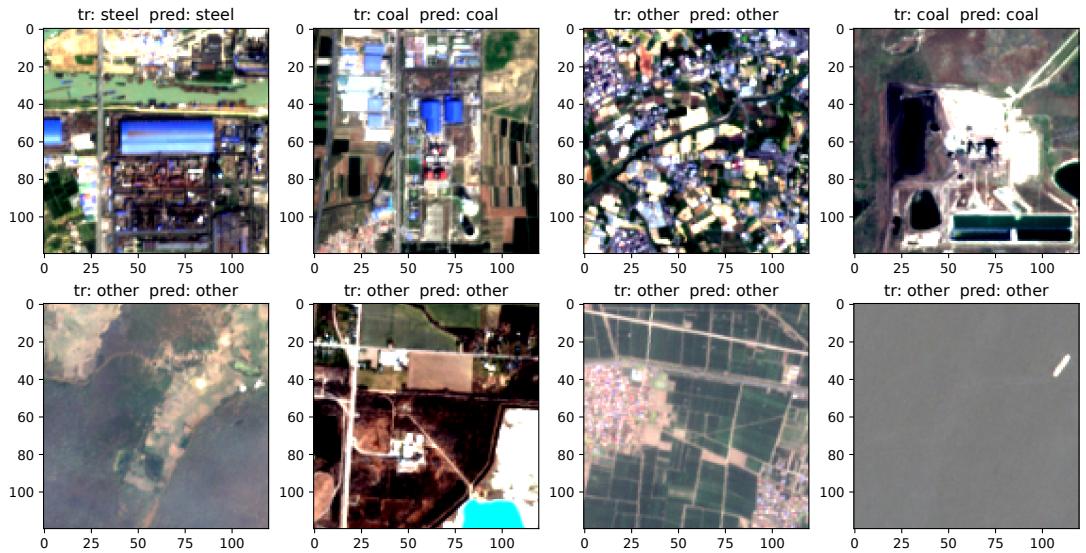
We also note that there is a growing interest in independent estimates of greenhouse gas emissions using satellite observations, and therefore in heavy industry sites as large emitters. Among others, [www.climatetrace.org](http://www.climatetrace.org) and its partners are accounting for global emissions. Recently, large point sources of methane have also been discovered from satellite imagery [Lauvaux et al., 2022].

This project is informed by recent advances in detecting large industrial sites and land cover classification: Sheng et al. [2020] use a deep neural network approach to identify oil refineries in the United States. However, their work uses satellite data at relatively high resolution of 2.5 m, which requires handling large amounts of data for continental-scale detection. For land cover classification, Sumbul et al. [2021] propose ‘BigEarthNet’, a deep neural network which classifies 10-60 m resolution Sentinel2 images into 19 land use classes. However, their ‘industrial and commercial units’-class is the one that performs second to worst. If the locations of power plants are known, Mommert et al. [2021] have shown that ResNet architectures are skillful at identifying their type, e.g. distinguishing nuclear, gas, and coal. With known locations, Hanna et al. [2021] propose a method to then estimate power plant CO<sub>2</sub> emissions.

Our work addresses the gap of detecting heavy industry in Sentinel2 data, with the aim of developing a method that can, in principle, be deployed globally. To this end, we propose a two-step approach in which we first perform classification on 10 m resolution data and then additional filtering of false positives at 1.2 m resolution, using Bing Maps [Microsoft Corporation, 2023].

## 2. Data

This work aims to extend some of the existing and freely available data sets of heavy industry. Namely, we base the project on the location data from Global Energy Monitor [2022a,b] and from McCarter et al. [2021] for coal and steel plants worldwide. These data sets include over 4000 samples of heavy industry sites. To include non-industry sites, we add the same number of random land cover patches by choosing the patch at a fixed distance of 0.1° longitude to the east of each industry site center.



**Figure 1.** Some example visualizations the RGB channels of the Sentinel2 patches in the data set. For a human, characteristic features of steel plants are e.g. the dark surface together with industrial structures like in the top first panel. For coal plants, cooling towers are common, note the bright crescent-shape in the top second panel, bottom right corner. Labels show truth (tr.) and prediction (pred.) of our best model..

We download Sentinel2 data for the 10 m resolution bands 02, 03, 04, and 08 (i.e. RGB and near infrared), in patches of 240x240 pixels centered around the locations from the above databases. Querying the database for cloud-free Sentinel2 overpasses, we use the date range from March to October 2021 and take the output with the lowest cloud cover in the scene and in the patch. The data is normalized to have zero mean and unit standard deviation in each channel. Leaving out the NIR band slightly reduces the performance during training, so all models include it.

As the Sentinel cloud mask is not completely accurate, we iterate over the data set twice, training only on half the data in each case. The models obtained that way are used to predict the labels of the held-out data. We review the most mismatched prediction/ground truth pairs by eye, which in some cases are cloud-covered, or do not show any industry (errors in the site data). This leaves a data set of 7,829 patches, with 28% of coal, 22% of steel, and 50% of non-industry samples. For training, we use 80% of the data with 10% held out for testing and 10% for validation. We employ data augmentation using random rotation, random crop, and random flip. We find that we obtain the best results with a center crop to 140x140 pixels and then a random crop to the target 120x120 pixel patches (the same size as those used by [Sumbul et al., 2021]). This cropping is in order to vary where the industry plant is with respect to the scene. It should always be at the center of the 240x240 patch, then center and random crops allow it to be offset from the center in the final 120x120 training patch. Each patch is assigned to one of three classes coal, steel, other/no industry. Examples of what this data looks like in the three visible channels can be seen in Figure 1.

For the second step, we download Bing Maps [Microsoft Corporation, 2023] aerial RGB imagery in the same locations as discussed above, and filter in the same way. The resolution here is 1.2 m, and we limit the image size to 1400 pixels, giving an on-the-ground size of just below 1200m, the same as the size of the Sentinel2 images. Splitting and augmentation are as above.

**Table 1.** Performance evaluation of the best model (ResNet50v2) on the held out test data for both the lower and higher resolution cases. Overall accuracy of the top-1 prediction is 0.94 for the lower, 0.77 for the higher resolution case.

Category	Lower res			Higher res		
	Precision	Recall	F1-score	Precision	Recall	F1-score
coal	0.96	0.95	0.95	0.72	0.82	0.76
steel	0.88	0.86	0.87	0.84	0.53	0.65
other	0.95	0.97	0.96	0.85	0.93	0.89
weight. avg.	0.94	0.94	0.94	0.76	0.77	0.76

### 3. Models and Training

For step one, following [Sumbul et al., 2021], we initially try a number of CNN architectures, trained with categorical crossentropy loss. These are: ResNet50, VGG16, EfficientNet, InceptionNet, and ResNet50v2. For each of the architectures, we add a dense layer of three neurons at the end of the network for the three classes. We find that among the models, ResNet50 (the original, and v2, [He et al., 2016]) converge the quickest. Compared to ResNet50, ResNet50v2 reaches slightly higher categorical accuracy. We train ResNet50v2 for up to 200 epochs, using early stopping after 50 epochs of no decrease in validation loss. We start training with a learning rate of  $10^{-3}$  and decay with  $e^{-0.03}$  each epoch after 30 epochs.

For step two on high-resolution data, we train the same ResNet50v2 architecture, using early stopping and a decaying learning rate as above. In both cases, training from scratch performs better than using pre-trained weights from imagenet [Jia Deng et al., 2009].

## 4. Results

### 4.1. Step 1: Lower resolution

The Sentinel2 model performs well on the held out data, with precision values of up to 0.96 for coal plants. Table 1 shows performance for all classes in terms of precision, recall, and F1-score. It seems that generally steel plants are harder to detect (lower recall), as they are often not as visually distinctive as coal plants. Figure 1 shows a random selection of predictions on the test data together with the ground truth. Figure A 3 shows a selection of patches that were misclassified by the model.

### 4.2. Step 2: Higher resolution

In the higher resolution case, the model does not perform as well on the test data, as shown in Table 1. This is surprising given the larger level of detail in the data. The model performs worst on steel plants, with the confusion matrix, shown in Table 2 indicating that it mistakes many of them for either coal plants or no-industry sites – of the 23% steel plants, just over half are correctly identified (compare also the recall of 53% in Table 1). The overall lower skill may be due to insufficient training samples to learn high resolution properties, different hyperparameters being needed for the high resolution case, or the model making more cautious predictions with a higher spread of probabilities. A promising result is that the model has high recall on the ‘other’ category, which is crucial for using it as a second stage filter on the positive predictions of the low resolution stage.

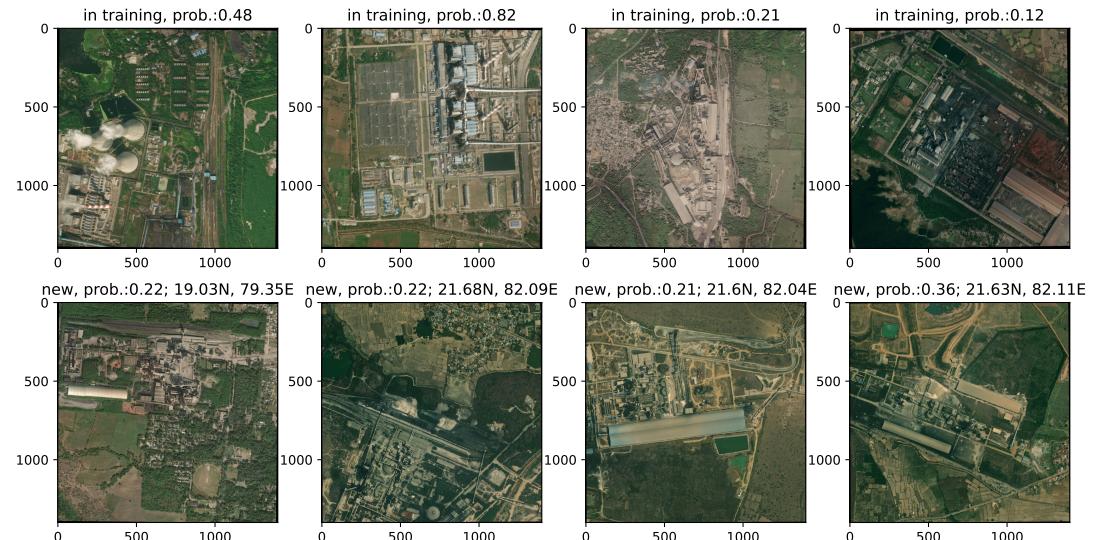
**Table 2.** High resolution model confusion matrix, normalised by the number of samples. Each line shows the fraction of samples belonging to a category that were predicted as each of the categories, i.e. the first line means out of all patches, 20% are coal and were identified as coal, 2% are coal and were identified as steel, 3% are coal and were identified as other. This is evaluated on the held out test data.

Truth	Prediction		
	coal	steel	other
coal	0.20	0.02	0.03
steel	0.06	0.12	0.05
other	0.03	0.01	0.48

#### 4.3. Large scale deployment

In a deployment test, we deploy our first stage model to a  $4^\circ \times 4^\circ$  area centered around ( $21^\circ$  N,  $82^\circ$  E) in north-eastern India, focusing on coal plants. Checking the results manually, we find that the model finds some coal plants that are present in the region, but also returns many false positives even at high probability thresholds. This is most likely due to the more imbalanced real world data, with a large fraction of no-industry tiles. A similar behaviour was also observed by [Sheng et al. \[2020\]](#).

Figure A 4 shows some example positive predictions for coal. For some false positives, we can guess why the model confuses them for coal plants. For instance, they are often situated near rivers (a), produce small clouds (b) near the cooling towers, show built structures that may look similar to urban fabric (c), or are close to open air mines (d) or coal piles. The figure also shows two true positives (e), with the characteristic dark surface, industrial structures, cooling tower, and river (f).



**Figure 2.** RGB images of the second stage higher resolution predictions in the deployment region. Scenes that the first stage classed as  $>50\%$  probability of coal plant are downloaded at higher resolution from Bing Maps and fed into the model. The second stage detects 23 out of 24 coal plants that were present in the training set (top row, labelled ‘in training’, and with probability output). It also detects five ‘new’ coal plants not in the data set from Global Energy Monitor, four of which are shown in the bottom row with their respective locations..

More quantitatively, of the 51 coal power plants in the study area which are present in the training data set, the model correctly identifies 24 (probability of >50% for coal). This seems low (recall of 0.47) given the recall of 0.95 on the testing data (cf. Table 1). This may be linked to poorer data quality in the deployment set, with sites off-centered, cut in half, or covered by cloud. The model makes 890 positive predictions in total, most of which are false positives (giving a precision of only 0.03, much lower than on the testing data, because of the imbalanced deployment data with many more ‘other’ patches). This is where the second stage, higher resolution model comes in: We download the high resolution Bing Maps scenes in the locations of the 890 positives from the first-stage and make predictions on them with the second stage model. Even though the accuracy of the high resolution model is lower on the test data, it performs well at this confirmation/rejection task: Setting a threshold of 10% probability for coal confirms 23 out of the 24 known coal plants (recall of 0.96), while reducing the 890 potential positives to 55 (precision of 0.42).

Reviewing the remaining positives, we find five additional coal plants, which were not present in the training, validation, or test set. Four of these, together with four of the 23 that were in the data sets, are shown in Figure 2. The five detected sites are cement production facilities (confirmed by an internet search of the location), which form a complex with coal power plants for their large energy consumption. (The fifth site, not shown in the figure is at 21.37N, 83.61E.) Two additional metal processing sites are detected which likely also use coal power. These new sites were probably identified from the architectural or surface elements that are there because they are coal powered, such as coal piles, smokestacks, and polluted, dark surfaces. The above values for precision and recall after each step were given without taking into account the new sites, and are therefore a conservative estimate of the actual values.

This demonstrates that our approach can be used to fill in gaps in existing data sets. At the same time it raises the question of how to treat complex industrial facilities which combine coal power with manufacturing of products like steel or cement.

## 5. Discussion and future work

The above results depend strongly on the choice of threshold values for detection in both stages of the algorithm. To illustrate the problem, true and false positive rates of the first and second stage coal plant detections are shown in Figure A 5. In the first stage, the priority is a low false positive rate, as we want to download as little high-resolution data as possible for the second stage. In order not to miss too many true positives, we chose a threshold of 50% here.

For the second stage, the priority is a high true positive rate, so that industry sites present in the high-resolution data are not missed. This is why we chose a (relatively low) threshold of 10% in this study, which still allows to filter out many false positives, as shown by the steep decline in false positives in Figure A 5. Ultimately, these thresholds should be chosen depending on computational resources (imposing a lower limit on the first stage threshold, as a low threshold will mean many positives that need to be verified in the second stage), as well as the resources for human review (imposing a lower limit on the second stage threshold, as a low threshold will mean many positives that need to be verified by eye).

We would like to further improve step one and lower its false positive rate. For this we propose: (i) Deploying on two or three Sentinel2 overpass dates for each tile, which would presumably eliminate the ‘small cloud’ false positives. (ii) Using human-in-the-loop learning, i.e. retraining the model with the correctly labeled false positives from the deployment stage after human review, or just increasing the number of non-industry cases in the training data. (iii) Re-running on shifted tiles to exclude the

case where a site is cut in half and therefore not recognized.

This last point is connected to the cropping hyperparameter choices: We start with 240x240 pixel scenes, then center crop to 140x140, and then random crop to 120x120. While the initial and final size are modelling choices, the center crop size should be seen as a hyperparameter. 140x140 pixels is the best choice for training and then testing on the data. It balances variation in where the site is in the final scene with ensuring enough of it is present for the model to learn important features. However, for the deployment, a less conservative choice, i.e. a larger value than 140, which will train the network with more off-center scenes, could do better on the random positions that it will be deployed to. This could potentially improve the model's false negative rate.

To improve step two, in order to achieve accuracy comparable to the low-resolution first step, independent hyperparameter tuning seems promising. Ultimately, the goal is to roll out this approach globally, and to fill the gaps in existing data sets. To do this, careful evaluation of detection thresholds at both stages is needed, balancing the threshold's effect on false positive and false negative detections.

## 6. Conclusions

There is a need for publicly available databases of locations of heavy industry, which act as large point sources of aerosol and greenhouse gas pollution. We show that heavy industry complexes can be detected in globally available 10 m-resolution satellite data. ResNet architectures are particularly well adapted to this task and reach high accuracy for both steel and coal plants. However, this approach has many false positive detections at deployment. This can be remedied by a second step, which makes use of the available high-resolution data, but only in the locations where detection seems likely based on low-resolution data. Our approach combines the advantages of being not too data intensive, not yielding too many false positives (high precision), and requiring minimal human review of detections. Furthermore, the method succeeds at finding new heavy industry sites not present in the existing data sets it was trained on. Finally, these efforts may contribute to the larger effort of independently and bottom-up quantifying not only air pollution, but also greenhouse gas emissions.

## A. Appendix

**Acknowledgments.** We would like to acknowledge the important work done by Global Energy Monitor which provided training data for this project. We thank Lilli Freischem for helpful comments on the manuscript.

**Funding Statement.** This work was supported by the European Union's Horizon 2020 research and innovation program under Marie Skłodowska-Curie grant agreement No 860100 (iMIRACLI). D.W.P. and P.S. were supported by the UK Natural Environment Research Council project ACRUISE (NE/S005099/1). P.S. additionally acknowledges support from the European Research Council Project RECAP under the European Union's Horizon 2020 research and innovation program grant 724602 and from the FORCeS project under the European Union's Horizon 2020 research program with grant agreement 821205. P. Man. acknowledges GAF AG for hosting a three month research visit to their Munich office.

**Competing Interests.** None.

**Data Availability Statement.** Replication code can be found on github: [https://github.com/ManshaP/sentinel\\_industry](https://github.com/ManshaP/sentinel_industry). Sentinel2 data is freely available and can be downloaded with the sentinelhub API. Bing Maps is available non-commercially and was downloaded using the code referenced above.

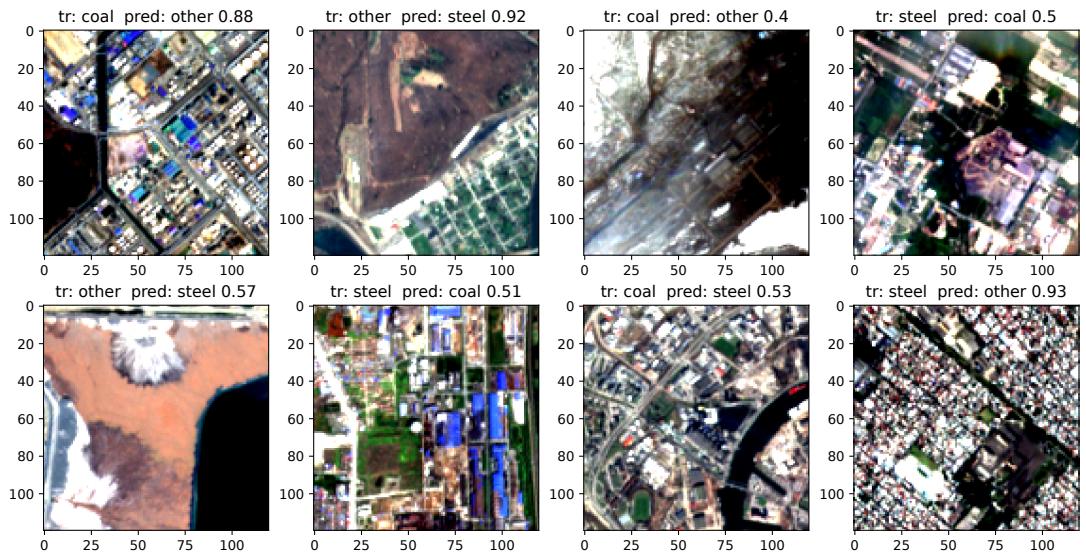
**Ethical Standards.** The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

**Author Contributions.** Conceptualization: P.Man., D.W.P., G.R., P.Mai., P.S. Methodology: all authors. Data curation: P.Man., L.W., S.J.M. Model training and analysis: P.Man. Data visualisation: P.Man. Writing original draft: P.Man. All authors approved the final submitted draft.

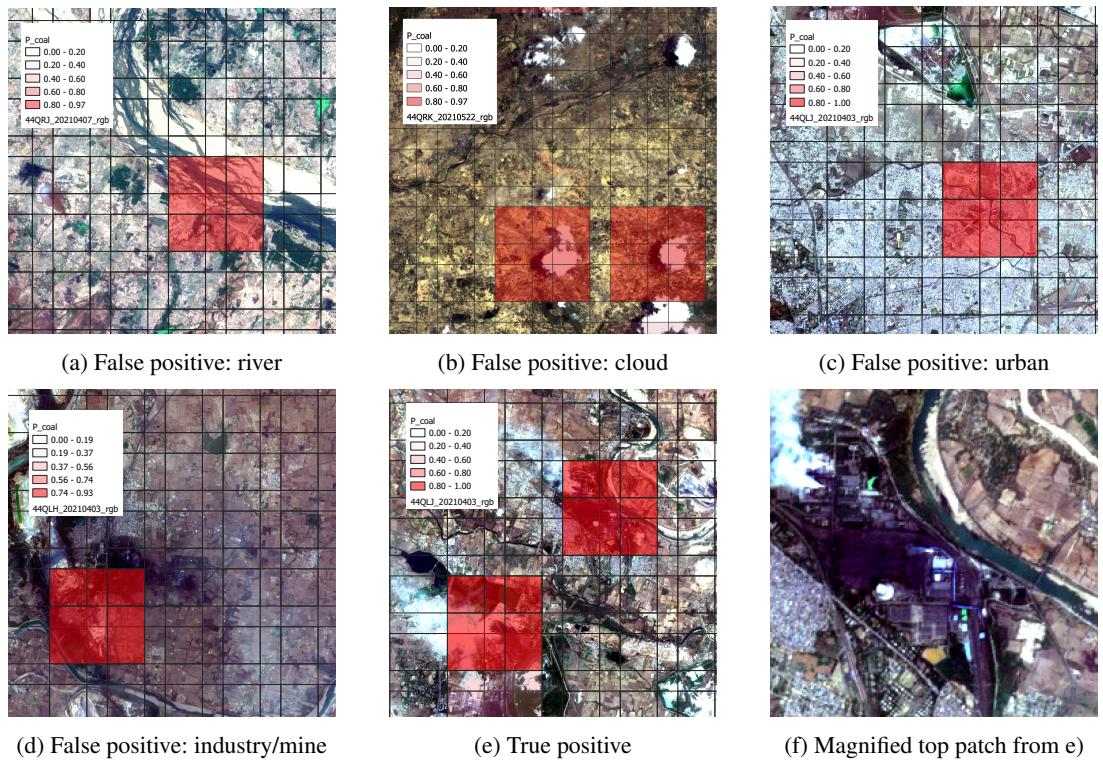
**Supplementary Material.** None.

## References

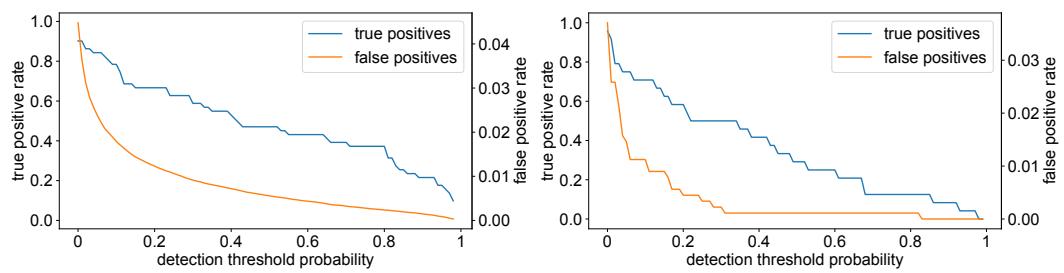
- M. Christensen, A. Gettelman, J. Cermak, G. Dagan, M. Diamond, A. Douglas, G. Feingold, F. Glassmeier, T. Goren, and D. P. Grosvenor. Opportunistic Experiments to Constrain Aerosol Effective Radiative Forcing. *Atmospheric Chemistry and Physics*, (August), 2021.
- J. A. Coakley and C. D. Walsh. Limits to the aerosol indirect radiative effect derived from observations of ship tracks. *Journal of the Atmospheric Sciences*, 59(3 PT 2):668–680, 2002. ISSN 00224928. doi: 10.1175/1520-0469(2002)059<0668:ltair>2.0.co;2.
- Global Energy Monitor. Global Coal Plant Tracker. 2022a. URL <https://globalenergymonitor.org/projects/global-coal-plant-tracker/>.
- Global Energy Monitor. Global Steel Plant Tracker. 2022b. URL <https://globalenergymonitor.org/projects/global-steel-plant-tracker/>.
- J. Hanna, M. Mommert, L. Scheibenreif, and D. Borth. Multitask Learning for Estimating Power Plant Greenhouse Gas Emissions from Satellite Imagery. *NeurIPS 2021 Workshop on Tackling Climate Change with Machine Learning*, 2021.
- K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem, pages 770–778, 2016. ISBN 9781467388504. doi: 10.1109/CVPR.2016.90.
- G. Janssens-Maenhout, M. Crippa, D. Guizzardi, F. Dentener, M. Muntean, G. Pouliot, T. Keating, Q. Zhang, J. Kurokawa, R. Wankmüller, H. Denier Van Der Gon, J. J. Kuenen, Z. Klimont, G. Frost, S. Darras, B. Koffi, and M. Li. HTAP-v2.2: A mosaic of regional and global emission grid maps for 2008 and 2010 to study hemispheric transport of air pollution. *Atmospheric Chemistry and Physics*, 15(19):11411–11432, 2015. ISSN 16807324. doi: 10.5194/acp-15-11411-2015.
- Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. pages 248–255, 2009. doi: 10.1109/cvprw.2009.5206848.
- T. Lauvaux, C. Giron, M. Mazzolini, A. D’Aspremont, R. Duren, D. Cusworth, D. Shindell, and P. Ciais. Global assessment of oil and gas methane ultra-emitters. *Science*, 375(6580):557–561, 2022. ISSN 10959203. doi: 10.1126/science.abj4351.
- F. F. Malavelle, J. M. Haywood, A. Jones, A. Gettelman, L. Clarisse, S. Bauduin, R. P. Allan, I. H. H. Karset, J. E. Kristjánsson, L. Oreopoulos, N. Cho, D. Lee, N. Bellouin, O. Boucher, D. P. Grosvenor, K. S. Carslaw, S. Dhomse, G. W. Mann, A. Schmidt, H. Coe, M. E. Hartley, M. Dalvi, A. A. Hill, B. T. Johnson, C. E. Johnson, J. R. Knight, F. M. O’Connor, P. Stier, G. Myhre, S. Platnick, G. L. Stephens, H. Takahashi, and T. Thordarson. Strong constraints on aerosol-cloud interactions from volcanic eruptions. *Nature*, 546(7659):485–491, 2017. ISSN 14764687. doi: 10.1038/nature22974.
- P. Manshausen, D. Watson-Parris, M. Christensen, and P. Stier. Invisible Ship Tracks Show Large Cloud Sensitivity to Aerosol. *Nature*, 610, 2022. doi: 10.1038/s41586-022-05122-0.
- V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Z. (eds.). IPCC, 2021: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Technical report, IPCC, 2021.
- M. McCarten, M. Bayaraa, B. Caldecott, C. Christiaen, P. Foster, C. Hickey, D. Kampmann, C. Layman, C. Rossi, K. Scott, K. Tang, N. Tkachenko, , and D. Yoken. Global Database of Iron and Steel Production Assets. Spatial Finance Initiative. 2021.
- Microsoft Corporation. Bing Maps, 2023. URL <https://learn.microsoft.com/en-us/bingmaps/>.
- M. Mommert, L. Scheibenreif, J. Hanna, and D. Borth. Power Plant Classification From Remote Imaging With Deep Learning. *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2021-July:6391–6394, 2021. doi: 10.1109/IGARSS47720.2021.9553219.
- H. Sheng, J. Irvin, S. Munukutla, S. Zhang, C. Cross, K. Story, R. Rustowicz, C. Elsworth, Z. Yang, M. Omara, R. Gautam, R. B. Jackson, and A. Y. Ng. OGNet: Towards a Global Oil and Gas Infrastructure Database using Deep Learning on Remotely Sensed Imagery. (*NeurIPS*):1–8, 2020. URL <http://arxiv.org/abs/2011.07227>.
- G. Sumbul, A. De Wall, T. Kreuziger, F. Marcelino, H. Costa, P. Benevides, M. Caetano, B. Demir, and V. Markl. BigEarthNet-MM: A Large-Scale, Multimodal, Multilabel Benchmark Archive for Remote Sensing Image Classification and Retrieval [Software and Data Sets]. *IEEE Geoscience and Remote Sensing Magazine*, 9(3):174–180, 2021. ISSN 21686831. doi: 10.1109/MGRS.2021.3089174.
- V. Toll, M. Christensen, J. Quaas, and N. Bellouin. Weak average liquid-cloud-water response to anthropogenic aerosols. *Nature*, 572(7767):51–55, 2019. ISSN 14764687. doi: 10.1038/s41586-019-1423-9. URL <http://dx.doi.org/10.1038/s41586-019-1423-9>.
- D. Watson-Parris, M. W. Christensen, A. Laurenson, D. Clewley, E. Gryspeerdt, and P. Stier. Shipping regulations lead to large reduction in cloud perturbations. *Proceedings of the National Academy of Sciences of the United States of America*, 119(41), 2022. ISSN 10916490. doi: 10.1073/pnas.2206885119.



**Figure A 3.** Wrong predictions of the lower resolution model on the test data set. Labels show truth, prediction, and probability. The model struggles with coal and steel plants that are not isolated but in an industrial or urban area (top first and bottom last), geological formations it mistakes for steel plants (top second, bottom first), as well as distinguishing steel and coal (top last, bottom second and third).



**Figure A 4.** Positives for coal detection in the deployment data at low resolution. Transparent red patches indicate where the model gives high probability of coal plant detection. However, only the red patches in e) are true positives (the top one magnified in f shows a characteristic cooling tower).



**Figure A 5.** Dependence of true and false positive rates on thresholds for detection of coal plants in the first stage (left) and the second stage (right).