Deep learning applied to CO_2 power plant emissions quantification using simulated satellite images

Joffrey Dumont Le Brazidec[1], Pierre Vanderbecken[1], Alban Farchi[1], Grégoire Broquet[2], Gerrit Kuhlmann[3], and Marc Bocquet[1]

[1]CEREA, École des Ponts and EDF R&D, Île-de-France, France
[2]Laboratoire des Sciences du Climat et de l'Environnement, LSCE/IPSL, CEA-CNRS-UVSQ, Université Paris-Saclay, 91198 Gif-sur-Yvette, France
[3]Swiss Federal Laboratories for Materials Science and Technology (Empa), Dübendorf, Switerzland

Correspondence: Joffrey Dumont Le Brazidec (joffrey.dumont@enpc.fr)

Abstract.

The quantification of emissions of greenhouse gases and air pollutants through the inversion of plumes in satellite images remains a complex problem that current methods can only assess with significant uncertainties. The anticipated launch of the CO_2M (Copernicus Anthropogenic Carbon Dioxide Monitoring) satellite constellation in 2026 is expected to provide high-resolution images of CO_2 (carbon dioxide) column-averaged mole fractions (XCO_2), opening up new possibilities. However, the inversion of future CO_2 plumes from CO_2M will encounter various obstacles. A challenge is the CO_2 plume low signal-to-noise ratio, due to the variability of the background and instrumental errors in satellite measurements. Moreover, uncertainties in the transport and dispersion processes further complicate the inversion task.

To address these challenges, deep learning techniques, such as neural networks, offer promising solutions for retrieving emissions from plumes in XCO_2 images. Deep learning models can be trained to identify emissions from plume dynamics simulated using a transport model. It then becomes possible to extract relevant information from new plumes and predict their emissions.

In this paper, we develop a strategy employing convolutional neural networks (CNN) to estimate the emission fluxes from a plume in a pseudo XCO_2 image. Our dataset used to train and test such methods includes pseudo images based on simulations of hourly XCO_2, NO_2 and wind fields near various power plants in Eastern Germany, tracing plumes from anthropogenic and biogenic sources. CNN models are trained to predict emissions from three power plants that exhibit diverse characteristics. The power plants used to assess the deep learning model's performance are not used to train the model. We find that the CNN model outperforms state of the art plume inversion approaches, achieving highly accurate results with an absolute error about half of that of the cross-sectional flux method and a absolute relative error of $\sim 20\%$ when only the XCO_2 and wind fields are used as inputs. Furthermore, we show that our estimations are only slightly affected by the absence of NO_2 (nitrogen dioxide) fields or a detection mechanism as additional information. Finally, interpretability techniques applied to our models confirm that the CNN automatically learns to identify the XCO_2 plume and to assess emissions from the plume concentrations. These promising results suggest a high potential of CNNs in estimating local CO_2 emissions from satellite images.

1 Introduction

The burning of fossil fuels, such as coal and oil, in power plants (PP), is a primary source of anthropogenic CO_2 emissions. Approximately 50% of worldwide fossil fuel CO_2 emissions originate from large facilities, which encompass PPs (IEA, 2019; Nassar et al., 2022). As a result, maintaining regular monitoring of these emissions and possessing the capacity to control their reporting is crucial.

Observations from satellites like OCO-2 provide valuable data that can be utilised to estimate CO_2 emissions (Nassar et al., 2017; Reuter et al., 2019; Chevallier et al., 2019; Wu et al., 2020; Zheng et al., 2020; Nassar et al., 2022; Chevallier et al., 2022). Specifically, satellite observations of CO_2 plumes, such as the plume transects obtained from OCO2-2 and OCO-3 satellites, offer a direct means of quantifying their source and complement other estimations (Cusworth et al., 2021). The upcoming launch of the CO_2M satellites in 2026 is anticipated to capture high-resolution images of CO_2 column-averaged mole fractions (XCO_2), aiming to have a much larger swath and more accurate CO2 estimations, further advancing our capabilities in this area. Leveraging these images, however, will present significant challenges (Wang et al., 2020).

CO_2 plumes are notoriously difficult to invert due to various factors, including 1) image integrity issues caused by cloud cover or satellite overpasses, which result in missing data in the images used for analysis. Additionally, the estimation of the emissions associated with a plume is further complicated by 2) the measurements low signal-to-noise (SNR) ratio. The noise component encompasses variations in the background as well as errors in the satellite measurements. The SNR problem stands as the main hurdle in the detection of the plume, a crucial step for inversion. Recent research conducted by Dumont Le Brazidec et al. (2022) has illustrated the remarkable ability of Convolutional Neural Networks (CNNs) to effectively overcome this obstacle. Lastly, another challenge stems from 3) the uncertainties in the transport and dispersion processes, specifically, when it comes to estimating the effective wind driving the plume and determining its shape (Kuhlmann et al., 2019).

This paper addresses the second and third problems (except for the errors in the satellite measurements) by employing deep learning techniques to perform inverse modelling of CO_2 plumes. In particular, we focus on developing techniques for inverting CO_2 plumes from PPs of different emission levels.

To assert the effectiveness of the method, the predictions of the deep learning model are compared against state-of-the-art techniques. Plume inversion methods include approaches that use an atmospheric transport model to simulate the plume and compare it to observation (e.g., Pillai et al. (2016); Broquet et al. (2018)). They also include techniques that quantify emissions from a hotspot based on plume detection in satellite observations (Koene et al., 2021). These methods can be based on time-averaged plumes, such as the divergence method (Beirle et al., 2019; Hakkarainen et al., 2022), or on instantaneous images. Varon et al. (2018) compared several of these approaches, namely the Gaussian plume inversion, the integrated mass enhancement and the cross-sectional flux method. In the CoCO2 project, several of these methods were compared using synthetic CO2M CO_2 and NO_2 observations (Hakkarainen et al., 2023). Here, we use the cross-sectional flux (CSF) method for comparison, which showed similar accuracy as other well performing methods such as Gaussian plume inversion and the light cross-sectional flux (LCSF) method. Finally, it should be noted that the CoCO2 project identified several potential improvements of the methods that may yield superior performance in the future (Hakkarainen et al., 2023; Santaren et al., 2024).

In this paper, the proposed plume inversion approach is based on convolutional neural networks. This research builds on earlier work in the field of remote sensing image analysis leveraging machine learning techniques (Lary et al., 2016; Finch et al., 2021; Jongaramrungruang et al., 2021; Joyce et al., 2023; Kumar et al., 2023). Here, plume inversion involves the analysis of an image to extract scalar or vector emissions data at different time steps. Therefore, this task can be framed as an image regression problem, where relying on CNNs can offer significant advantages (Chollet, 2017). CNNs, a type of supervised learning method, can be trained on a comprehensive dataset, where all input variables (images) and associated output variables (like emissions) are known. Once trained, these CNNs can effectively process and draw conclusions from unseen observational imagery. CNNs employ convolutional layers to extract essential features from images. Each filter is automatically trained to detect specific patterns, such as edges, corners, or other shapes within the image. By stacking multiple convolutional layers, CNNs become capable of learning intricate patterns, enabling them to capture increasingly complex features. The ability of CNNs to capture and learn spatial features in images makes them a popular choice for various image-related tasks, including image recognition, classification, and regression. Given the nature of our plume inversion task, they are particularly well-suited due to their ability to identify spatial features in images, such as plume shapes or intensity, that correspond to specific emissions. This feature extraction approach effectively harnesses the knowledge embedded in transport models, enabling this automatic capture of plumes dynamics. This ability to capture such features has already been demonstrated by Finch et al. (2021), or Dumont Le Brazidec et al. (2022) which study the segmentation of plumes in CO_2 images.

To train and test the CNN models, this paper relies on a synthetic dataset as CO_2M data will not be available until 2026. This dataset has been designed to possess similar key features as the forthcoming CO_2M satellite, such as resolution and the availability of NO_2 data. However, the influence of clouds or systematic error patterns is not considered in the analysis.

Before introducing the inversion methodology and results, we briefly describe the physical fields used to train and evaluate the CNNs. This includes the presentation of the simulated satellite fields in section 2 and of the model used to produce segmentation masks of the plumes used as inputs of the inversion model in section 3. The inversion methodology using CNN is described in section 4, specifying the problem statement, the model, the training process, and the alternative method employed for comparison. The subsequent section 5 delves into the application of the model to three specific PPs. In particular, subsection 5.3 places special emphasis on the interpretability of the trained CNNs. Discussions on the limitations and future directions of this study can be found in section 6, while the conclusions are outlined in section 7.

2 Dataset of XCO$_2$, winds and NO$_2$ images

XCO_2 and NO_2 images are taken from the concentration fields simulated for the SMARTCARB project (Brunner et al., 2019). Using the COSMO-GHG model, the SMARTCARB simulations were performed in a region centred around Berlin and covering several nearby coal-fired PPs. These simulations were used to produce synthetic observations of CO_2M and to evaluate various plume detection and inversion approaches (Kuhlmann et al., 2019, 2020, 2021; Hakkarainen et al., 2021, 2023). The data are hourly and cover an entire year. Their spatial resolution is $0.01°$ and sixty vertical layers ranging spanning from an
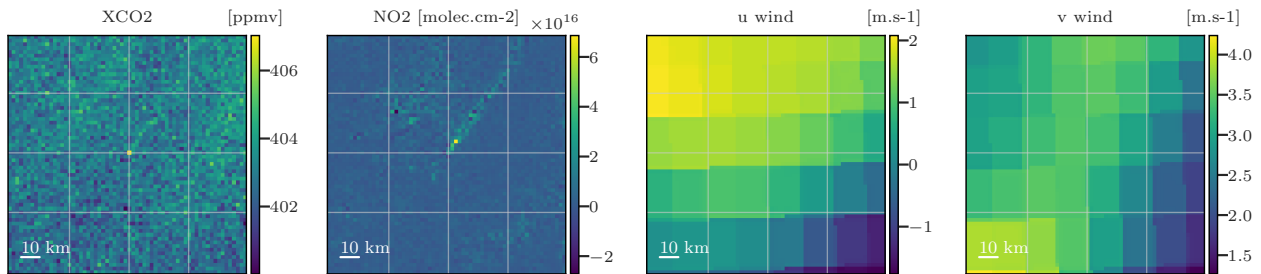
Figure 1. Examples of inputs used by the CNN model. The first, second, third and fourth columns represent the XCO_2 images, NO_2 images, vertically averaged u and v winds, respectively.

altitude of 0 to $24\,km$ are used. More information can be found in Dumont Le Brazidec et al. (2022), which presented more extensively a dataset very similar to the one used in this study.

Images used to train and evaluate the CNN inversion models consist of 64×64 pixels, with each pixel covering an area of $2 \times 2\,km^2$. Each image is extracted from the SMARTCARB COSMO-GHG simulated fields so that one hotspot is located in the centre. In addition, the selected size guarantees the inclusion of the majority of the central hotspot plume within the image. The mapping from the original SMARTCARB fields resolution to the $2\,km$ resolution is performed by cubic spline interpolation (Virtanen et al., 2020). The $2\,km$ resolution was chosen to be consistent with the resolution expected for CO_2M observations.

It is necessary to take into account the expected noise associated with the satellite instruments. For this, a Gaussian random noise of standard deviation $0.7\,ppm$, characteristic of the CO_2M (Meijer, 2020), is added to the XCO_2 images.

In addition to XCO_2, ancillary data can be used to assist in the inversion of XCO_2 plumes. Considering that CO_2M will provide measurements of NO_2 and the observed strong correlation between NO_2 and CO_2 plumes, noisy NO_2 fields are used in this study. The noise associated with a NO_2 field is implemented as the standard normal distribution multiplied by the NO_2 field values. The median standard value for NO_2 fields surpasses $1e15\,molec.cm^{-2}$, leading to an average noise level in the NO_2 field that exceeds the CO_2M NO_2 requirement (less than $1e15\,molec.cm^{-2}$). Furthermore, ERA5 winds are used: the original resolution of $28\,km$ is mapped to $2\,km$ to be consistent with the CO_2 and NO_2 images. To overcome the circular data limitation of statistical methods, the u and v wind fields are used instead of the direction and magnitude components. This limitation corresponds to the statistical model's inability to correctly interpret wind directions where the value of 360 degrees is equivalent to 0 degrees. More precisely, we use 2D u and v wind fields which are calculated as the average of the zonal and meridional wind fields over the 37 lower ERA5 vertical levels, respectively, which corresponds roughly to the lowest $4000\,km$ of the atmosphere. Figure 1 presents a series of potential inputs to the CNN.

This paper only addresses the retrieval of PP emissions, although the training dataset includes the city of Berlin. More precisely, depending on the PP evaluated, the training dataset might be composed of any hotspot in {Berlin, Jänschwalde, Schwarze Pumpe, Boxberg, Turow, Patnow, Lippendorf, Opole, Dolna Odra}. The primary rationale behind prioritising the

Hotspot	min	max	mean	std
Berlin	4.8	34.7	16.8	7.2
Jänschwalde	16.4	52.7	33.3	7.7
Boxberg	9.4	30.1	19.0	4.4
Lippendorf	7.5	24.1	15.2	3.5
Turow	4.3	13.8	8.7	2.0
Schwarze Pumpe	4.0	13.0	8.2	1.9
Dolna Odra	3.7	12.5	7.9	1.9
Opole	3.5	11.8	7.5	1.8
Patnow	2.9	9.2	5.8	1.3

Table 1. Emission statistics for the considered PPs and the city of Berlin. Fluxes are in Mt CO_2 /yr.

training of the model on PPs is the scarcity of cities in the dataset, which poses a challenge for the model to effectively learn and generalise for cities. However, Berlin is included in the training dataset as supplementary data to aid the model in its
115 learning process. In the SMARTCARB dataset, the modelling of anthropogenic emissions, incorporating fixed diurnal, weekly, and seasonal cycles, was performed using the TNO-MACC III inventory (Kuenen et al., 2014). The emissions range, mean, and standard deviation of each hotspot are given in Table 1 and locations of considered PPs and Berlin are described in 2. Moreover, data augmentation techniques are employed to expand the database, as detailed in section 4.2.1.

3 Application of the segmentation model

120 Utilising a segmentation algorithm to incorporate plume contours as additional prior information in plume inversion may yield significant benefits. In this section, we provide a brief description and application of the CNN-based method developed in Dumont Le Brazidec et al. (2022) that predicts plume contours in XCO_2 images. The methodology of Dumont Le Brazidec et al. (2022) involves employing an image-to-image U-net model, which generates images that are subsequently used as inputs for the CNN inversion model, as outlined in section 4.

125 Apart from a few specific points, the training and model choices are similar to those of Dumont Le Brazidec et al. (2022). A simpler encoder, with fewer neurons, is chosen since the NO_2 fields are used as inputs to the CNN. This simplification of the problem reduces the need for a complex encoder. In addition to NO_2 and XCO_2 fields, winds are also used to assist in the XCO_2 plume contour prediction, although experiments show that the addition of this data has very little influence on the predictions. Finally, the U-net models were designed to make predictions beyond the geographical region of their training data.
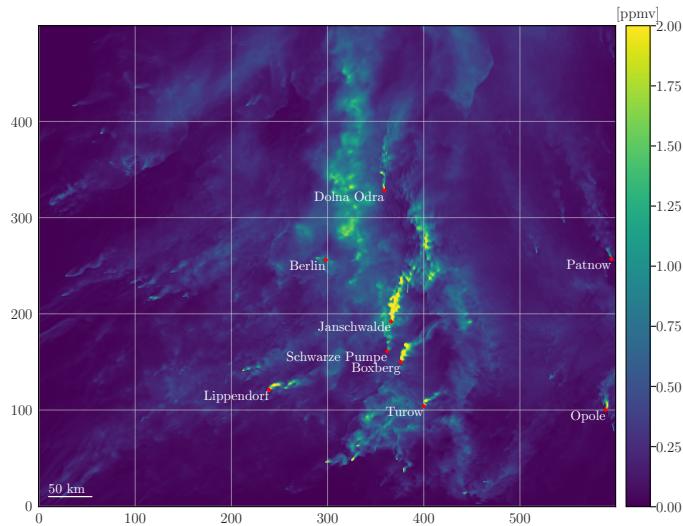
Figure 2. XCO_2 concentration map with the locations of Berlin and each considered PP within the complete SMARTCARB domain. The map consists only of the concentrations stemming from the major anthropogenic sources. Furthermore, to enhance plume visibility, as fluxes of power plants such as Jänschwalde are vastly superior to other fluxes, concentrations exceeding 2 ppmv have been capped at 2 ppmv.

130 Specifically, the model that learns to predict the mask of the Boxberg PP plume from an image centred at Boxberg PP is trained on a dataset excluding the images centred at Boxberg.

In Fig. 3, we show the application of a model trained to predict the positions of Turow plumes. It was trained on pairs of fields in the regions of {Boxberg, Berlin, Lippendorf, Patnow, Jänschwalde, Dolna Odra, Schwarze Pumpe, Opole}. The first and second columns of the figure show the XCO_2 and NO_2 fields, inputs to the CNN. The third column shows the target plume

135 as a reference point, while the fourth column shows the output of the CNN.

4 Deep learning method for the inversion of XCO_2

4.1 Inversion based on supervised learning

The inverse problem addressed here is the estimation of the CO_2 emissions accountable for the central hotspot plume observed in a given XCO_2 field image. To do so, a CNN is used, processing as input a given XCO_2 field and other additional fields

140 and resulting in a scalar output representing the emission rate of CO_2 in $MtCO_2.yr^{-1}$ at the hour corresponding to the image. This flux unique scalar representation choice was made for the sake of simplification, as the quality of the results is minimally impacted by the choice between targeting average, instantaneous fluxes, or a vector of instantaneous fluxes over the last N hours. This is due to the relatively slow hourly variation in the CO_2 emission rate of the PP. All future flux quantities are
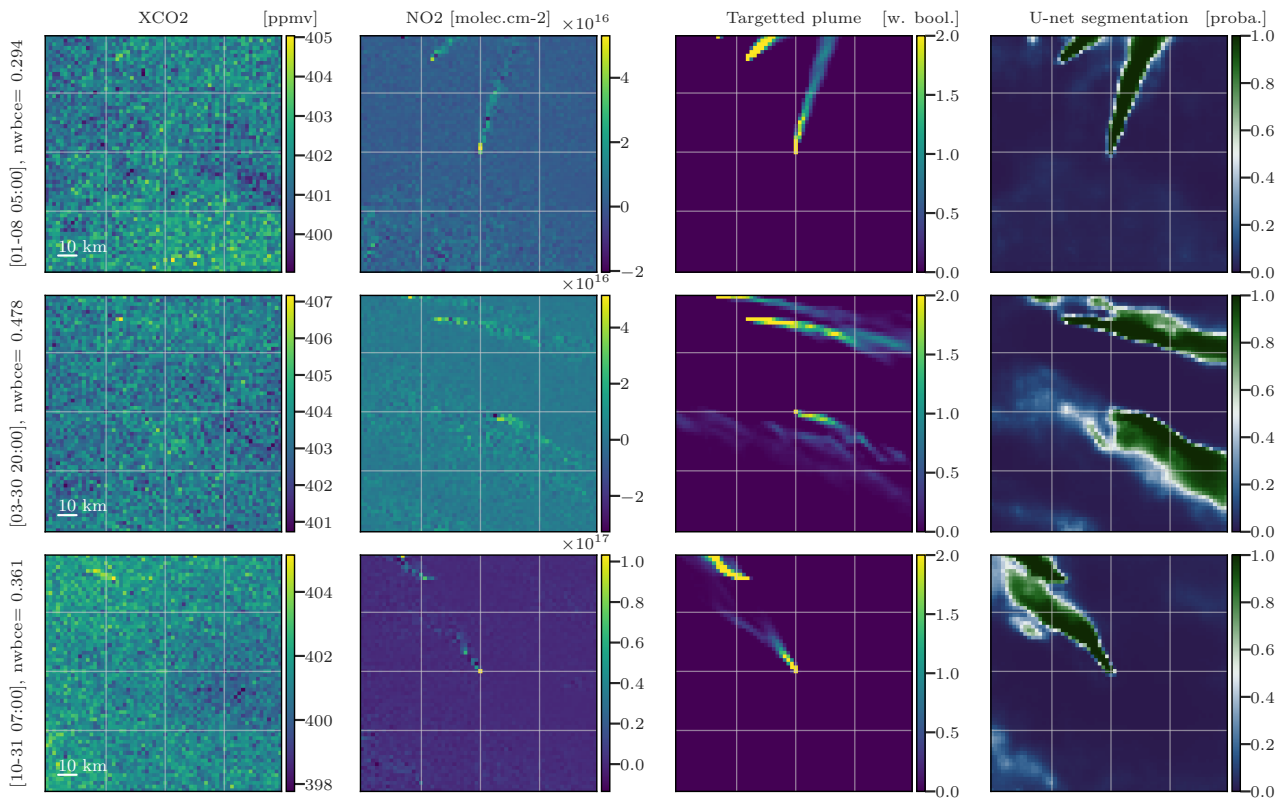
Figure 3. Examples of the U-net CNN model application on images from the test dataset centred at Turow. The first, second, third and fourth columns represent the XCO_2 images, NO_2 images, weighted Boolean plumes and U-net predictions as probability maps, respectively. All times are in UTC.

expressed in $Mt.yr^{-1}$. The image-to-scalar, or image regression, problem is depicted in Fig. 4. The CNN is trained using XCO_2 fields, ancillary data fields (winds, segmentation results and NO_2) and associated emission fluxes ranging from 3 to $53\,Mt.yr^{-1}$ across various times and targets.

4.2 CNN model and training parametrisation

In this section, we present and discuss the architecture of the model, the hyperparameters, and the learning methodology for the inversion task. In particular, the model is built from preprocessing layers and a core model. The preprocessing layers are used to augment/transform, construct, add noise, and normalise the input data before feeding it into the core model. The core model is designed to extract features from these transformed input data.
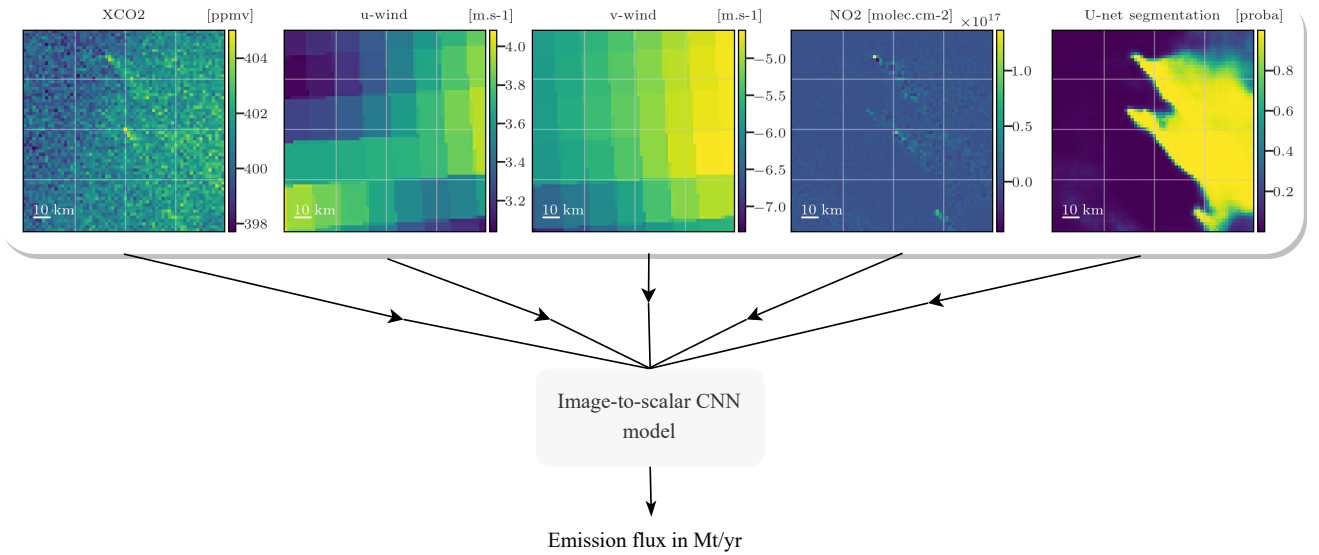
Figure 4. Various fields: XCO_2, winds, segmentation results and NO_2, and the emissions are used by a CNN that learns to estimate the emissions associated to the central plume concealed under the background. Note that the combination of NO_2 and segmentation results as inputs is never assessed in this manuscript.

4.2.1 Description of the preprocessing layers

The preprocessing layers consist of a six steps sequence, as presented in Fig. 5. The purpose of steps 1-2-3 is to extend the initial database, thereby enhancing the model's ability to generalise to unseen data. A pair of input/output data for training is constituted in this way:

1. a target CO_2 plume corresponding to a PP at a time t is chosen randomly. A background must be constructed to form the XCO_2 image with this plume. This background is chosen randomly and therefore does not necessarily correspond (geographically and temporally) to the chosen target plume. It may correspond to another PP as well as to another time. Furthermore, in SMARTCARB, the background is partitioned into multiple segments. In our case, the background is constructed from two randomly and independently drawn fields: a field containing the major part of the fluxes including the biogenic fluxes and a field containing a part of the anthropogenic fluxes of the SMARTCARB domain. Finally, potential additional fields (wind, NO_2, segmentations) corresponding geographically and temporally to the chosen CO_2 plume are selected;

2. then,

   – the target XCO_2 plume is multiplied by a random uniform scaling factor $p \sim U(0.25, 2)$. The corresponding true emissions are also multiplied: $y_{\text{truth}}^{\text{scaled plume}} = p \times y_{\text{truth}}^{\text{plume}}$;

- a random number $b \sim U(-3.5, 3.5)$ (in ppmv) is added uniformly to the main background. Extrema of this uniform distribution are chosen as approximately the standard deviation of an average XCO_2 background;

- a random uniform scaling factor $a \sim U(0.33, 3)$ is applied to multiply the field containing a part of the alternate anthropogenic fluxes.

Anthropogenic fluxes scaling factors p and a are chosen so that the resultant emissions still correspond to reasonable anthropogenic fluxes;

3. the XCO_2 field is built as the sum of the target plume, background, and alternate anthropogenic fluxes field components;

4. a Gaussian noise matrix of shape 64×64 (equal to the image shape) $\mathbf{G}_{Nx \times Ny} \sim N(0, 0.7)$ (in ppmv) is added to the XCO_2 field to simulate the satellite observational noise;

5. the noisy XCO_2 field is concatenated with the additional fields. If added, the NO_2 field is noised beforehand;

6. standardisation (i.e., Z-score normalisation) is performed independently on each channel (each physical field) of the concatenated input data.

These steps are carried out exclusively during the model training phase. The different operations in this process are performed to create a more robust and diverse training dataset. To ensure an accurate assessment of the performance of the trained model, the test dataset used for the evaluation consists only of pre-constructed, physically consistent simulated data. Specifically, no scaling factors are applied and the XCO_2 fields used for testing are always constructed from geographically and temporally consistent plume and background components.

4.2.2 Description of the core model

The chosen core CNN model, described in Fig. 6 is designed for image regression. It was chosen by comparing its performance with state-of-the-art models such as EfficientNet or Squeezenet (Tan and Le, 2020). These two models are deep neural networks designed primarily for image classification tasks. They incorporate modified versions of CNN, including features such as residual connections and depth-separable convolutions, in order to improve efficiency, speed and ease of implementation. As their initial implementation tended to overfit, we considered a smaller version with a reduced number of neurons in each layer. But even after tuning, the simpler model depicted below outperformed these more advanced models.

The chosen model takes 3 to 4 images of 64×64 pixels as input (which correspond to the XCO_2 field and ancillary data such as the winds). It is constructed as a succession of convolutional, max pooling, batch normalisation, and dropout layers where:

- convolutional layers aim to identify and extract relevant features by applying a set of learnable filters to the previous feature map. The 2D convolutional operations are applied with a filter size of 32 and a kernel size of 3×3;

- max pooling layers play a key role in reducing the resolution of feature maps while retaining the essential information. This reduces the computational complexity of the network and leads to the extraction of more complex features;
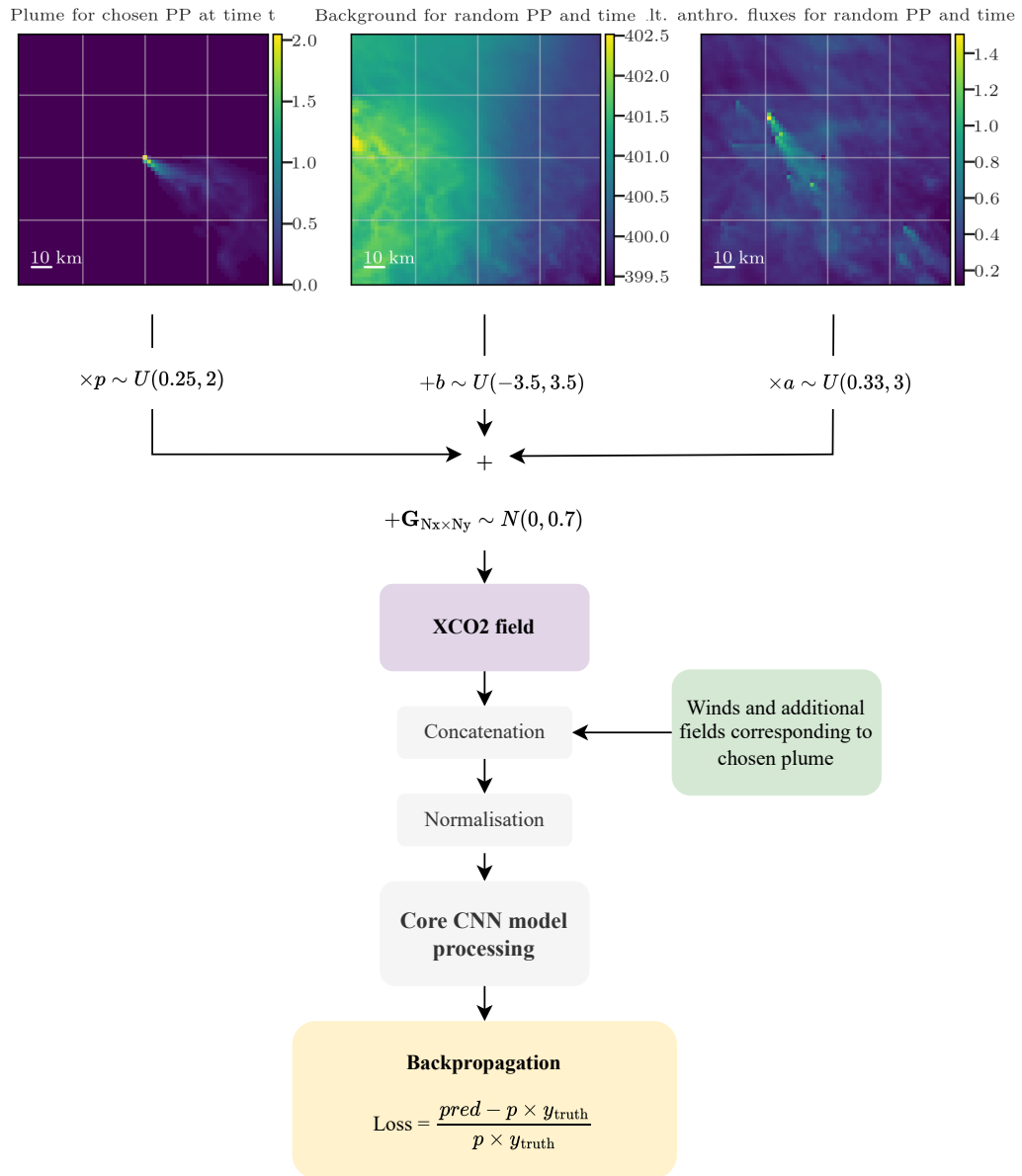
Figure 5. Description of the preprocessing layers as a sequence of six steps: 1) random choice of the XCO$_2$ field components, 2) scaling transformation of the components, 3) sum of the components, 4) satellite noise simulation, 5) concatenation with the additional data, and 6) normalisation.
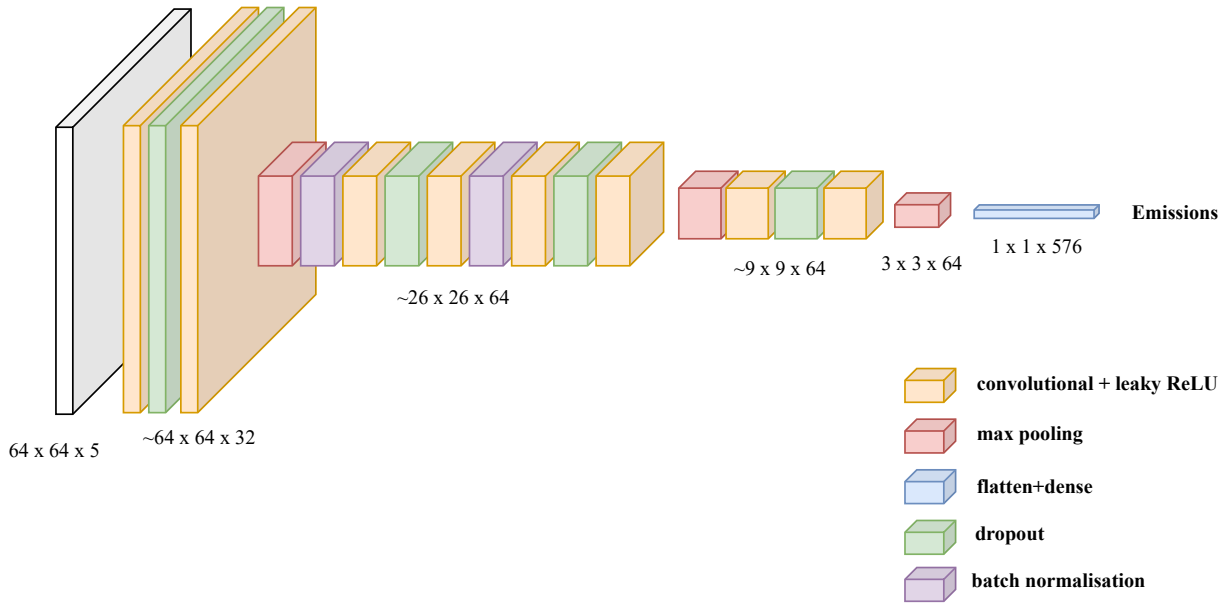
Figure 6. Description of the core CNN model, designed to extract features from the input data. The resolution given for each sequence of layers is approximate, as convolutional layers also slightly reduce the resolution.

- batch normalisation layers are used to improve the stability of the network and speed up its learning process;

- dropout layers randomly exclude a certain percentage of the neurons in the previous layer at each iteration to reduce overfitting. They are only activated during the training phase.

The output of the core model is flattened and fed into a fully connected final dense layer with a single output unit, which is activated by a leaking rectified linear unit (ReLU). In total, the CNN has approximately $\sim 186,000$ parameters.

4.2.3 Training parametrisation

The training hyperparameters, such as the optimiser, learning rate and batch size, have been determined through a combination of experimental investigation and adherence to standard practice. In accordance with customary practices, Adam's optimiser was employed with a fixed learning rate of 10^{-3} and a dropout rate of 0.2 was applied. The batch size, or number of samples before the model updates its weights, was set to 32. After analysis, to ensure model convergence, the epoch count was set to 500, which yields a total training time of approximately four hours using an Nvidia Quadro RTX 5000 16GB GPU. Furthermore, the default choice of loss function employed is the Mean Absolute Percentage Error (MAPE), which emphasises proportionate deviations between emissions predictions and ground truth. This loss function is equivalent to a absolute relative error between the truth and the predictions. It should be noted that the Mean Absolute Error (MAE) metric was used to train roughly half of the ensemble models (detailed in the subsequent section 4.2.4) in the Lippendorf case (5.1.1). MAE and MAPE metrics generally yielded similar performances.

4.2.4 Model ensembling

In many cases, deep learning models suffer from high levels of volatility due to their reliance on random initialisations of parameters and hyperparameter optimisation algorithms. To overcome these limitations and increase the stability of our predictions, we employed the technique of model ensembling. Specifically, for each considered configuration, we trained multiple instances of the same model architecture and parameters. This approach enabled us to produce a range of predictions that were subsequently averaged into a single estimate. To ensure sufficient convergence of these estimates, a minimum of five individual models were used for each considered configuration. Through these steps, we significantly reduced the variability observed across multiple models runs, thereby improving the accuracy and confidence in our system.

4.3 Geographical separation between the training and test datasets

We consider a strict geographical separation between the PPs of the training/validation and those of the test dataset. For example, to train a model to predict emissions from Boxberg plumes, we consider a training dataset consisting of images centred at all other available PPs except Boxberg. In this way, the model predicting Boxberg emissions never had access to the Boxberg emissions inventory.

During the training phase, to avoid overfitting or underfitting, a separate validation dataset is used to follow the model's performance on new data. In our case, the validation dataset is chosen as unseen data from the same geographical area as the training dataset.

Three distinct models are created and trained to predict emissions from three specific PPs. The target PPs are Lippendorf, Boxberg and Turow and are selected:

- to obtain a diverse range of emission rates: between 7 and $24\,\mathrm{Mt.yr}^{-1}$, with a mean of $15\,\mathrm{Mt.yr}^{-1}$ and a std of $3\,\mathrm{Mt.yr}^{-1}$ for Lippendorf, between 4 and $13\,\mathrm{Mt.yr}^{-1}$, with a mean of $9\,\mathrm{Mt.yr}^{-1}$ and a std of $2\,\mathrm{Mt.yr}^{-1}$ for Turow, and between 9 to $30\,\mathrm{Mt.yr}^{-1}$ with a mean of $19\,\mathrm{Mt.yr}^{-1}$ and a std of $4.\,\mathrm{Mt.yr}^{-1}$ for Boxberg;

- to account for the potential presence of multiple plumes in the same image: the images centred at Boxberg also include plumes from Jänschwalde, Schwarze Pumpe, and Turow which sometimes overlap;

- because their emission rates are neither at the highest nor the lowest extremes (Boxberg was selected over Jänschwalde because Jänschwalde has the highest emission levels in our dataset, potentially placing it outside the training distribution);

- due to their positioning away from the boundaries of the SMARTCARB domain. These PPs are less affected by border conditions. As shown in Fig. 1, Patnow or Opole are close to the borders and are influenced by border conditions.

Although these models differ in terms of the training datasets used, they all share the same architectural framework, including hyperparameters, CNN structure, and preprocessing layers. The objective behind creating these models is to demonstrate the effectiveness of the architectural framework, laying the groundwork for a universal model based on this architecture and capable

of generalising across future PPs. It is important to note that while the test dataset from one experiment appears in the training
dataset of another, each experiment was conducted independently. The model tuning was not influenced by the results obtained
with the test datasets. In particular, the selection of the hyperparameters such as the learning rate was made before the model
training.

4.4 Alternative method for comparison

The accuracy of our inversion method is compared to the cross-sectional flux (CSF) method implemented in the Python library
for data-driven emission quantification (ddeq[1]). The CSF method was recently compared with other state-of-the-art methods
and shows similar accuracy than other well performing methods (Hakkarainen et al., 2023). This method consists in

- detecting the plume and extracting it from the background. NO_2 fields are used to help in the detection;

- dividing the plume into a series of horizontal slices of known areas and heights;

- estimating the line densities of CO_2 by fitting Gaussian curves to the CO_2 and NO_2 concentrations within each slice;

- inferring the CO_2 fluxes as the product of the line densities and the wind speed at the sources;

- deriving the total emission rate by multiplying the flux estimation and the area of each slice, and then averaging all
  downstream fluxes.

The CSF method is limited by the need for accurate estimation of the effective wind (Kuhlmann et al., 2020). Two separate
wind estimates have been considered: the first derived from an average of the 37 lower levels of ERA 5 data, and the second
corresponding to the wind at 100 meters. The first estimate was chosen because of its superior performance.

5 Application: inversion of three power plants and model interpretation

In this section, we study the performance of a trained CNN under the conditions exposed in section 4:

- firstly, various ensembles of models with different sets of inputs are evaluated on the Lippendorf PP in section 5.1.1, the
  Turow PP in section 5.1.2 and the Boxberg PP in section 5.1.3. In each configuration, the training, validation, and test
  datasets involve 25152, 4608, and 6289 images, respectively;

- secondly, we investigate how the assimilation of segmentation fields or NO_2 affects the CNNs. Afterwards, since over-
  fitting arises in certain configurations, discussions and partial solutions to this issue are proposed;

- thirdly, we propose to interpret the CNNs using a gradient-based technique and by permuting the input features.

---

[1]https://gitlab.com/empa503/remote-sensing/ddeq

5.1 Inversion of plumes performance

We study the performance of various CNN model ensembles in predicting the emissions of the Lippendorf, Turow, or Boxberg PPs. More precisely:

- three model ensembles are trained on {Berlin, Jänschwalde, Schwarze Pumpe, Boxberg, Turow, Patnow, Opole, Dolna Odra} and tested on Lippendorf;

- three model ensembles are trained on {Berlin, Jänschwalde, Schwarze Pumpe, Boxberg, Lippendorf, Patnow, Opole, Dolna Odra} and tested on Turow;

- three model ensembles are trained on {Berlin, Jänschwalde, Schwarze Pumpe, Turow, Lippendorf, Patnow, Opole, Dolna Odra} and tested on Boxberg.

Lippendorf typical emissions fall between those of low-emission PPs (such as Dolna Odra or Turow) and those of high-emission (Boxberg, Jänschwalde). Turow emissions range between 4 and $14\,\mathrm{Mt.yr}^{-1}$, similar to PPs like Opole. This implies that most Turow CO_2 plumes are hidden under the background. The last studied PP is Boxberg which is characterised by high emissions and the presence of other PPs in the vicinity, which entails the presence of other high SNR plumes.

For each of the three target PPs, three model ensembles are trained. The input data for the first model ensemble are the XCO_2 field and the wind fields u and v. The second and third model ensembles use the same three base fields and the output of the segmentation model, or the NO_2 field, respectively, as fourth input. In the following, for simplicity, these ensembles of CNN models will be referred to simply as "models".

Kernel Density Estimation (KDE) plots over the 6289 images of the test dataset are drawn in Fig. 7 comparing with relative and absolute metrics the true emission rates and the predictions of the four models for each PP. While Fig. 7 employs signed relative error to assess prediction bias in the CNNs, Table 2 and the majority of the analysis rely on absolute relative error. KDE is a non-parametric statistical technique that estimates the probability density function of a continuous random variable by smoothing its observed data points using a kernel function. The three first ensembles of predictions consist of those of the trained CNNs, and the fourth corresponds to the CSF method application. The main statistics corresponding to these KDE plots are summarised in Table 2.

5.1.1 Lippendorf plumes inversion

As reported in Table 2, the utilisation of the CNN approach yields remarkably accurate predictions of Lippendorf emissions compared to the CSF method. The performance of all three CNN models shows a median absolute relative error of approximately 20% and a median absolute error of around $3\,\mathrm{Mt.yr}^{-1}$ (the average emissions for Lippendorf are $15.2\,\mathrm{Mt.yr}^{-1}$). In comparison, the CSF method exhibits a higher median absolute relative error performance of around 40%, and the absolute error performance is approximately double, at $6\,\mathrm{Mt.yr}^{-1}$. The CNN results are reliable, as the majority of errors are concentrated below $10\,\mathrm{Mt.yr}^{-1}$ or 50%, with very few exceeding 100%. This indicates that the models provide trustworthy estimates, with a relatively small margin of error.
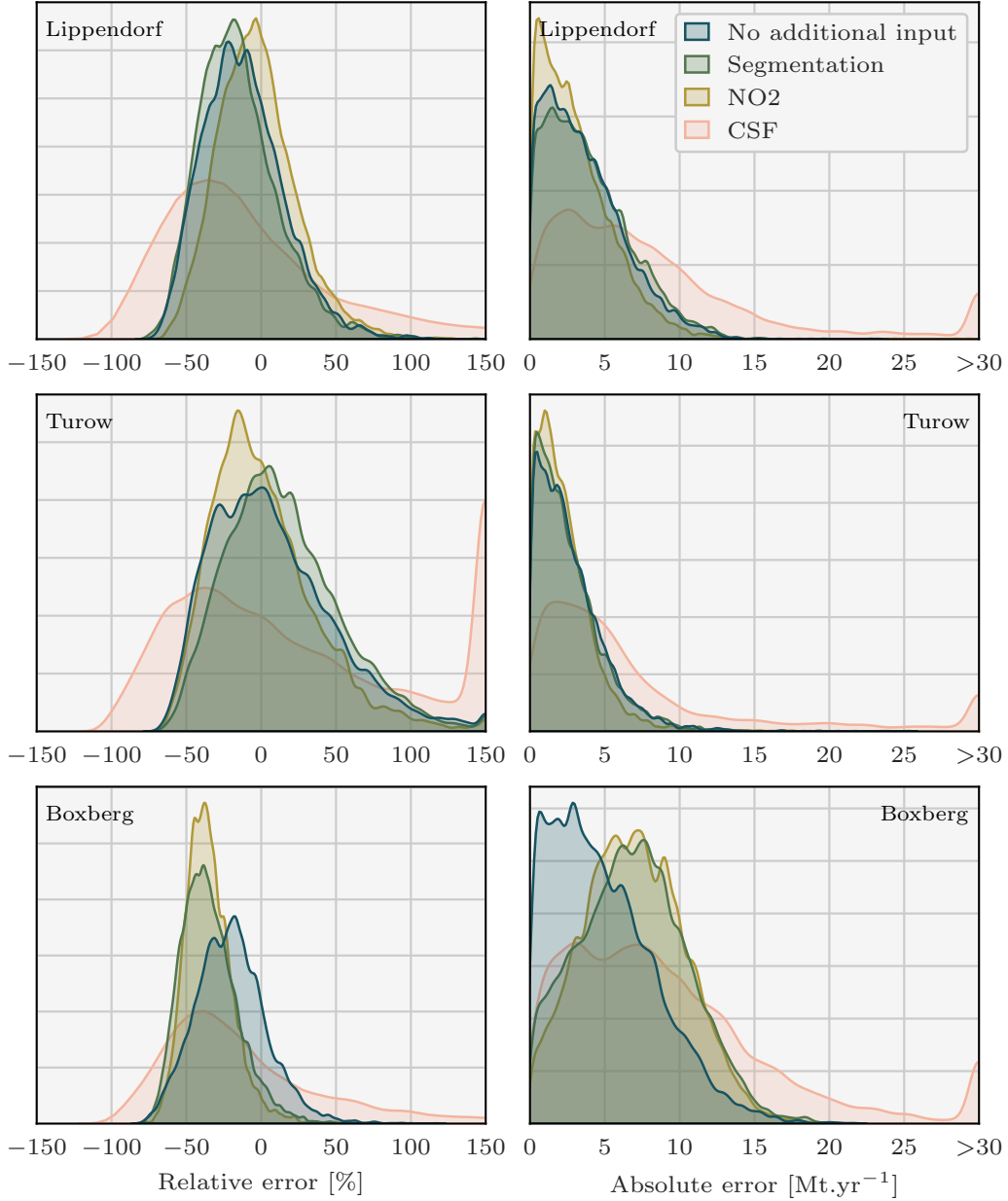
Figure 7. Density plots of the signed relative and absolute error between the predicted and the true Lippendorf emissions. Four sets of predictions are considered, corresponding to the three CNN models with three different sets of inputs and the CSF method. Each CNN model is trained with the XCO_2 field and the winds as inputs. Two of the models additionally assimilate the NO_2 field or the predictions of the segmentation model. Predictions with relative errors greater (in absolute) than 150% or absolute errors greater than $30\,Mt.yr^{-1}$ were set to ±150 or 30 to increase visibility. 3% of the CSF method predictions are missing. Those predictions correspond to Lippendorf plumes superimposed on other plumes, where the CSF method can not be applied.

	Absolute relative error [%]			Absolute error [Mt.yr^{-1}]		
	25%	Median	75%	25%	Median	75%
Lippendorf						
CNN with no additional input	9.8	21.3	35.6	1.4	3.1	5.2
CNN with segmentation	11.4	23.3	36.5	1.6	3.4	5.6
CNN with NO$_2$	8.6	18.1	30.3	1.3	2.7	4.5
CSF	21.0	42.8	70.1	3.1	6.3	10.5
Turow						
CNN with no additional input	12.3	25.9	43.2	1.0	2.2	3.8
CNN with segmentation	11.0	23.7	41.3	0.9	2.0	3.5
CNN with NO$_2$	10.9	22.9	38.0	0.9	1.9	3.2
CSF	26.3	52.1	92.0	2.2	4.5	8.1
Boxberg						
CNN with no additional input	11.7	23.5	37.2	2.1	4.4	7.1
CNN with segmentation	24.2	36.9	48.2	4.2	6.8	9.4
CNN with NO$_2$	26.4	36.9	45.4	4.5	6.8	9.3
CSF	21.7	41.5	63.5	3.9	7.7	12.3

Table 2. Absolute relative and absolute error statistics between predicted and true emissions for CNN models with various inputs and applied to various target PPs and the CSF method.

5.1.2 Turow plumes inversion

The CNN approach produces highly accurate predictions for the low SNR Turow plumes, exhibiting similar performance to the results obtained in the Lippendorf case. The three models yield a median absolute relative error performance of approximately 25% and a median absolute error performance of around $2\,\mathrm{Mt.yr}^{-1}$. The results can be considered reliable: 75% of the results fall below a threshold of $4\,\mathrm{Mt.yr}^{-1}$. By contrast, the CSF method exhibits a median absolute relative error performance of approximately $50-55\%$ and an absolute error more than two times larger. The inclusion of segmentation or NO_2 fields has a noticeable, albeit not significant, impact on the model's performance, resulting in an improvement on the order of a tenth of a $\mathrm{Mt.yr}^{-1}$. Notably, the addition of the NO_2 field appears to have a slightly greater impact compared to the inclusion of segmentation fields. This implies that when applying the model that assimilates the segmented XCO_2 fields (with the assistance of the NO_2 fields), there is a potential loss of information on the NO_2 fields compared to using the model directly assimilating the NO_2 fields. The observed phenomenon could be a result of potential overfitting when employing the segmentation model, as it is trained on the same dataset as the inversion model. Consequently, the segmentation predictions on the training dataset are typically superior to those on the test dataset. This discrepancy can lead to an over-reliance on the segmentation fields, subsequently causing overfitting.

5.1.3 Boxberg plumes inversion

In contrast to the inversion results obtained for Lippendorf and Turow PPs, there is a significant variation in Boxberg plume inversions performance. A CNN only trained with XCO_2, u and v winds demonstrates strong performance, comparable to those of the models estimating Lippendorf and Turow emissions. The median absolute relative error performance is approximately $20-25\%$, and the median absolute error performance is around $4\,\mathrm{Mt.yr}^{-1}$. The CSF method exhibits a median absolute relative error performance of around 40% and an absolute error performance close to $8\,\mathrm{Mt.yr}^{-1}$. But both models with segmentation or NO_2 fields show a significant decline in performance, which contradicts our expectations. This phenomenon can be attributed to overfitting and is examined in detail in the section 5.2.2.

5.1.4 Cross-sectional fluxes method results

The CSF results align with the findings reported in (Kuhlmann et al., 2021) and (Hakkarainen et al., 2023). Other methods, such as the light cross sectional fluxes technique, might yield predictions with absolute relative errors reduced by 5 to 10% (Hakkarainen et al., 2023). A notable constraint of the CSF method is that it hinges on the estimation of the effective wind speed inside the plume. Under ideal circumstances, the effective wind speed should be estimated using the wind profile, weighted by the CO_2 concentration profile, but in practical applications, estimating this profile presents a substantial challenge. In a separate experiment, the CSF method was utilised on a reduced dataset (compared with that used to study the CNN results). Among other changes, this dataset considers only the CO_2M overpass, or excludes the plume overlaps processed by the CNN. In this experiment, the exact emission profile used to simulate the COSMO-GHG fields of SMARTCARB was used to compute the

effective wind. This methodology resulted in a median absolute error around $4\,\mathrm{Mt.yr}^{-1}$, thereby indicating that the effective wind estimation significantly contributes to the errors associated with the CSF method.

5.2 Results analysis

335 This section presents an analysis of the results from the preceding section 5.1. It involves examining the trained models, or new models trained under the same configuration, across various datasets, including the test dataset. It is important to note that these analyses were conducted after obtaining the previous section's results and thus did not influence, e.g., the choice of model architecture or hyperparameters.

5.2.1 Study of the addition of the segmentation and NO_2 fields

340 Interestingly, the model designed to invert the Lippendorf plumes does not yield more accurate emission estimates when using segmentations as inputs, compared to the model without segmentations. One possible reason for this lack of improvement is that the Lippendorf segmentation fields are inadequate. The segmentation model assimilates NO_2 fields to segment the CO_2 fields and the presence of multiple alternative NO_2 plumes in the NO_2 fields hinders the segmentation model's ability to accurately delineate the contour of the Lippendorf CO_2 plume. However, the incorporation of NO_2 field as inputs slightly increase the

345 quality of the predictions made by the new CNN model. One hypothesis to explain the error discrepancy between the model utilising NO_2 and the one using segmentation fields is due to the segmentation model not capturing NO_2 or CO_2 plume amplitude variations. Precisely, the segmentation model does not discriminate between plume pixels with high amplitude and those with low amplitude. Consequently, if the NO_2 images are of poor quality due to the presence of numerous alternative NO_2 plumes, the segmentation model will struggle to distinguish the main plume accurately. Taking into account these segmentation

350 fields, which do not distinguish the background plumes in the test dataset, results in a degradation of the predictions.

Regarding the benefits brought by the integration of NO_2 fields, their essential contribution is their facilitation of the plume segmentation, rather than a direct enablement of inversion based on the NO_2 levels alone. Indeed, as stated in section 4.2.1, the NO_2 plumes are not scaled like CO_2 plumes. To further investigate this, it is possible to verify this hypothesis by modifying the scaling of the plume in the NO_2 fields. In Fig. 8 we draw the relative absolute error of the model at testing stage when taking

355 as inputs scaled NO_2 fields by a constant factor in $[0,3]$. The CNN model exhibits its highest performance within the scaling range of 1.0 to 1.3. As expected, performances gradually decrease for scaling below 1 and above 1.3. However, intriguingly, the model still achieves remarkably satisfactory scores for scalings ranging from 0.5 to 1.75. Notably, these scores outperform those of the CNN models without NO_2 field as input. If the model were utilising the amplitude of the NO_2 field as a predictor, deeply inaccurate results could be expected for scalings of 0.5 or 1.5. However, the graph contradicts this assumption, strongly

360 suggesting that the amplitude of the NO_2 field is not employed as a predictor by the model. Instead, it is likely that only the contour of the NO_2 plume and the ratios between different parts of the plume serve as predictors. In essence, the model's reliance on the NO_2 field for predictions appears to be based on the contour of the plume and the relative proportions of its various components, rather than on the absolute amplitude of the NO_2 field values. Finally, the model's inability to accurately
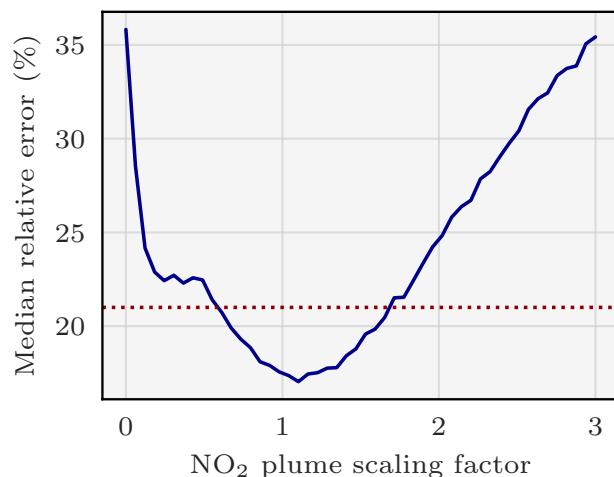
Figure 8. Effect of scaling the NO_2 plume of Lippendorf inputs on the performance of the CNN model. The x-axis corresponds to the scaling of the NO_2 plume: 1 corresponds to the original plume, 0 to no plume. The y-axis corresponds to the median absolute relative error of the CNN model evaluated for the given scaling of the NO_2 plumes. The CNN model is the same for each scaling (each dot of the x-axis) and corresponds to the CNN model having obtained the best absolute relative error score. The red dotted line approximately corresponds to the median absolute relative error of a model not learning with NO_2 fields.

estimate emissions for scaling factors exceeding 2 can be attributed to the unprecedented and extreme values that NO_2 plumes can reach, which lie beyond the range of what the model has been trained on.

5.2.2 Overfitting investigation

To understand the reasons behind the deviations between predictions and actual values, we conducted an analysis of the residuals in Fig. 9. This examination, which focuses on the disparities between predicted and true emissions, reveals a substantial underestimation of the actual Boxberg emissions by the model incorporating NO_2 or segmentation fields.

This observation suggests an overfitting issue (Zhang et al., 2022), since the majority of PPs used to train the model exhibit lower emission rates compared to Boxberg (6 out of 7: Schwarze Pumpe, Lippendorf, Turow, Patnow, Opole, and Dolna Odra). The low absolute relative error observed on the training dataset of the models with NO_2 or segmentation fields, as depicted in Fig. 10, further substantiates concerns regarding overfitting. An overfitted model tends to learn features that are overly tailored to the specifics of the training dataset. Consequently, when presented with images from the test dataset with slightly different features, the model struggles to generate accurate predictions.

Our hypothesis is that the model specifically overfits the low emissions PPs, which can be attributed to the information gained through the use of NO_2 or segmentation fields. When provided with new images, the model fails to recognise the new
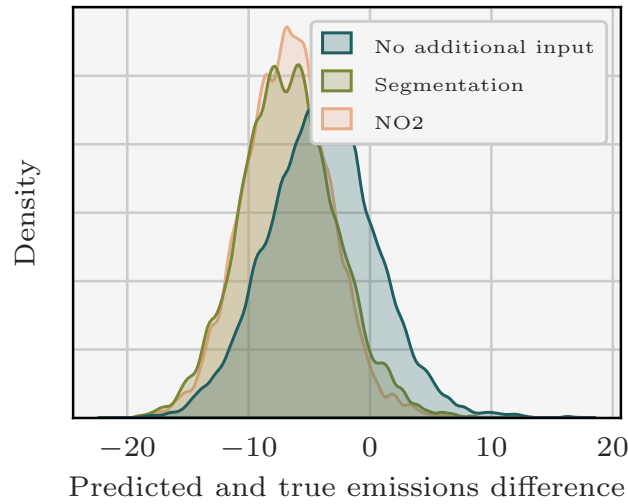
Figure 9. Residuals density between the true emissions of Boxberg PP and the predictions of the CNNs.
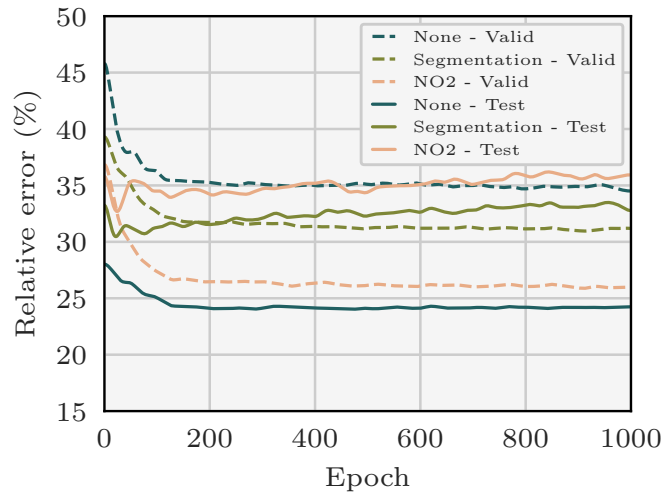
Figure 10. Evolution during training of the validation and test relative errors between the true emissions of Boxberg and the predictions of the CNNs. Specifically, for the need of this experiment, new models are trained and employed to predict the emissions corresponding to the validation and test fields at each epoch. Three models are considered: each is trained with the XCO_2 field and the winds as inputs. Two of the models additionally assimilate the NO_2 field or the predictions of the segmentation model. The validation error decreases monotonically with the number of epochs, while the test error does not, which suggests overfitting of the model.
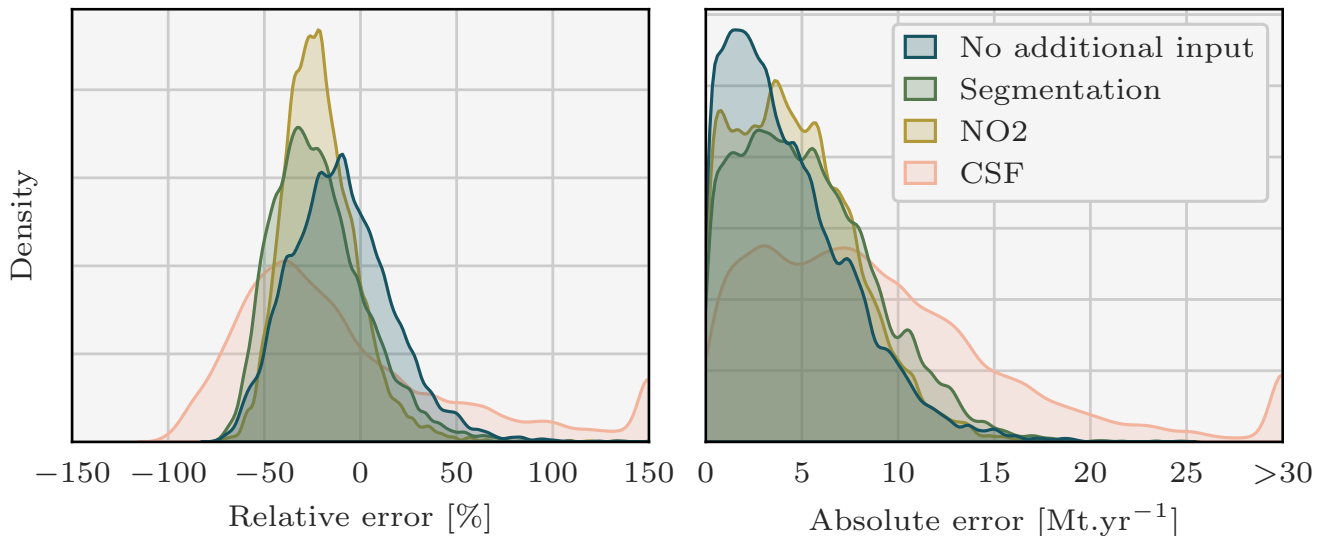
Figure 11. Density plots of the relative and absolute error between the predicted and the true emissions. Four sets of predictions are considered, corresponding to the three CNN models with three different sets of inputs and the CSF method. Each CNN model is trained with the XCO_2 field and the winds as inputs on a dataset composed of {Berlin, Jänschwalde, Lippendorf, Turow, Opole}. Two of the models additionally assimilate the NO_2 field or the predictions of the segmentation model. Predictions with absolute relative errors greater than 150% or absolute errors greater than $30\,Mt.yr^{-1}$ were set to 150 or 30 to increase visibility. 15% of the CSF method predictions are missing. Those predictions correspond to Boxberg plumes superimposed on other plumes, where the CSF method cannot be applied. A modified dataset is used to avoid overfitting.

features and consequently yields predictions that align more closely with the training dataset, which predominantly consists of emissions from the low emissions PPs. The failure to recognise new features is due to the incorporation of NO_2 field as inputs.

380   The model's ability to learn highly specific features is limited when no additional input is provided. Conversely, when the model incorporates the NO_2 field, it gains access to more information and can acquire more intricate features. Consequently, the model's capacity to generalise worsens in the latter case.

Next, we examine the performance of a new ensemble of models trained on a more balanced dataset, achieved by removing three out of the five low-emissions PP from the original dataset. The new dataset is composed of {Berlin, Jänschwalde, Lip-

385   pendorf, Turow, Opole}, i.e. of two medium or high emissions PPs, two low emissions PPs and Berlin. Furthermore, the choice of the factor scaling the plumes during training (see section 4.2.1) varies depending on the ensemble member considered (see section 4.2.4). Specifically, the uniform distribution is defined with a minimum scaling factor of 0.25 or 0.5 and a maximum scaling factor of 2 or 3. Once more, we examine the outcomes for three models obtained from ensembling, as depicted in Fig. 11, and summarised in Table 3.

	Absolute relative error [%]			Absolute error [Mt.yr^{-1}]		
	25%	Median	75%	25%	Median	75%
CNN with no additional input	9.5	20.4	33.8	1.8	3.7	6.3
CNN with segmentation	14.1	26.6	40.6	2.5	4.8	7.6
CNN with NO$_2$	13.2	23.8	34.2	2.3	4.4	6.6
CSF	21.7	41.5	63.5	3.9	7.7	12.3

Table 3. Relative and absolute error statistics between predicted and true Boxberg emissions for three CNN models (assimilating three different sets of inputs) and the CSF method. A modified dataset is used to avoid overfitting.

390   The three CNN models demonstrate very good performance, although the inclusion of NO$_2$ or segmentation fields still leads to a degradation in results, albeit to a lesser extent than in section 5.1.3. For example, the median absolute relative error for the CNN with NO$_2$ as additional input is 23.8%, comparing to 36.9% in section 5.1.3. On the one hand, the median absolute relative error of the CNN model trained on NO$_2$ and wind fields stands at approximately 20%, while the median absolute error remains below 4 Mt.yr^{-1}. On the other hand, the degradation of the results when adding the NO$_2$ or segmentation fields still

395   can be regarded as overfitting. The model learns features from NO$_2$ or segmentation fields that are not general enough to cover the case of Boxberg. Furthermore, it fails to acquire compensatory generalisable features such as in the case of Turow, where the model probably gains information about the plume contour from the NO$_2$ field, which is not straightforwardly apparent in the Turow XCO$_2$ field.

5.3   Interpretation of the CNN inversion models

400   In the two following sections, we introduce and apply two methods to gain a deeper understanding of the behaviour and decision-making processes of the CNNs discussed in this paper. These methods offer valuable insights into the significance of input features in the predictions made by the CNNs:

-    the integrated-gradient method allows us to examine the importance of individual pixels across channels;

-    the feature permutation method enables us to assess the importance of the channels, i.e. the fields used as inputs.

405   **5.3.1   Gradient-based study of the pixels**

Integrated Gradients is a gradient-based method for the interpretability of neural networks that enables the assessment of pixel importance in CNN prediction. It calculates the sensitivity of a model's predictions to input features (here pixels), assigning relevance scores to them. By analysing how changes in the input pixels affect the model's output, the method provides insights into the importance of each pixel in the decision-making process.

One unique aspect of this gradient-based method is that the scores it assigns are relative to a baseline. More precisely, these scores are computed as integrated gradients along a linear interpolation from a blank image to the input.

In Fig. 12, we apply the Integrated Gradient approach to study four different models specific to various sources. The first and second models are built to invert the emissions from Lippendorf (see section 5.1.1) and Boxberg plumes (see section 5.1.3), respectively, whereas the two last models target Turow plumes (see section 5.1.2). The first three models only use the XCO_2 field and the winds as inputs, whereas the fourth model considers an additional input, the NO_2 field. To apply the Integrated Gradient method, a random plume from the target PP is selected for each model. Columns of Fig. 12 represent the XCO_2 field, the corresponding XCO_2 plume and the Integrated Gradient between the model predictions and the inputs.

In order to simplify the analysis, we choose the model that exhibits the best performance on the test dataset, rather than using the ensemble of models. It is worth noting that such a model yields similar performances to the ensemble of models.

The Integrated Gradient technique reveals that the CNN model learns to estimate the emissions of a source based on the pixels of the plume from this source. In the first row, we examine a plume from Lippendorf PP. The Integrated Gradient technique identifies the most important pixels, which correspond to the plume pixels. This indicates that if the pixels associated with the plume were to deviate, the estimated flux rate would be significantly affected. This demonstrates that the CNN model effectively makes inference from the crucial parts of the image. In the second row, we focus on an image centred on Boxberg. The gradients reveal that the model concentrates exclusively on the Boxberg plume in the centre, disregarding the other plumes when inferring the emissions of Boxberg PP. In the third row of the figure, we examine a plume from Turow PP. Although the precision is lower compared to the Lippendorf or Boxberg cases, the pixels in the general direction of the Turow plume are the main ones used to estimate the emissions. In the fourth row, we display the gradients associated with the same Turow image but for a model trained with NO_2 fields as additional inputs. In this case, the model clearly identifies the pixels corresponding to the plume as critical, as indicated by the amplitude and contour of the gradients. This reinforces the hypothesis that the improved estimation of Turow emissions when the model is trained with NO_2 fields, can be attributed to the enhanced assessment of the plumes.

In conclusion, the model consistently identifies the target emission plume situated at the image's centre, indicating it implicitly understands the relationship between the plume and targeted emissions.

5.3.2 Feature permutation analysis

Feature permutation is a technique used to determine the importance of input channels used in a model (Molnar, 2022). As input variables used here are not independent, the interpretation of the following permutation analysis should be taken with caution. The principle is to i) permute the values of a feature (e.g. exchange the u wind field corresponding to an XCO_2 image with another random u wind field), and ii) use the model to predict emissions for the given input, which includes the XCO_2 field, other associated inputs and the random u wind field. By comparing the performance of the model on the original dataset with the performance on the permuted dataset, we can measure the impact of each feature on the performance of the model. The more the permutation of a feature affects the performance of the model, the more important that feature is. In Table 4, we present the outcomes of the permutation of the features for nine ensembles of models (for each PP and each ensemble of inputs
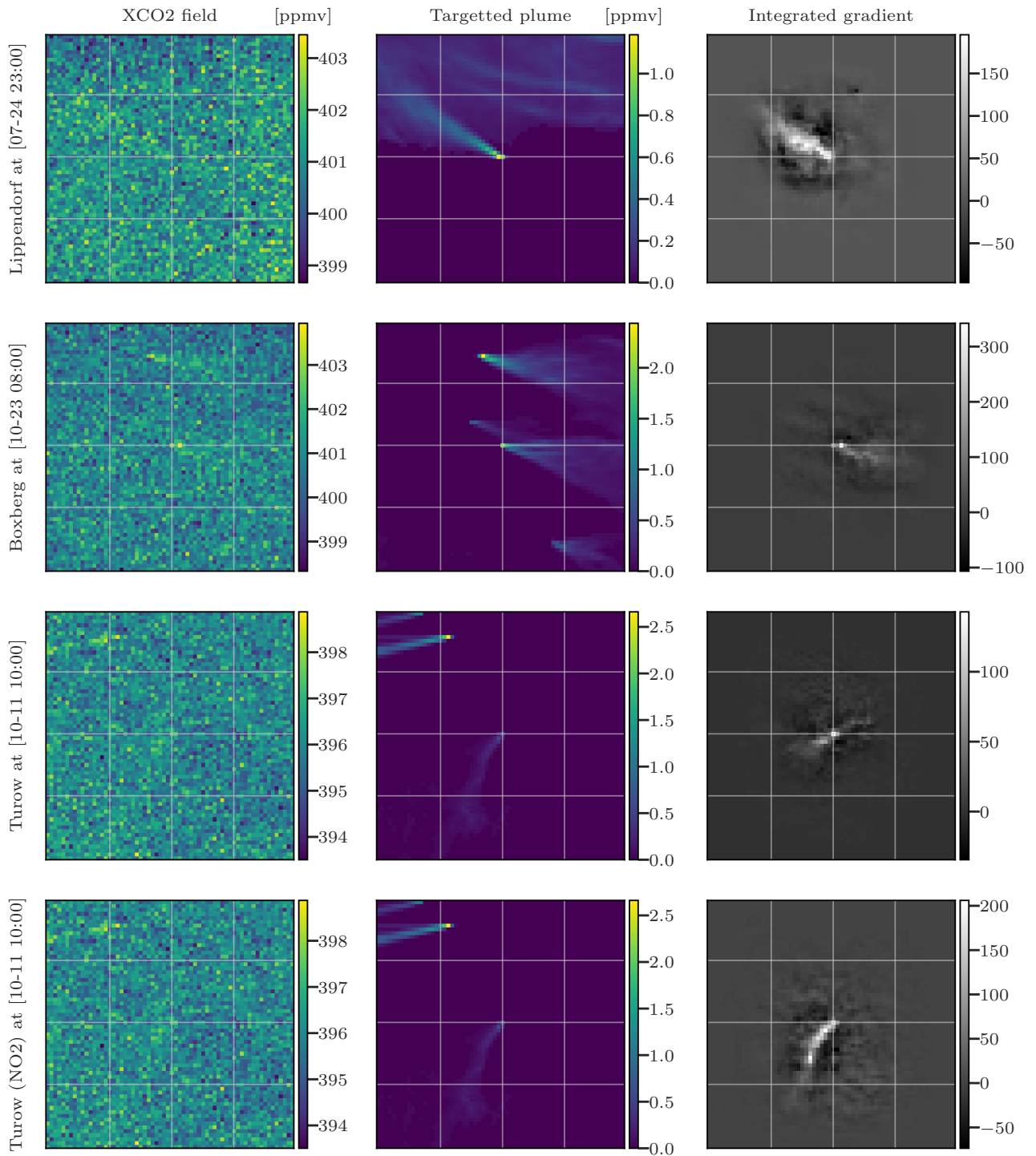
Figure 12. Evaluation of four CNN models using the Integrated Gradients method on four input sets. Columns 1, 2, and 3 represent the XCO$_2$ field, the corresponding XCO$_2$ plume and the Integrated Gradient between the model predictions and the inputs, in that order. The rows represent the four models and corresponding test fields.

PP	Lippendorf			Boxberg			Turow		
Feature	None	Seg.	NO$_2$	None	Seg.	NO$_2$	None	Seg.	NO$_2$
XCO$_2$	18.3	27.6	32.7	19.6	29.5	34.5	12.6	9.8	22.7
u-wind	11.6	10.4	3.3	12.4	9.7	4.2	6.6	5.4	0.9
v-wind	4.9	3.7	1.8	5.3	4.9	3.0	3.6	1.7	0.6
Add. input	n/a	22.7	30.3	n/a	28.9	34.6	n/a	6.9	21.0

Table 4. Evaluation of the degradation in the average absolute relative error of the model when the corresponding feature is permuted. Each column corresponds to a model, and each row corresponds to a permuted feature. For example, we estimate the emissions corresponding to Boxberg plumes with the CNN model trained with the XCO$_2$ field, the winds and no additional input. We then estimate the degradation of the predictions of this model on the test dataset with the u-wind fields permuted: the degradation in absolute relative error is 12.4%. Seg. is for Segmentation. Add. input corresponds to either None, NO$_2$, or Segmentation fields depending on the considered column.

considered). Each entry in the table represents the degradation in the average absolute relative error of the model (associated with a specific column) when the corresponding feature is permuted (related to the respective row).

We can formulate hypotheses on several groups of cells in Table 4:

– comparison of the first row with the others shows that the XCO$_2$ feature consistently holds (or shares) the highest importance among other features;

– examining the second and third rows reveals a pattern where the u-wind seems to have a more significant influence on the inversion process compared to the v-wind. It might be due to the bigger variance of u ($\sim 51\,\mathrm{m}^2.\mathrm{s}^{-2}$) comparing to v ($\sim 19\,\mathrm{m}^2.\mathrm{s}^{-2}$): u is the dominant wind in this situation;

– the XCO$_2$ field input permutation seems to have a lesser impact for the CNN targetting Turow (12.6% in absolute relative error degradation comparing to 18.3 and 19.6% for Lippendorf and Boxberg);

– the importance of the XCO$_2$ feature differs for models with different inputs: specifically, the XCO$_2$ feature importance increases when additional data are used. For example, the XCO$_2$ feature permutation degrades the performance of the model by 32.7% when NO$_2$ is used, and by 18.3% when no additional input is used. This observation is consistent with the overfitting tendency of the CNN models when trained with NO$_2$ fields. When confronted with inconsistent data (non-corresponding XCO$_2$ and NO$_2$ fields), the acquired complex features of the model exhibit a total absence of correspondence with these inputs. As a consequence, the model predicts nonsensical emissions;

– the wind inputs systematically hold greater importance in the model without additional input, followed by a relatively diminished importance in the model with segmentation fields, and finally, it exhibits the least significance in the model with

NO$_2$ fields. This suggests that the inclusion of NO$_2$ fields is more advantageous for inversion compared to segmentation fields. Furthermore, it indicates that the winds compensate the absence of additional data.

6   Discussions and limitations

465   The approach developed in this paper carries certain limitations: firstly, it focuses exclusively on European PPs, which questions its generalisability to PPs in other regions with different climatic conditions. Secondly, the study emphasises the importance of a balanced dataset, as highlighted by the section 5.2.2, and the need to be able to identify and address potential overfitting issues. It should also be noted that the study did not specifically investigate outliers such as PPs with exceptionally high emission rates. Despite incorporating plume scaling strategies, the model may struggle to generalise to such outliers. A last limitation is

470   the absence of zero-emission source in the dataset. However, the inclusion in the training dataset of very low emission power plants and of a plume scaling approach generating near-zero emission plumes indicates that incorporating zero emission cases would likely not markedly change the outcomes.

In terms of future research, several areas should be explored such as the challenge posed by clouds. In this respect, CNNs can be trained to ignore missing data caused by cloud cover and to make effective use of the available data. Another aspect

475   to consider is the presence of noise in CO$_2$M data. While Gaussian noise may not pose significant issues, if the satellite noise exhibits structured patterns, it would becomes crucial to develop robust noise modelling techniques to enable CNNs to accurately distinguish and remove such noise. Furthermore, in real-world applications, models trained on synthetic datasets may face challenges when applied to real datasets due to differences in data distribution. Strategies such as importance weighting, specific data augmentation techniques, transfer learning, or active learning methods may be necessary to account for these dif-

480   ferences and ensure reliable performance. Finally, the method could be modified to extract CO2 emissions from plume imagery at more detailed scales, necessitating the use of resource-intensive Large-Eddy Simulation (LES) models.

7   Conclusions and perspectives

The future availability of CO$_2$ satellite imagery, through missions such as the Copernicus CO$_2$ Monitoring (CO$_2$M) mission, heralds new opportunities and challenges for the evaluation of local CO$_2$ emissions. Emissions can certainly be retrieved from

485   CO$_2$ plumes of hotspots in the satellite images. But the emission estimation is hindered by two primary obstacles: low signal-to-noise ratio plumes which cannot be extracted straightforwardly from the background, and uncertainties in the transport or dispersion processes which hampers the assessment of the emissions from the plumes.

In this work, we assess the ability of Convolutional Neural Networks (CNNs) to invert a plume from satellite imagery using simulated XCO$_2$, NO$_2$, and wind fields with similar characteristics to the future CO$_2$M images. The fields used to train and

490   evaluate the CNNs are based on the SMARTCARB dataset and possess the same resolution, satellite noise level and ancillary data availability to the future CO$_2$M images. Each synthetic XCO$_2$ image encompasses the anthropogenic plume from at least one power plant, a background arising from other biogenic and anthropogenic fluxes, and a random Gaussian noise intended

to simulate the errors inherent in satellite instruments. But the model evaluation was conducted using simulated data. This approach does not account for all the challenges that real satellite images present, specifically issues related to cloud cover and systematic error patterns due to surface reflectance and the aerosol dependency of retrievals.

Our source emission estimation model is an image-to-scalar CNN model, which infers from the full XCO_2 field a flux rate, estimation of the anthropogenic emission corresponding to the plume. The first layers of the model consist of preprocessing steps which transform the input data at training time, in particular by adding noise and by scaling the plume and emission rate. The core model is a CNN composed of approximately 200,000 neurons divided into convolutional, max pooling, dropout and batch normalisation layers.

We highly suggest that the design of a "universal" CNN, trained on a small power plant subset and highly accurate on all of them, is possible. To do so, we evaluate the model's ability to generalise to unobserved images from another region. Explicitly, three CNN models trained on different datasets but sharing the exact same structure (hyperparametrisation, architecture, ...) are tested on plumes from three sources: Boxberg, Lippendorf and Turow. The training/validation dataset for each CNN is restricted to a dataset consisting of all other power plants except their target. The CNNs are highly accurate in each case and the addition of NO_2 fields often improves the results slightly. Precisely, the median absolute relative errors for the CNN models are on average close to 20–25 %. Moreover, the median absolute error is generally half that obtained with the CSF method: an alternative and state of the art inversion approach. This highly suggests that it is possible to build a universal neural network (which can generalise to all targets) using this methodology.

By using interpretability tools, we demonstrate that the predictions made by the CNNs are grounded in the physically meaningful components of the features. The Integrated Gradient method shows that the CNNs learn to predict the emissions corresponding to a plume from the pixels making up the plume. The feature permutation technique highlighted several aspects of the models, such as the expected high importance of the XCO_2 fields compared to the used ancillary data.

Future prospects of the CNN plume inversion method from satellite images encompass the challenges of clouds, cities and real satellite images. Concretely, the method should be able to handle missing data caused by clouds. Additionally, the CNN approach should incorporate the second important category of hotspots: cities. Finally, the method should be tested on real satellite data once it becomes available.

Data availability. The datasets used in this paper are available on a compliant repository on https://zenodo.org/record/8096616 and originate from https://zenodo.org/record/4048228. The weights of the CNNs are available on https://zenodo.org/record/8095487. The algorithms are available on Zenodo (https://zenodo.org/record/10100338) and Github at https://github.com/cerea-daml/co2-images-inv-dl.

Author contributions. Joffrey Dumont Le Brazidec: Conceptualisation, Methodology, Software, Investigation, Formal analysis, Visualisation, Resources, Project administration, Writing - Original Draft; Pierre Vanderbecken: Investigation, Formal analysis, Writing - Review; Alban Farchi: Conceptualisation, Methodology, Project administration, Writing - Review; Marc Bocquet: Conceptualisation, Methodology,

Project administration, Funding acquisition, Writing - Review; Grégoire Broquet: Conceptualisation, Writing - Review; Gerrit Kuhlmann:
525 Resources, Writing - Review.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This project has been funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 958927 (Prototype system for a Copernicus CO_2 service). CEREA is a member of Institut Pierre-Simon Laplace (IPSL). We would like to acknowledge Tobias Finn for the valuable thought-provoking discussions, as well as Evan D. Sherwin and our second reviewer
530 for their insightful comments.

References

Beirle, S., Borger, C., Dörner, S., Li, A., Hu, Z., Liu, F., Wang, Y., and Wagner, T.: Pinpointing nitrogen oxide emissions from space, Science Advances, 5, eaax9800, https://doi.org/10.1126/sciadv.aax9800, publisher: American Association for the Advancement of Science, 2019.

Broquet, G., Bréon, F.-M., Renault, E., Buchwitz, M., Reuter, M., Bovensmann, H., Chevallier, F., Wu, L., and Ciais, P.: The potential of satellite spectro-imagery for monitoring CO_2 emissions from large cities, Atmospheric Measurement Techniques, 11, 681–708, https://doi.org/10.5194/amt-11-681-2018, publisher: Copernicus GmbH, 2018.

Brunner, D., Kuhlmann, G., Marshall, J., Clément, V., Fuhrer, O., Broquet, G., Löscher, A., and Meijer, Y.: Accounting for the vertical distribution of emissions in atmospheric CO_2 simulations, Atmos Chem Phys, 19, 4541–4559, https://doi.org/10.5194/acp-19-4541-2019, publisher: Copernicus GmbH, 2019.

Chevallier, F., Remaud, M., O'Dell, C. W., Baker, D., Peylin, P., and Cozic, A.: Objective evaluation of surface- and satellite-driven carbon dioxide atmospheric inversions, Atmospheric Chemistry and Physics, 19, 14233–14251, https://doi.org/10.5194/acp-19-14233-2019, publisher: Copernicus GmbH, 2019.

Chevallier, F., Broquet, G., Zheng, B., Ciais, P., and Eldering, A.: Large CO2 Emitters as Seen From Satellite: Comparison to a Gridded Global Emission Inventory, Geophysical Research Letters, 49, e2021GL097540, https://doi.org/10.1029/2021GL097540, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2021GL097540, 2022.

Chollet, F.: Deep Learning with Python, Manning Publications Co., USA, 1st edn., 2017.

Cusworth, D. H., Duren, R. M., Thorpe, A. K., Eastwood, M. L., Green, R. O., Dennison, P. E., Frankenberg, C., Heckler, J. W., Asner, G. P., and Miller, C. E.: Quantifying Global Power Plant Carbon Dioxide Emissions With Imaging Spectroscopy, AGU Advances, 2, e2020AV000350, https://doi.org/10.1029/2020AV000350, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2020AV000350, 2021.

Dumont Le Brazidec, J., Vanderbecken, P., Farchi, A., Bocquet, M., Lian, J., Broquet, G., Kuhlmann, G., Danjou, A., and Lauvaux, T.: Segmentation of XCO2 images with deep learning: application to synthetic plumes from cities and power plants, Geoscientific Model Development Discussions, pp. 1–29, https://doi.org/10.5194/gmd-2022-288, publisher: Copernicus GmbH, 2022.

Finch, D., Palmer, P., and Zhang, T.: Automated detection of atmospheric NO2 plumes from satellite data: a tool to help infer anthropogenic combustion emissions, Atmos. Meas. Tech., pp. 1–21, https://doi.org/10.5194/amt-2021-177, 2021.

Hakkarainen, J., Szeląg, M. E., Ialongo, I., Retscher, C., Oda, T., and Crisp, D.: Analyzing nitrogen oxides to carbon dioxide emission ratios from space: A case study of Matimba Power Station in South Africa, Atmospheric Environment: X, 10, 100110, https://doi.org/10.1016/j.aeaoa.2021.100110, 2021.

Hakkarainen, J., Ialongo, I., Koene, E., Szelag, M., Tamminen, J., Kuhlmann, G., and Brunner, D.: Analyzing Local Carbon Dioxide and Nitrogen Oxide Emissions From Space Using the Divergence Method: An Application to the Synthetic SMARTCARB Dataset, Frontiers in Remote Sensing, 3, https://doi.org/10.3389/frsen.2022.878731, 2022.

Hakkarainen, J., Tamminen, J., Nurmela, J., Santaren, D., Broquet, G., Chevallier, F., Koene, E., Kuhlmann, G., and Brunner, D.: D4.4 Benchmarking of plume detection and quantification methods | CoCO2: Prototype system for a Copernicus CO2 service, Technical Report 4.4, https://coco2-project.eu/node/366, 2023.

IEA, I. E. A.: Fuel share of CO2 emissions from fuel combustion, 2019 – Charts – Data & Statistics, https://www.iea.org/data-and-statistics/charts/fuel-share-of-co2-emissions-from-fuel-combustion-2019, 2019.

Jongaramrungruang, S., Matheou, G., Thorpe, A. K., Zeng, Z.-C., and Frankenberg, C.: Remote sensing of methane plumes: instrument tradeoff analysis for detecting and quantifying local sources at global scale, Atmospheric Measurement Techniques, 14, 7999–8017, https://doi.org/10.5194/amt-14-7999-2021, publisher: Copernicus GmbH, 2021.

570 Joyce, P., Ruiz Villena, C., Huang, Y., Webb, A., Gloor, M., Wagner, F. H., Chipperfield, M. P., Barrio Guilló, R., Wilson, C., and Boesch, H.: Using a deep neural network to detect methane point sources and quantify emissions from PRISMA hyperspectral satellite images, Atmospheric Measurement Techniques, 16, 2627–2640, https://doi.org/10.5194/amt-16-2627-2023, publisher: Copernicus GmbH, 2023.

Koene, E., Brunner, D., and Kuhlmann, G.: Documentation of plume detection and quantification methods, Technical Report 4.3, Empa, https://www.coco2-project.eu/node/329, 2021.

575 Kuenen, J. J. P., Visschedijk, A. J. H., Jozwicka, M., and Denier van der Gon, H. a. C.: TNO-MACC_II emission inventory; a multi-year (2003–2009) consistent high-resolution European emission inventory for air quality modelling, Atmos Chem Phys, 14, 10 963–10 976, https://doi.org/10.5194/acp-14-10963-2014, 2014.

Kuhlmann, G., Broquet, G., Marshall, J., Clément, V., Löscher, A., Meijer, Y., and Brunner, D.: Detectability of CO_2 emission plumes of cities and power plants with the Copernicus Anthropogenic CO_2 Monitoring (CO2M) mission, Atmos. Meas. Tech., 12, 6695–6719, 580 https://doi.org/10.5194/amt-12-6695-2019, 2019.

Kuhlmann, G., Brunner, D., Broquet, G., and Meijer, Y.: Quantifying CO_2 emissions of a city with the Copernicus Anthropogenic CO_2 Monitoring satellite mission, Atmos. Meas. Tech., 13, 6733–6754, https://doi.org/10.5194/amt-13-6733-2020, 2020.

Kuhlmann, G., Henne, S., Meijer, Y., and Brunner, D.: Quantifying CO2 Emissions of Power Plants With CO2 and NO2 Imaging Satellites, Front. remote sens., 2, https://www.frontiersin.org/article/10.3389/frsen.2021.689838, 2021.

585 Kumar, S., Arevalo, I., Iftekhar, A. S. M., and Manjunath, B. S.: MethaneMapper: Spectral Absorption Aware Hyperspectral Transformer for Methane Detection, pp. 17 609–17 618, https://openaccess.thecvf.com/content/CVPR2023/html/Kumar_MethaneMapper_Spectral_Absorption_Aware_Hyperspectral_Transformer_for_Methane_Detection_CVPR_2023_paper.html, 2023.

Lary, D. J., Alavi, A. H., Gandomi, A. H., and Walker, A. L.: Machine learning in geosciences and remote sensing, Geoscience Frontiers, 7, 3–10, https://doi.org/10.1016/j.gsf.2015.07.003, 2016.

590 Meijer, Y.: Copernicus CO2 Monitoring Mission Requirements Document, Earth and Mission Science Division, 84, https://esamultimedia.esa.int/docs/EarthObservation/CO2M_MRD_v3.0_20201001_Issued.pdf, 2020.

Molnar, C.: Interpretable Machine Learning: A Guide For Making Black Box Models Explainable, Independently published, Munich, Germany, 2022.

Nassar, R., Hill, T. G., McLinden, C. A., Wunch, D., Jones, D. B. A., and Crisp, D.: Quantifying CO2 Emissions From Individual Power Plants From Space, Geophysical Research Letters, 44, 10,045–10,053, https://doi.org/10.1002/2017GL074702, _eprint: 595 https://onlinelibrary.wiley.com/doi/pdf/10.1002/2017GL074702, 2017.

Nassar, R., Moeini, O., Mastrogiacomo, J.-P., O'Dell, C. W., Nelson, R. R., Kiel, M., Chatterjee, A., Eldering, A., and Crisp, D.: Tracking CO2 emission reductions from space: A case study at Europe's largest fossil fuel power plant, Frontiers in Remote Sensing, 3, https://www.frontiersin.org/articles/10.3389/frsen.2022.1028240, 2022.

600 Pillai, D., Buchwitz, M., Gerbig, C., Koch, T., Reuter, M., Bovensmann, H., Marshall, J., and Burrows, J. P.: Tracking city CO_2 emissions from space using a high-resolution inverse modelling approach: a case study for Berlin, Germany, Atmospheric Chemistry and Physics, 16, 9591–9610, https://doi.org/10.5194/acp-16-9591-2016, publisher: Copernicus GmbH, 2016.

Reuter, M., Buchwitz, M., Schneising, O., Krautwurst, S., O'Dell, C. W., Richter, A., Bovensmann, H., and Burrows, J. P.: 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, Atmos Chem Phys, 19, 9371–9383, https://doi.org/10.5194/acp-19-9371-2019, 2019.

Santaren, D., Hakkarainen, J., Kuhlmann, G., Koene, E., Chevallier, F., Ialongo, I., Lindqvist, H., Nurmela, J., Tamminen, J., Amoros, L., Brunner, D., and Broquet, G.: Benchmarking data-driven inversion methods for the estimation of local CO_2 emissions from XCO_2 and NO_2 satellite images, Atmospheric Measurement Techniques Discussions, pp. 1–52, https://doi.org/10.5194/amt-2023-241, publisher: Copernicus GmbH, 2024.

Tan, M. and Le, Q. V.: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, arXiv:1905.11946 [cs, stat], http://arxiv.org/abs/1905.11946, 2020.

Varon, D. J., Jacob, D. J., McKeever, J., Jervis, D., Durak, B. O. A., Xia, Y., and Huang, Y.: Quantifying methane point sources from fine-scale satellite observations of atmospheric methane plumes, Atmospheric Measurement Techniques, 11, 5673–5686, https://doi.org/10.5194/amt-11-5673-2018, publisher: Copernicus GmbH, 2018.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, I., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors, Vijaykumar, A., Bardelli, A. P., Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C. N., Fulton, C., Masson, C., Häggström, C., Fitzgerald, C., Nicholson, D. A., Hagen, D. R., Pasechnik, D. V., Olivetti, E., Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G. A., Ingold, G.-L., Allen, G. E., Lee, G. R., Audren, H., Probst, I., Dietrich, J. P., Silterra, J., Webber, J. T., Slavič, J., Nothman, J., Buchner, J., Kulick, J., Schönberger, J. L., de Miranda Cardoso, J. V., Reimer, J., Harrington, J., Rodríguez, J. L. C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma, M., Newville, M., Kümmerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N. J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, P. A., Lee, P., McGibbon, R. T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., Oshima, T., Pingel, T. J., Robitaille, T. P., Spura, T., Jones, T. R., Cera, T., Leslie, T., Zito, T., Krauss, T., Upadhyay, U., Halchenko, Y. O., and Vázquez-Baeza, Y.: SciPy 1.0: fundamental algorithms for scientific computing in Python, Nature Methods, 17, 261–272, https://doi.org/10.1038/s41592-019-0686-2, 2020.

Wang, Y., Broquet, G., Bréon, F.-M., Lespinas, F., Buchwitz, M., Reuter, M., Meijer, Y., Loescher, A., Janssens-Maenhout, G., Zheng, B., and Ciais, P.: PMIF v1.0: assessing the potential of satellite observations to constrain CO_2 emissions from large cities and point sources over the globe using synthetic data, Geoscientific Model Development, 13, 5813–5831, https://doi.org/10.5194/gmd-13-5813-2020, publisher: Copernicus GmbH, 2020.

Wu, D., Lin, J. C., Oda, T., and Kort, E. A.: Space-based quantification of per capita CO2 emissions from cities, Environmental Research Letters, 15, 035 004, https://doi.org/10.1088/1748-9326/ab68eb, publisher: IOP Publishing, 2020.

Zhang, A., Lipton, Z. C., Li, M., and Smola, A. J.: Dive into Deep Learning, https://doi.org/10.48550/arXiv.2106.11342, 2022.

Zheng, B., Chevallier, F., Ciais, P., Broquet, G., Wang, Y., Lian, J., and Zhao, Y.: Observing carbon dioxide emissions over China's cities and industrial areas with the Orbiting Carbon Observatory-2, Atmospheric Chemistry and Physics, 20, 8501–8510, https://doi.org/10.5194/acp-20-8501-2020, publisher: Copernicus GmbH, 2020.