Discussion: Segmentation of XCO_2 images with deep learning: application to synthetic plumes from cities and power plants

Joffrey Dumont Le Brazidec[1], Pierre Vanderbecken[1], Alban Farchi[1], Marc Bocquet[1], Jinghui Lian[2,3], Grégoire Broquet[2], Gerrit Kuhlmann[4], Alexandre Danjou[2], and Thomas Lauvaux[2]

[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]Origins.S.A.S, Suez Group, Île-de-France, France
[4]Swiss Federal Laboratories for Materials Science and Technology (Empa), Dübendorf, Switzerland

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

In the following, the referees comments are in italics and in blue.

Report 1

We would like to thank the anonymous Referee 1 for her/his constructive comments and suggestions, which allowed us to clarify several points in the manuscript.

This manuscript presents a work of developing a deep-learning-based model for plume segmentation of XCO2 over cities or power plants in European countries. They evaluated the model for model generalization on new data from the same region and model extrapolation on unseen data from another region. The results indicate the proposed segmentation model outperforms the usual segmentation technique based on thresholding. In general, the presentation of the paper is clear, and the potential of this technique is well-suggested. However, further explanation is needed on how this technique can be applied to estimate emissions from satellite imagery.

Detailed comments:

- *1)In the introduction section,*
  - *The additional reference is needed for that NO2 can be a proxy to CO2 and with NO2, the plume detection capabilities are significantly improving.*
  - *Since CO2M is a satellite mission, the author is considering applying this technique; a more detailed explanation of CO2M is needed, such as the spatial resolution, channel information, etc.*

  - A reference to (**?**) has been added.

  - Swath, resolution, and channel details on CO2M have been added in the introduction section (from https://www.eoportal.org/satellite-missions/co2m).

- *2) In the 2.2 section, page 5.*

  - *The data for Paris are selected for Jan., Mar., and Aug. Is there any specific reason to use these three month?*

  We have chosen two winter months and one summer month. Plumes in winter are in principle more visible, whereas plumes in summer are less visible. With these three months we cover both cases. These simulations are costly and these three months correspond to the Paris data already simulated and available at the time of writing.

- *2) In the 2.2 section, page 5.*

  - *How much has the results performance improved using data augmentation techniques?*

  The following paper introduced the data augmentation technique for weather applications considering major wind direction. Like this, have you considered the domain characteristics in data augmentation methods? "Seo, Minseok, et al. "Domain Generalization Strategy to Train Classifiers Robust to Spatial-Temporal Shift." arXiv preprint arXiv:2212.02968 (2022)."

  The data augmentation techniques used in the document are critical. The overall performance is much improved with data augmentation techniques. Without these techniques, the model overfits the training data and generalises poorly, with a nwbce score close to 1 (depending on the case).

  Seo, Minseok, et al. 2022 propose to choose physically-consistent augmentations for weather forecasting (based in particular on regionality). We preferred a data-driven approach to calibrate our data augmentation choices. Several models trained on differently augmented datasets were competed on the validation set.

  We have added the following in the manuscript to precise: "The selection of the data augmentation techniques used and their characteristics was based on experimentation."

- *3) In the 3.4 section,*

  - *The results showed when the concentration is low or signal-to-noise is small, the performance is significantly degraded. The author mentioned NO2 is helpful for that in the introduction section. Then, why is NO2 data not used as an additional input to solve this problem?*

  Thank you for this pertinent question. This study is limited to the segmentation of XCO2 plumes using only XCO2 data. It shows that for many plumes, additional NO2 data are not needed to get good predictions. We wanted to demonstrate this before investigating the use of NO2. We will test the use of NO2 for segmentation a in a subsequent publication. (Additionally, other recent missions planned for CO2 do not have NO2 instrument. We have added a reference to CO2Image, another planned CO2 satellite without an additional instrument to measure NO2.)

- *3) In the 3.4 section,*

  - *In the deep-learning approaches, the data split is important. Generally, the training and validation dataset are randomly split, while the test is separated from the training and validation. It would be best if you used separate datasets, not days in the middle of the same month used in the training dataset. And please indicate how many datasets are in each training, validation, and test dataset.*

  Here, train, validation, and test data sets are totally disjoint: there is no overlap. We did not randomly split the data because two plumes from the same hotspot at two consecutive hours may be similar. It is therefore preferable to randomly split the full period of simulation for a given source into blocks of several days (here, two). Furthermore, note that in the "extrapolation" case, the test dataset is completely dissimilar from the train and validation datasets as the test hotspot (Berlin) is not considered in the train and validation sets.

  We have added the following paragraph in the manuscript: "The numbers of data for training, validation and test differ for each test case. In the last case (extrapolation to Berlin), there are about 23,000 images in the training dataset, 4,000 in the validation dataset and 7,000 in the test dataset. It is worth mentioning that data augmentation techniques enable us to use a significantly greater number of training images in practice."

- *4) In the results,*

  - *Most plume smoke shapes are long-tailed, and when the smoke does not spread and gathers in the middle, the segmentation results are not as good as those from long-tailed shapes. There has been a bias towards the plume shape. It seems necessary to analyze whether the result of having a higher wbce score was influenced by the shape of the plume.*

  In line with your suggestion, we investigate in the following whether the model is biased towards the shape of the plume. More precisely, we investigate whether plumes that stack in the middle are less well reconstructed than long-tailed plumes. For this study, we calculate for each plume a quantity called "ratio centre mass". This quantity is the sum of the plume concentrations in the centre of the image divided by the sum of all plume concentrations. The dimensions of the images span from 0 to 160 in both the x and y directions. The center of the images is defined as the pixels located within the range of [40:120, 40:120]. We can then group the plumes into four categories according to their "ratio centre mass" and plot the kernel densities (histograms) of the nwbce of these four categories in Fig. 1. The best reconstructed plumes are those with a medium "ratio centre mass". Furthermore:

  - low "ratio centre mass" plumes is the category with the worst nwbce. A large plume spread (due to a strong wind) is correlated with a smaller amplitude (as the plume is spread). It is therefore reasonable to observe that a large spread (i.e. low "ratio centre mass") is correlated with lower model performance;

  - plumes with high "ratio centre mass" (e.g. plumes that stack in the middle) also have a significant nwbce

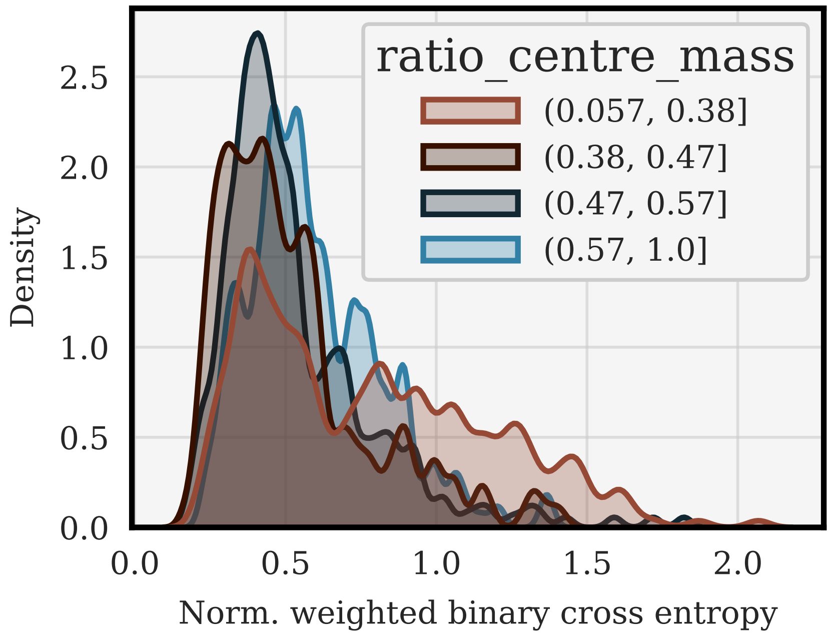

Figure 1. Histograms of the nwbce image scores over all test images of Lippendorf. The plumes are classified in four equivalent clusters according to the ratio centre mass.

This second fact seems in line with the bias hypothesis, but the first fact shows that other phenomenons (e.g. the amplitude of the plume) have a higher influence on the performance of the model. Since optimal results are obtained for intermediate ratios, we can assume that the competence of the model is a compromise between amplitude and a plume-like shape. For a more complete analysis, it would be necessary to create much more specific categories of plumes according to their shape but this is beyond the scope of this manuscript since it would require much more experimenting.

Finally, we propose to add this additional study in supplementary material.

- *4) In the results,*
  - *How you get the emission amount in the Figure 13.*

The indicated emissions are the inventory emissions that were used to generate the data. A precision has been added in the manuscript.

Thank you very much for all these suggestions.

Report 2

We would like to thank the anonymous Referee 2 for her/his technical comments and suggestions on improving the manuscript.

Segmentation of XCO2 images with deep learning: application to synthetic plumes from cities and power plants" by Dumont Le Brazidec et al. describes a new method for segmentation of CO2 plumes in satellite imagery that could improve methods for quantifying CO2 emissions from anthropogenic sources. The authors develop and train a convolutional neural network (CNN) for XCO2 plume segmentation from satellite observations that outperforms the more common thresholding approaches, when demonstrated on modelled plumes with random noise and non-uniform backgrounds. The manuscript applies atmospheric models to address a relevant scientific question that has implications for the Copernicus CO2M mission and monitoring, verification and support (MVS) efforts to support climate policy. The CNN approach developed is novel and represents a substantial advance for the field. In general, the method is described well and the manuscript overall is well-written, well-structured and well-presented.

Specific Points

- *Line 1: It would be best if expanded text matched acronym: "CO2 Monitoring Verification and Support" system and "CO2MVS" system.*

  This is true. It has been corrected. Thank you.

- *Line 2: "Amount of CO2" should be replaced with "Distribution of CO2"*

  We made the correction.

- *Line 35: It is my understanding that 2 of 3 satellites CO2M are scheduled for launch in 2026, with the 3rd to follow by a year or more. (The authors should confirm whether this is correct and if not, revise with the most up to date information).*

  As far as we know, there is no certainty about the dates. We have therefore replaced "by 2026" with "from 2026".

- *Line 113: including only January, March and August barely the minimum for representing "seasonal variability" with 3 of 4 seasons. A more thorough effort could have been made here.*

  This is the case for the Paris dataset. However, we use a full year for the SMARTCARB dataset which represents most of our data (Berlin, and 3 other power plants presented in this manuscript).

- *Line 120: Some basic information about CO2M observing characteristics would be useful for the reader, at minimum, the nominal image pixel size (2x2 km2) and swath width warrant mentioning. Furthermore, it is unclear whether the 0.7 ppm Gaussian random noise is applied at the model resolution of 1.1x1.1 km2 or the CO2M imagining pixel size, which was never mentioned. 0.7 ppm noise at 1.1x1.1 km2 is equivalent to a lower noise level at 2x2 km2.*

  Details on the swath and resolution of CO2M have been added in the introduction section (from https://www.eoportal.org/satellite-missions/co2m).

This manuscript is mainly motivated by the sole real mission now planned for launch with very large swath (i.e CO2M). But its focus is on CO2 images in general. We have included clarifications in the introduction to prevent any potential confusion. For our analysis, we used images of approximately 1km resolution, which differs from the CO2M resolution. This resolution is the native resolution of the SMARTCARB dataset. Consequently, we applied Gaussian random noise of 0.7 ppm at the model resolution.

– *For improved readability, capitalization of "WBCE", "NWBCE" and "DDEQ" acronyms is strongly recommended throughout the entire manuscript.*

  WBCE and NWBCE have been capitalised in the manuscript. ddeq has been kept in lower case for consistency with other manuscripts and the python page.

– *It might be difficult to reproduce these results from only the brief description about how the plumes were modified to generate the training dataset.*

  We have added the following paragraph (end of section 3.2) to specify how the plumes are modified to generate the training dataset: "In practice, during the training phase, the plumes undergo a two-step transformation process. Firstly, they are transformed using the weight function described in Eq. 3. Subsequently, they undergo further transformation using the data augmentation techniques specified in section 2.2. The resulting transformed plumes are subjected to the loss function defined in Eq. 4 during training." In addition, we have made available the code used to modify the plumes during the training phase.

– *Figure 13 seems to be the only mention of the magnitude of emission sources in the manuscript. For some perspective the authors should mention either the annual emissions for the sources in the study (Paris, Berlin and power plants). As a suggestion for additional perspective, the authors can cite the recent real world example of quantifying CO2 emissions from Europe's largest power plant using satellite observations (https://doi.org/10.3389/frsen.2022.1028240), consistent with the high end of the blue scale in Figure 13.*

  As Figure 13 is about Berlin, we have added the following sentence before the study of the right histogram: "Plumes that we assess are the consequence of a variety of emission levels. For example, Berlin emissions range from approximately 4 to 35 Mt/yr."

  We have provided in the second section of the manuscript the average emissions (plus the standard deviation) for all sources studied. "emission range variability across different locations and times. In Berlin, the average emissions based on the inventory is 16.8 Mt/yr with a standard deviation (std) of 7.2 Mt/yr. In Jänschwalde, the emissions average is 33.3 Mt/yr with a std of 7.7 Mt/yr, while in Boxberg, the average is 19.0 Mt/yr with a std of 4.4 Mt/yr. The Grand Paris emissions average is 20.7 Mt/yr with a std of 9.5 Mt/yr;"

  We have also cited https://doi.org/10.3389/frsen.2022.1028240 in the conclusion.

Thank you very much for all these very clear and helpful remarks.

References

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, https://doi.org/10.5194/amt-12-6695-2019, 2019.