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

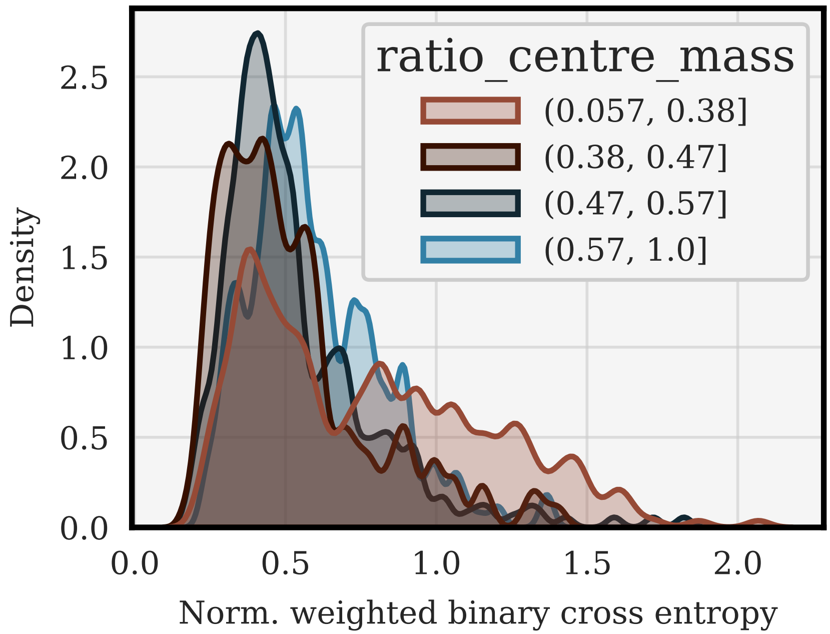

Figure 1. Histograms of the nwbce image scores over all test images of Lippendorf. The plumes are classified in four equivalent clusters according to the ratio centre mass.

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.

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.