Response to Review
Pascal Hedelt

The manuscript „A machine-learning reference dataset for SO2 plumes observed by TROPOMI: uncertainties and emission estimates" from Dougls P. Finch and Paul I. Palmer presents a machine learning approach to detect SO2 plumes in TROPOMI SO2 data using a U-Net image segmentation model.

In my view, the title of the manuscript is somewhat misleading. While it suggests that a "reference dataset" is provided alongside the paper as supplementary material, the dataset must instead be requested directly from the authors. Therefore, the dataset should be provided with the paper or the title should be changed. Moreover, it remains unclear, why this dataset should be considered a "reference", given that the underlying method to detect SO2 pixels lacks clarity in several aspects including the error characterization. It is also not evident how the proposed method is an improvement over the existing SO2 detection flag included in the operational TROPOMI products. The authors state, that the model was trained "with a precise plume mask manually created by the lead author" but provide no further details (see discussion below). In my opinion, the training of the algorithm could be substantially improved by incorporating the operational TROPOMI SO2 detection flag into the training samples.

With respect, we disagree that the title is misleading. The dataset is a reference to what can be detected using machine learning on the TROPOMI data. We have clarified our methods in the updated manuscript to expand on error characterization and model training. Given these revisions and the inclusion of open-access data, we believe the current manuscript title remains appropriate.

We agree that published data, particular a reference dataset, should be available without having to request the data from the authors and that was our intention after our study had been accepted for publication. To address the immediate request, we have uploaded the data to the following repository**:** https://zenodo.org/records/18302024 . These data are intended to be open access.

Although the authors acknowledge funding from ESA, the manuscript does not adequately reference the relevant publications and documentation associated with the ESA TROPOMI SO2 product and its characteristics.

We have included more references to relevant publications and documents associated with the TROPOMI SO2 product including the references mentioned below in this review.

Overall, I believe the paper would benefit from restructuring, rewriting, and improvements to the training approach prior to acceptance, as detailed in the following comments.

We thank the reviewer for their thoughtful comments. Many of the suggestions have been incorporated into the revised manuscript. We would like to emphasise that this paper is intended as a demonstration of what is currently achievable with machine learning approaches and as a foundation dataset for future development. We do not believe that retraining the model is necessary for the present study, and doing so would be unfeasible given the labour-intensive nature of constructing a new training set. Future projects that adapt, retrain, or extend the model would be valuable to the community and represent a natural next step building on the work presented here.

Detailed comments:
Introduction:
This section should already mention the SO2 detection flag in the operational TROPOMI product (this appears only in the Results section 3.3) and should discuss why a image segmentation model could potentially be better in detecting SO2 plumes

We have added the following to the data and methods section:

An existing SO2 detection flag is provided in TROPOMI files, as described in the Sentinel-5P/TROPOMI Algorithm Theoretical Basis Document (Theys et al., 2023), built on a detection algorithm (Brenot et al., 2014). This detection algorithm combines the SO2 observation, solar zenith angle, VCD error and proximity to other detections to assign a flag to each pixel. This flag uses five categories: (0) no enhancement, (1) general SO2 detection, (2) near a known volcano, (3) near a known anthropogenic source, and (4) a potential false positive due to a high solar zenith angle. This study does not include this flagging data in training the model as we want the model to learn SO2 enhancement and plume shapes without relying on existing methods. The flags indicating if the detection are near a known source (flags 2 and 3) also rely on proximity to a known source, which could adversely affect the ability of model to detect plumes from new sources. The method presented in this study does not assign a threshold to a potential plume and tests whether a model can learn to correctly identify plumes from the TROPOMI data without pre-defined limits.

Section 2.1
Although the authors correctly summarize the TROPOMI SO2 L2 product overall, they do not follow standard practices and cite the related publications: Theys et al. 2017) and the TROPOMI SO2 Algorithm Theoretical Baseline Description (ATBD). This section should also describe the SO2 detection algorithm in detail, so the authors can refer to it later in their results and discussions.

We have added the relevant references to this section. The description of the SO2 detection algorithm is addressed in the above comment.

Section 2.2
Lines 99-100: The authors write that they train their model with a "custom database … with a precise plume mask manually created by the lead author." Please add details how the database was generated, how the plume mask was generated, including SO2 thresholds, how many samples are used for the training, which time period was covered, etc.

We have now added an appendix with a detailed description of how we created the training dataset of plumes and plume masks. Along with the appendix we have added:

"These images were sampled randomly from all available TROPOMI files across the full study period and the global domain to minimize the risk of introducing geographical biases".

The number of samples is already provided in this section (just over 1000 images).

Is there a reason, why the authors did not use the operational TROPOMI SO2 detection flag for generating training samples? This would improve the algorithm significantly, especially for weak plumes as well as extended plumes, which the algorithm is currently struggling with.

The purpose of the ML model is to create a system that does not rely on prior knowledge of emission sources. As described in the ATBD, the detection flag uses proximity to known sources to attribute an emission label. While this is very useful in many cases, training the model using these data may lead to the model not capturing any new or unknown sources of SO2. We acknowledge that the detection flag may help the ML algorithm determine whether some detections are true or false (particularly with low SZA), but we believe it would also introduce errors relating to plumes detected in regions away from known sources. As this method for creating the training dataset relies on the judgment of the authors, including more variables (e.g. SZA, albedo & cloud cover) would not necessarily result in a more accurate model as the initial plume judgement may be incorrect.

An error characterization of the image detection model should be included in this section.
We have included the precision and recall for the model in the manuscript. We have now added the F1 score (0.69) for further characterisation.

Figure 2: Please add a colorbar with SO2 VCD values. Does the figure only show 32x32 pixel subimages which contain a detected plume? Consider showing the actual pixel-wise detection mask.

We do not believe a colour bar would add value as the purpose of the figure is to show the model performance in detecting plumes. The absolute value of the observations in this case is not important and adding a colourbar would clutter the plot and reduce the size of the images.

These images are not the 32x32 input images, but the result of the output from the model which can be any size within a swath. The red line shows the boundary for the plume detection (i.e. anything within the boundary is considered the plume). We believe a pixel-wise mask would obscure the images.

Section 2.3
Line 140: How was the mass enhancement calculated? I assume you used the sulfurdioxide_total_vertical_column variable in the main PRODUCT group of the TROPOMI SO2 L2 files... Please note that this VCD is representative for a SO2 pollution source close to the ground. Actual volcanic SO2 VCDs for assume plume heights of 1,7,15km can be found in the DETAILED_RESULTS subgroup of the L2 product and should be used to calculate mass enhancements of volcanic plumes.

This study does use the main product variable and not the 1,7 or 15km products. At the suggestion of another reviewer, we have reduced the emission estimate section down to an example of what could be done with the product as the assumption in the calculation are too large to include in the manuscript and therefore, we no longer need to calculate mass enhancement.

We have also added the following statement:

As this study uses the full VCD, not the 1, 7 or 15km products, we acknowledge that this product is more sensitive to SO_2 pollution nearer the ground and therefore less sensitive than the individual altitude products. As the quality flag limit of 0.5 is for the VCD, observations of SO_2 above the cloud layer (particularly applicable to volcanic emissions) may be missed. The full impact of using the VCD instead of individual layers is not assessed here but could be incorporated into future model developments.

Commented [PP1]: Revise based on my other edit.

Lines 148-152. It is correct that the 10m U and V wind fields only serve as a first estimate, but when you analyze volcanic plumes ,the 10m wind fields are not applicable at all. Here the SO2 layer height in the product (please refer to Hedelt et al. 2019 and Koukouli et al 2021 for details) can be used to interpolate in ERA-5 altitude-resolved wind field data.
We did investigate this but the computing time to attain the ERA-5 wind field data was prohibitive for this study. At the suggestion of another reviewer, we have reduced the emission estimate section down to a single example of a ground level source to reduce the wind-associated errors.

Section 3.1
I am missing a clear statement in this section of what is the lowest SO2 amount the algorithm can detect here.
Now we have removed the section on emission estimates from the manuscript, we no longer have a figure for this. We previously estimated this to be an emissions rate of around 500 kg/hr based on statistical inference.

Lines 184ff: From daily TROPOMI SO2 images and news entries it is clear that the Shiveluch volcanic eruption produced the huge SO2 plume in period “I”, which was detected by the algorithm.
Thank you for alerting us to this eruption. We have included this in Table 1. We have also kept in the analysis over Norilsk as the outflow from the Shiveluch eruption does not reach this area but there remains a spike at the same time, which will contribute to the overall spike in detections. The section has been reworded to reflect this.

Figure 9: The map shown is not from 08 August 2024 but from 24 August 2024. Attached is a map of pixels identified by the operational SO2 detection flag of TROPOMI. It is clear that the operational TROPOMI flag identifies many more pixels as enhanced SO2 compared to the image classification. Therefore, I would suggest that the authors train their model using the operational detection flag to improve their results.
Thank you for pointing out the mistake for the dates; this has now been amended. We have addressed the suggestion of retraining the model in a previous comment.

Section 3.3.
Line 216ff: As described before, the TROPOMI flag should be introduced much earlier in the manuscript together with the corresponding references.
This has now been addressed in the data and methods section.

Figure 11. The distribution of "no enhanced detection" shows many false positive detection scattered around the globe with an enhancement over the mid-latitudes (high cloud cover) as well as many detections at high latitudes (either high SZA and/or high albedo from snow cover). I suggest to refine the model and restrict it to low SZA below about 60-65 degrees, and also take into account additional information in the training (e.g. SZA, cloud cover, surface /cloud albedo)

This would be a valuable line of investihation for future developments to the model but retraining the model and reperforming the analysis is beyond the scope of this paper. This work is also intended as an experiment in what can be achieved by machine learning when only limited additional information is available beyond the observations and the QA flag.

Line 237ff "This analysis demonstrates that while the source labelling provided in the TROPOMI files is informative, it fails to capture the complete picture of emission attribution…" This summary is very questionable and misleading. The operational TROPOMI SO2 flag is flagging pixels with enhanced SO2, with a low threshold of >0.2DU and a distance-based labelling based on known volcanic and anthropogenic locations. It is therefore not at all designed for emission attribution. The same applies to the new plume detection algorithm presented by the authors. The new algorithm does not even indicate the potential source (anthropogenic or volcanic). So therefore the whole statement should be carefully rewritten.

We have rewritten the statement as:

This analysis demonstrates that while the source labelling provided in the TROPOMI files is informative, there is potential for our plume detection algorithm to complement the current flagging through masking and grouping plumes, facilitating analysis into repeated new detections.

MISSING REFERENCES

Theys et al. 2017: https://doi.org/10.5194/amt-10-119-2017
TROPOMI SO2 L2 specifications, dataset, ATBD: https://doi.org/10.5270/S5P-74eidii
Hedelt et al. 2019 https://doi.org/10.5194/amt-12-5503-2019
Koukouli et al 2021 https://acp.copernicus.org/articles/22/5665/2022/acp-22-5665-2022-discussion.html

These have now been included in the manuscript.