Public justification (visible to the public if the article is accepted and published):
Dear authors,

I have had a chance to review your original and revised manuscripts as well as reviewers' comments and your responses for egusphere-2025-5900: A machine-learning reference dataset for SO2 plumes observed by TROPOMI: uncertainties and emission estimates. Thank you for taking the time to respond to each of the comments by reviewers.

I feel that although you may have responded to reviewers' comments, responses for some are not reflected in the revised manuscript. With the comments, reviewers' expectations are expanded reasoning and revisions for each of the comments. It will be helpful if your responses could include specific page and line numbers for the changes made in the manuscript.

Reviewers have expressed concerns with the approach for estimating emissions. Concerns are also with the use of "reference dataset" in the title. Since the emission part is substantially reduced, the title should be modified considering both. With the emission part removed from Section 3, I wonder if Section 2.3 is still relevant.

We have changed the title to:

Towards Automated Near-Real-Time Global Monitoring of Atmospheric SO2 Plumes from Satellite Data using U-Net Segmentation

We have removed Section 2.3.

Below I compile comments from the 5 reviewers that require your revisions in the manuscript in response to those comments. I have also noted reviewers' questions/suggestions that may require your additional analysis to fully address their concerns.

Reviewer #1:
• SC2 - It is understood that the authors trained their model only using images that show the presence of a plume. Why not using even images without a plume for the training?

We have added the following to line 135:
"Unlike classification models (e.g., Finch et al 2022), including images without plumes is not necessary for the training of segmentation models. As we are training the model to be interested in part of an image, the model learns the background of images (i.e. outside the area/object of interest) as the 'no plume' case."

• SC3 - The size of the images used for the segmentation model is such that it would correspond to plume straight paths of < 90 km, for which plume ages may be 1-25 hours, for typical ranges of transport speed. Continuous sources will most likely produce longer plumes, and this may result in the tendency of the method to split big plumes into smaller ones. The size of the images makes also the method very sensitive to "polluted" scenes, i.e. scenes with noisy background, which may be dominant for unsteady plumes, changing winds, or big emission events. What would be the computational cost or other penalty for extending the size of the input images?

We have added the following to line 179:
"We also estimate the length of the plume, based on taking the length of the primary axis of an ellipse fitted to the plume outline. We find the detected plumes have a median plume length of 37.5 km, much shorter than the length of the model input image."

We have added the following to line 139 in regard to the computational cost:
"The computational cost increases slightly more than linearly with image size. Doubling both dimensions typically increases the cost by about four times.."

• SC5 - Source attribution is also a problem with this method. Several sources, especially volcanic, are evidently wrong, e,g, Iztaccíhuatl -> Popocatépetl, Cerro Bravo -> Nevado del Ruiz, Nyiragongo -> Nyiragongo/Nyamuragira, Chimborazo -> Sangay, Ampato -> Sabancaya. I think it is correct to keep this wrong attribution as a consequence of the potential pitfall of the method, but a column to the most likely source should also be provided. The pitfall is possibly originated from big column densities being discarded in the L2 data products of TROPOMI, or from unsteady emission.

We have amended Table 2 to include these corrections where we believe the reviewer was right in their attribution.

• SC7 - All figures presenting maps should include latitude, longitude, time, scale of column density, name of source.

We have now updated figure 2 to also include latitude, longitude, date and scale.

• L79 - What were the filters used for VCD, cloud cover, SZA or number of pixels at the edge of the swath?
We have added the following to line 79:
"No further filtering was applied to the data."

As we refer to the TROPOMI level-2 product user manual and Algorithm Theoretical Basis Document which explain the use of VCD, cloud cover etc in calculating the quality flag, we do not need to include a description in the manuscript.

• L170 - Make sure that pixel size was homogeneous, since the resolution of TROPOMI pixels changed during the period of study. Were these values resampled before their use for training and testing of the algorithm?
We have added the following to line 77:
"(increased from a resolution of 3.5x7 km from the 6th of August, 2019)"
And the following to line 227:
"There is no notable impact on the plume detection results due to the increase in the across-track spatial resolution of the TROPOMI SO2 product on the 6th of August 2-19 from 7.0 km - 5.5 km. There is larger variation in pixel size between the nadir and edge of swath observations than the increase in resolution and the model copes well with all these variations."

Reviewer #2:
• Model Performance: In line 104, the authors say the model's precision and recall are 65.7% and 74%. Does this mean 35% of predictions are incorrect and 26% of cases are missed? I assume this is decent for a U-Net model, but more explanation would help readers who are not familiar with U-Net understand the model's performance. Also, how does this accuracy compare to other plume detection techniques, including the authors' previous work? Providing precision and recall for volcano sources alone would be useful, as performance is likely better for larger sources. This would show that some errors come from the signal-to-noise level of SO2 observations.

The manuscript addresses the first section of this comment on line 146:
"Although these scores indicate the performance of the model, the precision and recall percentages are not directly interpretable as plumes missed or not, as segmentation models work on a pixel per pixel basis. For example, the model may correctly detect a plume in the test dataset but draw a smaller or larger mask than the test data and would therefore be penalised."

We have added the following onto the end of that sentence:
"It is for this reason that we cannot directly compare these performance metrics to classification style models (e.g. Finch et al (2022))
As explained to the reviewer, it is not possible to filter out volcano sources in our model as it would rely on previous estimates of volcanic emission plumes, which would introduce a further uncertainty in the model.

Reviewer #3:
• Which SO2 product version was used? Recently, the TROPOMI SO2 product has

switched to the COBRA algorithm which is more sensitive to weak SO2 emissions. Did the author use this data? If not, why not?
We have added the following to line 74:
"...from version 3 of the offline analysis product"
And the following to line 83:
"A further development to explore in future work is the use of the Covariance-Based Retrieval Algorithm (CORBA) developed by Theys et al (2021). The COBRA retrieval method has been shown to reduce noise and biases in the SO2 column when compared with the DOAS method and would therefore likely impact the plume detection results. The comparison between DOAS and COBRA is not presented here due to data availability at the time of analysis."

• It is not clear to me what is the added-value of the proposed plume detection compared to the detection flag. The selective detection of SO2 from a hyperspectral instrument like TROPOMI is relatively straightforward, and because the SO2 background is negligible it is easy to identify the plumes. The proposed method would perform better for species like CH4 for which the background level is significant.

We have added the following to line 95:
"Our plume detection model is designed to be distinct from the detection flag by not relying on proximity to known sources and looking at plumes as a whole structure, rather than a pixel-by-pixel basis."

We do not think we need to amend the manuscript to discuss whether CH4 would be a better choice as we have already justified why understanding SO2 plumes is important in the introduction.

• l184: about Peak I, if it is related to Norilsk, I don't see why it appears only for this year.

We have written in the revised manuscript that we also don't know why this happens, and although we don't have an explanation, it still warrants inclusion. We have also stated that it is likely at least some of the plumes spotted are from an eruption of the Shiveluch volcano in Russia in April (line 204).

• l195: about the high background SO2 concentrations. Over China, the SO2 levels are quite low. Indeed, there have been many regulations on SO2 emissions in China, and I think this statement is not true. Later, the main author argues on the high SO2 levels in China based on EDGAR inventory but it is not clear to me if EDGAR is up-to-date regarding the SO2 emissions level over China or not.

We have added the following to line 217:

“Although emissions of SO2 have declined in China over the past couple of decades, ambient SO2 levels are still relatively high compared to the rest of the globe (Wang et al., 2025), potentially making this effect more prominent in this region.”
As we’re no longer using the EDGAR inventory in this study, the second half of this comment is now redundant.

Reviewer #4 (Pascal Hedelt):
• 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.

We feel we have answered this question in the original response to the review.

• 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 have now added a colourbar to the images. As explained in the response to the reviewer, a full pixel-wise mask would obscure most of the image, so we have kept with the red boundary for the plume outline.

We have emphasised that the output images are not limited to the 32x32 pixel image size with the following addition to line 170:
“The images in this figure show the varying sizes of the plume images, and that they are not limited to the 32x32 pixel training image size. “

• 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).

As our initial response to the reviewer states, this would be a valuable line of investigation 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.

Reviewer #5 (Alexander Ukhov):
• A minor concern is the reliance on a relatively small set of manually drawn plume masks (1000 plumes) for training. Manual plume labeling is very subjective. One possible workaround is to generate plume masks using a Lagrangian dispersion

model (e.g., FLEXPART-WRF driven by WRF winds). Incomplete background removal can bias plume-based top-down, which is relevant to your discussion of regions where individual plumes cannot be isolated. Question: how does your post-processing handle overlapping or merged plumes from closely spaced sources?

We have added the following to line 133 regarding the size of the training dataset:
"To maximize the effectiveness of training, we augmented this dataset through rotations and flips, yielding a final training pool of over 4,000 images and corresponding masks. This dataset size is sufficient to demonstrate the capabilities of this method."

We have added the following to line 127 regarding the comment on using modelled plumes:
"While it is possible to generate modelled plumes and corresponding masks for training, providing labels and complete background removal, this approach may introduce errors by misrepresenting observed plumes characteristics. Consequently, we restrict training to SO_2 plume data derived from TROPOMI data."

We have added the following to line 130 regarding the question of merged plumes:
"The model is designed to detect single plumes and is not currently able to distinguish between multiple plumes that have merged (e.g. from nearby sources) Training the model to identify and separate merged plumes would require hundreds to thousands of additional examples, with merged plumes subdivided at the labelling stage, thereby introducing further potential for subjective error."

I look forward to receiving your revised manuscript and updated responses.

Sincerely,

Lok Lamsal
AMT Associate Editor