Response to Reviews

Anonymous Reviewer #1

General comments:

The manuscript entitled "A machine-learning reference dataset for SO_2 plumes observed by TROPOMI: uncertainties and emission estimates", co-authored by Douglas P. Finch and Paul I. Palmer, presents a machine-learning method aimed to quantification of SO_2 plumes from point sources. The authors consider this method as a demonstration of the transformative role of machine-learning for validation of emission inventories and for effective environmental management.

My impression is what this study really demonstrates is a novel and efficient algorithm for a first-guess detection of SO_2plumes (less than 15 minutes for the entire globe! according to the authors' estimate). But, it does not demonstrate a reliable quantification of emission. In this sense, I think the title of the manuscript should highlight the method ("machine learning for detection") and not the product ("reference dataset").

Despite this limitation, the approach to detect plumes globally and efficiently is an important contribution as it may provide a quick screening and identification of plumes. This would be particularly important for the current and upcoming geo-stationary satellite missions (GEMS, TEMPO, Sentinel-4), as indicated by the authors. However, the paper only shows its application to TROPOMI, and it is very likely that the application of the method to other missions will require a lot of mission-specific adaptations.

I find the method for estimation of emission strength overly simplified to provide reliable results, and therefore a meaningful comparison with emission inventories is not possible. Moreover, there are evident limitations regarding source attribution, especially for volcanoes, and disregard to recent literature on the "classical" methods for SO_2 emission quantification.

Although the manuscript is easy to follow, the figures need more details and there are some errors, as indicated in the "technical corrections".

As a structural change, if these comments are not addressed, it would be better to skip the emission quantification and comparison with inventory, and concentrate on the plume-identification method.

We thank the reviewer for their comments and have addressed the specific issues below. The primary contribution of this study is to demonstrate an efficient algorithmic approach for rapid processing of EO SO_2 data. We view this as a foundation on which more robust, scalable, and operational algorithms can be built, enabling fast, consistent data handling across diverse sensors and missions. We acknowledge that some work would be needed to adapt this for other satellite missions and consider this to be a starting point or framework for future analysis of up-and-coming missions.

We have taken on board the suggestion for a structural change and have reduced the emission estimate section down to a single example of what could be done, using a low altitude source of SO2 to reduce errors associated with wind fields.

Specific comments:

SC1 - In my opinion, the manuscript's main methodological innovation is the implementation of the U-net architecture for plume segmentation. As such, it would be very informative to provide more details of how the architecture was conceived, e,g. what guides the choice of hyperparameters, e.g. the number of blocks in relation to memory constraints, the type of activation or normalization, etc.

To address this point, we have added the following to the manuscript:

This architecture was based upon the original U-Net designs, described in Ronneberger et al (2015), then manually adapted through iterative tuning to fit this study. We found training the model for 20 epochs sufficient for our relatively simple image problem. We found no further model improvement when it was trained over more epochs. As our training dataset was relatively small (~1000 images), we chose a batch size of 64 to balance over-fitting with training efficiency. The model uses a sigmoid activation as we wanted the model to make a binary pixel-wise classification of within a plume or not.

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?

Unlike classification models, including images without plumes are not necessary for the training of segmentation models. As we are only 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. It would likely do no harm to include images without a plume, and may improve the model, but in this case of making the dataset we decided it was not needed and would introduce further labour costs.

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 agree there are limits to use the 32x32 pixel images as mentioned by the reviewer, and this is evident in the example shown in figure 9. We find the median plume length detected by our model to be 37.8 km, far below the limit of the image size. If the image resolution did artificially split up a plume, we would expect this length to be nearer the 90km image scale. Although we find some evidence of this happening, most plumes do not show lengths consistent with resolution-driven breakup. Instead, the dominant effect is physical fragmentation, where a single plume evolves into multiple distinct structures (as illustrated in figure 9).

It is difficult to get an accurate measure on increased computational costs when increasing the image size although it is often slightly larger than linear. In our case, we may have been able to increase the image size by a few pixels in both directions and been able to roughly maintain the speed of processing, but a large increase (e.g. double in both directions) would have increased the costs by at least 4x.

SC4 - The emission estimation is based on fitting an ellipse to the identified plume, dividing the mass of the masked plume by the major axis of the fitted ellipse and multiplying this ratio by the

wind speed (here taken as the ERA5 10-m wind fields provided with the L2 TROPOMI data products). The metric used to validate the approach is the ratio of the plume area to the fitted area. I method too simplistic, or erroneous for several reasons:

- SC4-a - Quantifying the emission rate as the total mass in the area of the plume divided by the plume length and multiplied by the plume mean velocity is not erroneous by itself and it is used by other established methods. However, the way that the variables are estimated in this study may only result in reasonable emission estimates for "well-behaved" plumes that result from steady emission at stable altitude and with stable winds. In reality, weak plumes (e.g. from industrial sources or passive volcanic degassing at low altitude) may be severely distorted by interaction with the ground, and large plumes may be too big to adapt to the assumptions behind this method. Variable source and wind conditions would results in plumes with heterogeneous shapes. Still the area of these plumes may be similar to the area of a fitted ellipse, without this meaning that the mass to length ratio are similar. I think it would be better to divide the measured mass above background inside the masked plume by an "equivalent plume length". This plume length could be estimated by estimating a mean mass density (kg/m2) inside the fitted ellipse, so that the product of this mass density and the area of the ellipse be equal to the observed mass of the plume. From this relation you could obtained the equivalent plume length (mass divided by pi and by a factor for the eccentricity obtained from the fitting).

This would be an interesting method to help improve the emissions estimates. However, as we are now removing the section from the paper, we will not explore this method in this study.

- SC4-b - The wind speed at 10-m would not be representative for most volcanoes, and providing statistics of this wind field is not informative because the winds at that level will most likely not be correlated with the winds at plume level.

I understand that implementing corrections to varying plume altitude (important to correct column densities and for the choice of wind speed) may be too time-consuming, going against the advantages of this method. Therefore, I conclude that the method, as presented here, may be good at identifying potential plumes, but not on quantifying the emission.

Following the reviewer comments, we have decided to shift the focus of the paper away from the emission estimates and onto the plumes detected: we have removed the emission estimates section and replaced by a simple single example.

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.

These attributions are not part of the method per se. Instead, they are our best guesses where the plume clusters may originate. While we do not believe this mislabelling is “evident”, we appreciate the reviewers’ corrections, and we will amend these in the manuscript. We don’t agree that the incorrect attributions should also be included, which reflect mistakes by the authors not the machine learning model.

SC6 - Given that the manuscript introduces a novel algorithm, it seems important to present in relation to other established methods for plume identification and quantification. There is a clear lack of relevant references, even to the documents describing the TROPOMI products, and to the several approaches used for proper SO_2 quantification of emission (wind rotation, divergence, delta-M, back-trajectory, disk method, etc.)

We have now added the relevant references to the TROPOMI products. As we are no longer focusing on the emission estimates for this paper, we have chosen not to discuss other emission estimation methods as they are no longer relevant.

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

We have amended figure 4 to include latitude and longitude. We do not agree that figure 2 needs this additional information or a scale of column density because the figure is designed as a showcase for the capabilities of the model. The absolute value of the observations, or their location, is not important and adding a colour bar or coordinates/location information would clutter the plot and reduce the size of the images.

Technical corrections:

L19 - SO_2 lifetime is too dependent on environmental conditions to provide a single, representative estimate.

We state in the manuscript that this estimate of lifetime is within a clean troposphere. We have clarified this statement by adding "…although this varies on environmental conditions."

L34 - Provide full name and reference for the EDGAR inventory.

This has been added.

L53 - Reference to Carn et al., 2017 seems misplaced here (Arellano et al., 2021?)

Thank you for pointing this out. The references were in the wrong places and have been corrected.

L57 - Missing reference for the Network for Observation of Volcanic and Atmospheric Change (Galle, Arellano, …)

This has now been added.

L79 - What were the filters used for VCD, cloud cover, SZA or number of pixels at the edge of the swath?

As we use the quality flag provided in the TROPOMI data, this includes VCD precision, cloud cover, SZA etc in its calculation as to what it considers a reliable observation. The full calculation for this can be found in the TROPOMI Algorithm Theoretical Basis Document. We apply no further filtering of pixels in our analysis.

L106-107 - Sentence not justified (see above).

This sentence has been removed.

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 do not find any discrepancy with the change of resolution of the TROPOMI product pre and post 6th of August 2019. As the given resolution for TROPOMI (5.5 km x 3.5 km) is for observations at nadir, the resolution of pixels along the swath increases towards the edges of the swath to around 5.5 km x 14 km. This far outweighs the change in resolution in 2019 for which the model copes well. We ensure all distance calculations are based on latitude and longitude coordinates, not pixel resolution.

L184 - Check reference to "?".

Thank you for spotting this formatting error. This now reads “(Vernier et al., 2024, and references therein)”

L187 - Add year to reference to "Ester et al.".

This has been fixed.

Fig9 - More correct to refer the activity to the Svartsengi volcanic system.

This has been corrected.

L252 - The definition of the coefficient of variation is wrong. If this definition was used, then the conclusions are also wrong.

Thank you for spotting this error in the manuscript. The initial analysis uses the correct definition, and this has now been amended in the text.

L256 - I recommend to compare with the CAMS-GLOB-VOLC dataset, which is based on satellite- and ground-based observations. However, the entire section may just as well be entirely discarded, considering that the emission estimates are too unreliable for a proper comparison.

In line with your suggestion and decided to reduce the emissions section down to a single example of what could potentially be done with this method given more information and computing power.

The reviewer thanks the Editor of AMT for the opportunity to review this manuscript.

Anonymous Reviewer #2

The authors use a U-Net image segmentation model to detect SO2 plumes and measure their emissions from TROPOMI data. The detection part is well-done and creates a valuable database of detected SO2 plumes. The emission calculations, while promising, could benefit

from some refinement. I suggest reducing the focus on emissions unless the wind-related concerns are thoroughly addressed. The manuscript is on the right track for publication after the revisions below.

We thank the reviewer for their comments. We have addressed the suggestions below.

- *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.*

These 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 this to the manuscript.

Comparisons against other models (including our previous work) would be misleading as the scores are not directly comparable. This is particularly the case for our previous work using a classification method (where we could say if the model found a plume or not) but also true if there were other plume detection models as factors such as image size impact the results.

To provide precision and recall for just volcano sources we would have to rely on previous knowledge of volcanic plumes to test the model. This would then introduce another uncertainty on the accuracy of the volcanic plume labelling.

- *Training Data: The training truth is based on manually selected grid cells. While this makes sense due to lack of better options, manual labeling introduces uncertainties. Discussing these uncertainties would be helpful.*

We added the following statement in the manuscript: “Because the training dataset was created manually and relied on the authors’ judgement of what constitutes a plume, the model cannot outperform the quality of this dataset. Any biases or recurring misclassifications present in the training examples may be learned by the model and subsequently propagated into the final results. An alternative approach would be to augment the training dataset with model-simulated SO_2 plumes, but for this study we chose to rely exclusively on real data to maximise the diversity of scenes and plume morphologies that may not be captured in model simulations.”

- *Emission Calculation: Volcanic SO2 emissions have been estimated by Carn et al. and Vitali et al. More details can be found here: https://so2.gsfc.nasa.gov/. How is your method different? My main concern is the use of 10-meter U and V wind fields. As pointed out by the authors themselves, near-surface winds are not suitable for volcano plumes. While concerns about computation cost are valid, using winds at the correct height is essential for emission estimates. I recommend a sensitivity study to show the impact is limited, at least for one case. Otherwise, the emission estimates*

are less reliable. Comparing volcanic emissions with estimates by Carn et al. would also be strongly recommended.

As suggest by other reviewers, we have decided to reduce the emission estimate section down to one example of what could be done and improved on in future iterations. If these improvements are made in future work, then we will compare estimates with Carn et al.

- *Data Availability: I suggest the authors make the plume database publicly available to support further research.*

We have uploaded the data to the following repository**:** https://zenodo.org/records/18302024 . This has been added to the manuscript. These data are open access.

Anonymous Reviewer #3

Finch et al. presented an SO2 data set based on machine learning identification of plumes and emission estimation from TROPOMI. It is an interesting study, and I believe the work could be published after revision. The paper is well written and structured. Overall, the figures are of sufficient quality.

My main reservation is that the SO2 fluxes derived are not evaluated against other available estimates. Over the last years, several studies have reported SO2 top-down estimates using TROPOMI, but the main author does not cite those papers. In particular, Fioletov et al. 2023 (https://doi.org/10.5194/essd-15-75-2023) provides SO2 emission estimates and I would like to see a comparison between the emissions from this work and the results from Fioletov. This can be done for several representative SO2 sources (anthropogenic and volcanic), with stable emissions. This should come with a discussion of pro and cons for the presented method. Apart from that, I agree with all the comments raised by Pascal Hedelt (some are repeated below) and the author should address them in the replies and revised manuscript.

We thank the reviewer for their comments and have addressed their suggestions below.

Introduction

Line 33: "… with recent years showing lower values". Please add a reference.

We have moved the reference to Soulie et al (2023) to the end of the sentence to include this statement.

Line 57: NOVAC. Please add a reference to Galle et al. 2010 https://doi.org/10.1029/2009JD011823

This has been added.

Methodology: section 2.1

-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 used version 3 of the offline TROPOMI data, using the DOAS retrieval for the SO_2 product. We did not use the COBRA retrieval for this paper because, at the time of data download and processing, the full COBRA dataset was not yet available. Although the complete dataset has since become been released and published evaluaions indicate that it performs well, incorporating it here would require re-processing the entire analysis at a computational cost beyond the scope of this paper.

-In line with P. Hedelt comment, more information must be given on which SO2 column product was used (the main VCD and/or the 1,7,15km VCD product) and what is the impact of this choice on the final result.

This is addressed in the response two points below this one.

-A reference to the main papers, ATBD and product read me file should be added.

These have now been added to the manuscript.

-The quality flag >0.5 applies to the main VCD product which assumes an SO2 profile from pollution. This flag removes much of the cloudy pixels which are still very useful for volcanic events where the SO2 plume lies over clouds (in this case, the 1,7,15km VCD product are more appropriate). This is not assessed or discussed in the paper.

We have added the following statement:

Because this study uses the full vertical column density (VCD) rather than the 1, 7 or 15km layer products, we acknowledge that some potential SO_2 plumes located above cloud tops may be missed when applying a quality-flag threshold of 0.5. The broader implications of using the full VCD instead of layer-specific products are not assessed here, but this could be incorporated into future model developments and retrieval-selection strategies.

Methodology: section 2.2

-line 100: in line with P. Hedelt, the 'manual creation of a precise plume mask' deserves a thorough description.

We have now added a new appendix that includes a detailed description of how we created the training dataset of plumes and plume masks.

-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.

This method defines the outlines of plumes whereas the detection flag is on a per pixel basis. Further analysis is needed with the detection flag to determine boundary thresholds or distinct plumes. We acknowledge that the detection flag does a good job a lot of the time and is

adequate for many uses but do not consider this a reason not to try and develop other methods of plume detection.

The ESA EOPlumes project, from which this study originates, included parallel analyses for NO_2 and CH_4. We chose to publish the SO_2 component first because it was the most mature and provided the clearest demonstration of how machine learning methods can accelerate the processing of EO data.

Methodology: section 2.3

-Line 35: typo: "To estimate the the emission"

This has been corrected.

-it would be informative to compare the elipse main axis direction with the wind direction used to estimate the SO2 emissions. Do they compare well?

We did try this and the results were mixed. As with the other issues with the emission calculations, the wind is likely not representative if the plume is at a high altitude. At the request of another reviewer, we have reduced the section on emissions down to a small example of what could potentially be done so the issues with the wind are no longer relevant to this study.

Section 3

-line 183: typo – a question mark appears in (Vernier et al., 2024;?).

Thank you for spotting this formatting error. This now reads "(Vernier et al., 2024, and references therein)"

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

We agree it is unusual behaviour and we have been unable to find a clear reason for this, but we consider it interesting enough to include in this study. Any data on change in production in the area, which is the most likely cause, is unavailable to us.

It has also been pointed out that the spike in detections also corresponds to an eruption in eastern Russia which has now been included in the manuscript.

-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.

While there have been many successful regulations implemented in China, and SO_2 levels have reduced dramatically in the past two decades, the levels of SO_2 are still higher than compared with many other countries. We agree the EDGAR inventory may not be up to date, we were unable to find any other suitable inventory that covers the study period. We do state in the paper that this is our hypothesis and not drawn from thorough analysis.

-section 3.5: the presentation of the emission database is minimal. It consists mainly of Figure 12 which is not very informative. I would like to see at least maps of emissions (global, regional, per emission type, etc). The comparison with EDGAR is also weak in my opinion. The author mainly describes why it is presumably not possible to compare. As a reader, it is not clear to me what this section is about.

At the request of another reviewer, we have reduced this section down to an example of an emission calculation to show what could potentially be done as a next step in this project.

Conclusions

The last sentence about the usefulness of the approach for the VAACs is doubtful. A simple VCD threshold mask (or detection flag) is enough to isolate the plumes.

We acknowledge that the final sentence uses some hyperbole and have rewritten it emphasise the potential usefulness of this product.

A threshold mask would be on a per pixel basis and therefore need further analysis to determine clusters of pixels and what constitutes a single plume or background noise. This method has been designed to try and eliminate the need for that step on every detection and provide a quick and simple assessment of a plume.

Data availability

The dataset should be available as supplement or in a data repository

We have uploaded the data to the following repository**:** https://zenodo.org/records/18302024 . This has been added to the manuscript. These data are open access.

Acknowledgements

Please provide information on the ESA project supporting this work.

We have added the name of the project (EOPlumes) into the acknowledgements.

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.

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.

Alexander Ukhov

I enjoyed reading this manuscript. The construction of a large SO2 plume database using a U-Net segmentation approach has value for the Community.

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).

We used FLEXPART-WRF in our recent work (Ukhov et al., 2025, JGR Atmospheres, https://doi.org/10.1029/2025JD043334) for SO2 point sources and emphasized that dense source clusters

and 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? For example, how do you avoid double-counting in the emission-rate estimate?

We thank the reviewer for their comments on our paper.

We agree that plume labelling is subjective and does result in the model only being as accurate as the person labelling the training data. However, we decided to use real data instead of modelled data to try and ensure we capture a full range of plume shapes and sizes that may not be represented in a modelled dataset. Including modelled plumes as

well as observational data in the training dataset would likely enhance the model performance but is outside the scope of this paper.

We find 1000 images (augmented to 4000 images) was sufficient to get our desired result for this project and paper. More iterations of training with more images will improve the model but with diminishing returns. As manually labelling plumes is labour intensive, we determined the 1000 would be enough to showcase the capabilities of this method.

A limitation of the machine learning model is that it cannot differentiate between merged plumes. For the model to be able to do this, it would need many (100s-1000s) examples of merged plumes which we do not currently have. With our method it would also require splitting the merged plume into two (or more) at the labelling stage which would introduce errors. As you rightly point out, another option would be to use modelled plumes, which could be used to train a model to detect and correctly identify merged plumes. This is currently outside the scope of this paper but would be a very interesting expansion to the project.

If the current model mislabelled two merged plumes as one large plume, it would not double count the emissions but calculate the total SO_2 from all merged sources. It is worth noting that the emission estimate would likely be inaccurate as a merge plume is unlikely to be a standard plume shape (e.g. gaussian distribution in broadly one direction).