Response to Review
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.