Review of Yuan et al. (2023) (Reviewer #1)

Yuan et al. demonstrate a novel machine-learning model for the detection of ship-emitted NO2 plumes using input data from the TROPOMI satellite instrument. The manuscript explains the relevance of such a model, its design, and that the model performs well on test data. The text is mostly well structured, but some sections require more in-depth explanations or restructuring. Particularly section 4 („Discussion and conclusions") suffers from shortcomings, as it lacks critical reflection on the presented model. Language- wise, the quality of the text is average, but can easily be improved e.g. by inspection of the technical comments below. The scientific value of the manuscript is adequate for publication, and it fits well into the scope of AMT.

General remarks:

I find the scientific concept behind this manuscript convincing. It is expected that a convolutional neural network (CNN) generally outperforms analytic methods of image processing, and Yuan et al. demonstrate that their model works as intended. I agree with the choice of the method and (with some uncertainties relating to text interpretation) could not identify any methodological mistakes.

However, I have concerns regarding the text of the manuscript. I believe the manuscript should be more detailed in certain places (especially section 4). Some descriptions are very brief and allow for multiple different interpretations, which makes it hard for the reader to identify possible mistakes on the authors' side. I further disagree with some statements made by the authors. I have attached a full list of related specific comments below. Although I consider this a major revision, I believe that it only requires text work, and that the scientific content of the paper can remain as is.

Specific comments:

l. 13: This is not explicitly demonstrated in the manuscript. Consider rephrasing to „(...) could be useful (...)".

Changed.

l. 18-19: The sentence „However, their magnitudes still have large uncertainties" insinuates that this problem is addressed at some point in the manuscript, which is not the case. The proposed model can localize shipping routes, but not quantify ship emissions.

Modified to clarify that this point is just a motivation and does not imply we are directly addressing it here.

l. 88-89: If the authors wish to describe how the NO2 SCD is retrieved, they should mention the term Differential Optical Absorption Spectroscopy (DOAS). The term does occur in l. 275, but should be mentioned here instead.

Changed.

l. 93-94: The statement on robustness wrt. different retrievals should be reconsidered. The authors describe later on, that there is a sensitivity of their model to cloud cover and the stratosphere-troposphere separation, which are both aspects of the satellite retrieval. The authors can neither know whether their model will work well with future retrievals, nor is the proclaimed robustness towards different existing retrievals shown in the paper. This statement is also conflicting with l. 26-27.

We modified the statement to clarify what we meant. The reviewer is correct that such robustness is not valid under all circumstances.

Fig. 1: Consider expressing NO_2 VCDs in units of molecules cm^{-2}, which to my knowledge is the standard unit in the satellite remote sensing community.

Changed.

Fig. 1: Consider moving Fig. 1 to section 3, because its right subplot is already a demonstration of the model.

We want to keep it this way because Figure 1 serves two purposes. One is visualizing data and the other is showing its corresponding model output on a map.

l. 111: Consider changing the section title to „Model design" or something similar.

Accepted and changed.

l. 112-121: In my opinion, it is not necessary to explain, what an encoder-decoder (ED) model is, and how it works. Consider giving a reference to a more extensive description in literature instead. On the other hand, the authors should explain, why an ED is required in the first place. The usual purpose of an encoder is feature extraction and/or dimensionality reduction. However, a CNN is essentially a feature extraction mechanism itself. Without further insight, it is not clear why the important features can not be extracted by the CNN alone. The output mask shown in Fig. 2 looks like a typical example of binary image segmentation, which is usually achieved without an additional ED. Please elucidate the purpose of the ED. Ideally, this explanation should extend to the rest of the model, because the text is quite unclear on the purpose of the individual model components.

We added more discussions on this topic as suggested by the reviewer in this section. CNN is used here (ResNet and the dilated convolutional layers). They are building blocks of our model.

l. 122-131: Extending on the previous comment, the authors should explain, how they ended up with this particular model architecture. Why are exactly three dilated convolutional layers with these specific dilation rates used? How did the authors identify ResNet-50 as a good base-model among the many alternatives found in literature? Without further information, it must be assumed that this model design was developed (partly) via trial and error. From a methodological standpoint this is fine, but it should be mentioned. If a systematic

hyperparameter search was conducted instead, it would be sufficient to mention that. If not, it bears strong potential for improvements in the future and should be mentioned as an outlook.

This is a good point. We added more discussions on this topic as well. We cannot afford a full hyperparameter search. Instead, we adopt the best practices and known values that work in the literature, and only carried out a few runs that changed the depth of our network, the learning rate, and the number of convolution layers at each level.

l. 126: It is unclear, whether some form of interpolation to a regular grid takes place before feeding the data to the neural network. Please elaborate.

We used the native data and did not do interpolation on them.

l. 149: The explanations in l. 118 describe the use of the L1 norm as the loss function, not the cross-entropy. If the authors are using a mixed loss function, they should elaborate further.

Thank you for spotting this. We clarify that we use || || denote a loss function, not specifically L1.

l. 154: The formula for the IOU is not rigorous, because a fraction of two sets is undefined. I assume the authors are referring to the cardinality of the sets. Furthermore, please evaluate the IOU on the training set as well.

l. 172-177: I disagree with this statement. The fact that binary classifiers are noisy around their classification threshold is a standard observation. This alone would explain the fine-scale variations described by the authors. The idea that „The trained model, however, appears to understand the essence of what defines a plume and can trace out fine-scale variations by itself" can not be supported without further, detailed discussion, and I would highly recommend to drop this notion altogether. In particular, there is no explanation from where the neural network is supposed to learn a better definition of a plume, than from what is present in the training set in the form of the manual labels.

The reviewer has a valid point and we have modified this discussion and added clarification on what we want to convey here.

l. 176-177: I disagree with this statement. Model generalization refers to the ability of a model to maintain its prediction quality when confronted with new instances absent in the training set. What the authors describe is something conceptually different, namely the model's (supposed) ability to exceed the performance of the data labelling process (here: a human expert). I would recommend removing this statement entirely, also due to the issues raised in the previous comment. If the authors wish to emphasize good model generalization, they can do so easily with the standard method of comparing the IOU on the training and test set or unseen data.

The reviewer is right. We have changed the text to reflect this.

l. 193: Why was the filtering of land pixels not applied to Fig. 3? It is pleasant to see that the model produces no false-positives over land, but this irrelevant if land-data can be filtered out at any given time.

We did filter out land data include those used for Figure 3.

l. 209: Here, a section begins where the authors apply their model to TROPOMI data of the year 2019. At the same time, l. 91-92 suggest that TROPOMI data of the year 2019 was also used for training the model. It is not explained anywhere, by how much the training data and the data used in sect. 3.3 intersect. Ideally, with regards to evaluating the model's capability to generalize, they should not intersect at all.

Training data is from 2019, but only covers the Indian Ocean area and the evaluation is done on data that are not used for training and from other areas. This is now clarified in the text. Thank you for raising this point.

l. 212: If the plume mask resolution is really 1° x 1°, then it is far lower than the spatial resolution of TROPOMI. In that case, the authors should elucidate, why they chose not to train the model to produce the plume mask on a similar resolution as TROPOMI.

Our model is run on native resolution of TROPOMI. The model's prediction results are aggregated into 1x1 data for plotting.

Fig. 4: Please explain, why the MODIS cloud data is used instead of TROPOMI cloud data.

We were not aware of TROPOMI's cloud data. Also, MODIS's resolution is 250m at the highest and its cloud product has spatial resolution of 1km. It is much better suited for cloud detection. Please also note that this is annual mean data.

Fig. 4: The upper middle plot is labelled „NO2 Emission" but has the units of a column density, while in the figure caption, the authors refer to it as „NO2 concentration".

Fixed.

Fig. 4: The figure shows mean cloud fractions of > 0.5. According to the TROPOMI PUM, all pixels with cloud fraction > 0.5 are removed if f_{qa} > 0.75. This is conflicting with l. 95.

The cloud map is shown for annual mean. For a grid whose annual mean cloud fraction is greater than 0.5, it is possible that its CF is lower than the TROPOMI retrieval threshold, which is why there are NO2 retrievals for these grids and therefore our model ourputs.

l. 231-233: I disagree with this statement. What the authors describe here is not a demonstration of good model performance, but rather of the model's limitations. It is expected, that the neural network does not identify any ship plumes under cloud contamination. This is not because it somehow internalized „physical

sense", but because there is simply no input signal. The fact that the model does not work under cloud contamination requires no justification on behalf of the authors, but it can not be declared a beneficial feature of their model.

We remove rear half the sentence. Here the physical sense refers to what the reviewer said here: these two should be anticorrelated due to the blocking of signals by clouds.

l. 266-271: This text section is essentially an outlook, and it is not suitable to start a discussions and conclusions section. Consider moving it to the end of the section. Furthermore, consider beginning section 4 with a brief summary of the new model, similar to l. 287-293.

l. 268-271: I would recommend rephrasing these statements as ideas for the future, but not to claim them as certain until tested.

Thanks for the suggestions. We rearranged this section considering the comments here.

l. 273-285: This explanation of the satellite retrieval should occur much earlier (e.g. in section 2.1). Moreover, it is unclear, how the authors come to the conclusion, that the stratosphere-troposphere separation is a critical issue. The fact that the shipping routes can be easily labeled by a human expert indicates the contrary. Furthermore, CNNs can be surprisingly robust (through data augmentation) to this kind of noise. If the authors have examples of orbits which support their claim, then these should be showcased, e.g. in section 3.

Good suggestion on moving the discussion forward. On the point of stratospheric component, it is only talking about the retrieval process. The point is that slight errors in stratospheric component can severely impact the troposphere retrieval, which can mask out plume signals or make the signal less clear. It is not affecting the model capability per se. We added a sentence to clarify this.

l. 289-290: See my comment above, regarding the possibility of filtering out land-pixels in general.

*l. 293: The authors use the words „high" and „low" cloud cover in an ambiguous way. l. 202 reveals that „low cloud cover" means „low **altitude** cloud cover". In l. 193, „high cloud cover" apparently means „a large amount of clouds".*

Fixed.

Section 4 is generally too brief, and lacks critical reflection on the proposed model. Besides the discussion of the stratosphere-troposphere separation (which was mentioned here for the first time), no other weaknesses or ideas for improvements of the model are adequately discussed. I urgently recommend to extend section 4, and I have listed some suggestions below:

- *Consider discussing the possibility of improving the model by hyperparameter optimization. This is a standard ingredient of any mature machine learning model, but has not been mentioned throughout the paper.*

- *The model in its current form leaves room for improvement by inclusion of further input variables, such as cloud fraction, water body classification, or meteorological data.*
- *Human labelling of training data is usually considered unproblematic, if there is no reason to doubt the proficiency of the human expert (e.g. when labeling everyday-objects in photographic images). The ship plumes, however, are a bit more ambiguous. The human expert must make an educated* guess*, and the uncertainty of this guess is implicitly adapted by the model. Consider discussing this topic.*
- *Consider discussing the relatively low resolution of the ship plume mask produced by the model. It is an obvious limitation, for which no reasoning or discussion was provided.*

Excellent suggestions. We have added discussions on some of these points. The last point stems from a confusion about data resolution as discussed above. We use native TROPOMI data, i.e., the highest possible resolution.

Technical comments:

- *The spelling of NO2 (NO_2) and NOx (NO_x) is inconsistent throughout the paper, including the figures, the „Key points" section, and the „Plain Language Summary".*

Fixed

- *There is a general inconsistency of grammatical tense within the individual sections of the paper.*

Fixed

- *Many figures of this paper lack proper labelling (e.g. (a), (b), (c), ...)*

*l. 17: Change to „Ship emission are **an** important contributor", and replace „interacting" by „interaction". l. 23: Change to „block signals from **reaching** the sensor".*
l. 23: Remove „Indeed,".

l. 23: Change „complements" to „complement".
l. 35: „over the global" is erroneous.
*l. 42: Change to „exerting an importing influence on **the** marine boundary layer".*

l. 46: The „, features" seems to be unintended.

*l. 49: Change to „and **are** the dominant anthropogenic source (...)". l. 51: Change to „to identify **ship emissions** will complement".*
l. 63: Change „ground resolution" to „spatial resolution".
l. 66-67: The latter sentence is grammatically incorrect.

l. 73: Change „uses" to „use".
l. 76: Rephrase „in other words", because it is colloquial.
*l. 87: Change to „irradiance data with **a spectral resolution** of".*

l. 89: „ground footprint“ is not a suitable term. Consider replacing with „pixel size“.
l. 90: Change „nadir“ to „nadir viewing geometry“.
l. 92: Change „products“ to „product“.
*l. 94: Change to „demonstrate the applicability **of** our method“.*
l. 112-114: The second sentence of this section has a somewhat colloquial character.
l. 115: Change „encoded“ to „encodes them“.
l. 115-116: If FW represents the neural network, then Y = FW(X). The bias term is part of the neural network forward pass.
l. 117: The term „features“ usually represents the input (not the output) of a neural network (here, X would be the features).
l. 117: See previous comment regarding the notation of the bias term.
l. 120: The formulae for the optimal configurations are not displayed correctly.
l. 128: Change „lever“ to „level“.
l. 129: Change „upsampled by 4 times“ to „upsampled by a factor 4“ or something similar.
l. 139: Change „Training data contains“ to „The training data contains“.
l. 139: Consider replacing „blocks“ with „images“.
l. 168: The text goes from the first to third to second example. Consider changing the order.
l. 183: Change „First example“ to „The first example“.
*l. 200: Change to „Applying the model **to** one year of TROPOMI **data**“.*
Fig. 4: Change „NO2 Pixel Frequency“ to „Plume pixel frequency“.
l. 209: Change „TROPOMI data during 2019“ to „TROPOMI data of the year 2019“.
l. 216: Change „the east Asia“ to „east Asia“.
l. 225: Change „is“ to „are“.
l. 238: Change „NO2 density“ to „plume density“.
l. 240: Change „for“ to „of“ or „in“.
Fig. 5: Change y-label to „Number of plume mask pixels“.
l. 257: Replace „cloud situation“ with „cloud contamination“.
l. 280: Start the sentence with „The“.
*l. 281: Make the notation of „1.0 deg * 1.0 deg“ consistent with l. 212.*

We thank the reviewer for the careful job. All are fixed