Response to Referee Comment 1 (RC1) from Anonymous Referee # 2 (*Referee comment in Italic*, response in blue).

RC1: 'Comment on amt-2022-135', Anonymous Referee #2
Li et al. presented an new machine learning scheme and applied it to OMI SO2. It is an interesting preliminary study and the work should be published after minor revision. The paper is well written and structured. Overall, the figures are not of sufficient resolution.

We thank the referee for the review and for raising several important points. Below please find our point-to-point response. As for the figures, we have made changes to Figure 7 to make it more readable. We will also provide high-resolution files to AMT for the final production, if our paper is accepted.

Main comment: *although it is an interesting study, I am concerned by the fact that the NN function is heavily weighted by SRR. The SRR, as defined by the author, actually contains the desired result. Therefore, it is not a surprise that this works. The only real task of the NN is to resolve the RMS dependence. Also, the fact that the noise gets reduced is in a way artificial, as it is the direct result of the constrain on the clean pixels that should be SCD = zero. The reduction of the bias poses also the problem of a possible overcorrection. On long-term averages, are the weak emissions sources detected by OMI PCA still present in the NN data set?*

In my opinion, what would be much stronger is to train the NN directly with the corresponding radiances. This is done to some extend (through a PCA transformation) but it is coming at the end of the paper. It is pity it is not put more in front.

We agree that the NNs are heavily weighted by SRR. At the same time, based on the correlation coefficient and RMSE in Table 1, a model based on just SRR will produce fairly large errors. Also as shown in sect. 4.3 and Figure 9, a more complex model based on SRR and the additional spatial context (monthly mean SRR) still underestimates SO_2 over polluted areas. In that sense, we feel that the NN and the additional variables have certainly helped to improve the results.

We also agree that assigning zero SCDs to clean pixels is a fairly strong constraint, which is to some extent out of necessity given the dearth of measurements over vast remote regions. In a way, our analysis method can be viewed as a more advanced version of the Pacific sector correction (PSC), a quite common and well-established practice to reduce retrieval artifacts for species such as SO_2 and HCHO. In the PSC approach, retrievals over the remote Pacific are averaged as a function of latitude. The latitude-dependent mean VCDs or SCDs (or the differences between retrievals and model simulations) over the remote areas are considered artifacts and subtracted from retrievals for all pixels, regardless of their location. Our approach, on the other hand, considers more factors for each individua pixel. We have added this point to the revised paper.

As for the overcorrection, we have taken caution (mainly through the test of different setups for the pixel classification scheme) to reduce it. The test on the annual emission estimates in the paper shows

that the weaker sources are still present in the annually averaged NN analyzed data. We have added the potential overcorrection as a caveat in the revised paper.

Finally, we agree that training the NNs to do retrievals from radiances would be interesting and valuable. But at this point there is still a lot of room for improvement as clearly shown in sect. 4.4 and Figure 10 (sect. 4.5 and Figure 11 in the revised paper). We plan to conduct follow-up studies to explore ways to further improve the NN-based retrievals from radiances. Also the lack of high-quality training data is a major obstacle in training NNs for retrievals. And in this sense, our current study contributes to such efforts, by providing training data with improved quality (as compared with the original retrievals). We have added this point to the revised paper.

Minor comments

L55: I agree with the necessity to improve the retrievals but is a 10% noise increase a real problem for addressing long-term trend monitoring? I do not think so. The appearance of instrument issue like row anomaly is more of a problem.

The point we try to make here is that with reduced emissions (and signals), the increase in noise makes the long-term monitoring more challenging, especially if the signal is relatively weak to begin with. We have clarified this point in the revised text.

Figure 2b is confusing. Is the SRR unit less? The SCD is DU and RMS has no unit, thus SRR should be expressed in DU (?).

Thank you for pointing this out. We have made the correction to Figure 2b in the revised paper.

Section 2.2. The classification of pixels is quite complex. As explained in the text, the parameters used (a1,a2) have been determined by testing. However, it would be good to illustrate the impact of the (a1,a2) settings on the final results. Currently, it is hard to judge if the complexity is worth, compared to a more simple classification.

We started from a simpler classification scheme (with fixed a1 and a2) but found some deficiencies. For example, we found signs of overcorrection (negative bias) for sources at lower latitudes. We also found relatively large positive bias for high latitudes. This is why we moved to the more complex scheme. We have added this discussion to the revised paper.

Per your suggestion and also the comment from Referee #1, we have conducted additional sensitivity tests by altering a1 and a2 by +/-10%. Overall, the final results show some (but not overly large) sensitivity to the a1 and a2 parameters. We have added these test results and the discussion to the revised paper.

Section 2.3: the processing is not performed separately for each row. Why not? Would it improve/degrade the results?

The main reason is to ensure a large enough sample of polluted pixels in the training data. On average, each row would only have ~200 polluted pixels per day. In the PCA-based retrievals, we process each row separately because each has different measurement characteristics (e.g., wavelengths, instrument spectral response function). After the PCA SCD retrievals, the row-to-row difference in SCDs and SRRs are relatively small (see figure below). And we do not expect major changes in the training results if we process the data separately for each row.

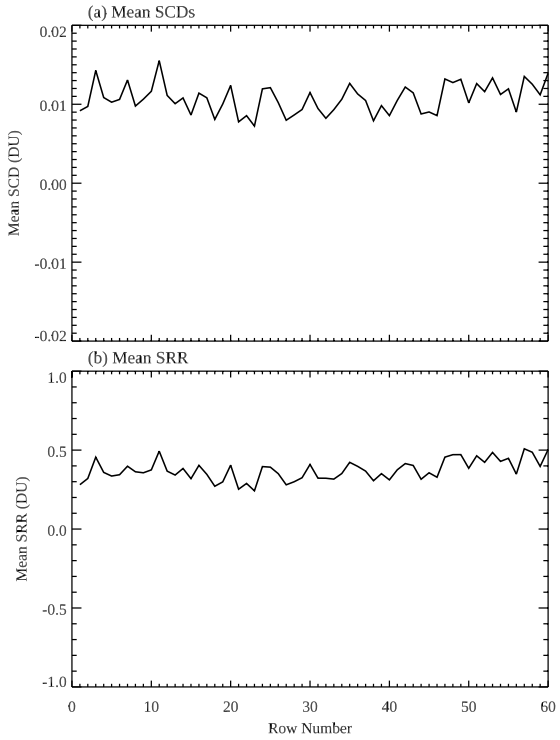

Figure. Mean SO_2 SCDs and SRRs for each of the 60 OMI rows for April 16, 2005, calculated from pixels with SZA < 70, outside of the SAA areas, and outside of polluted regions (monthly mean SRR < 3).

Section 3.1: the SCDs over SAA are much better in the NN analysis, which is surprising. Any idea why?

One potential reason is that retrievals over SAA areas tend to have relatively large fitting residuals (and RMS). The use of SRR partially cancels out the relatively noisy SCDs. We have added this point to the revised paper.

Figure 7: the figure quality is not sufficient. When zooming over the subplots it is hard to see the emission patterns described in the text.

Per your comment and also the suggestion from Referee #1, we have revised Figure 7 to include fewer subplots. We have also moved some subplots (as separate figures for each SRR range) to the supplemental information. The text has also been updated accordingly.

Figure 10b : the PCA-NN results seem to show striping features, although the analysis is performed separately for each row. Why?

The NN models are trained separately for each row but using the same architecture, and the training performance varies for different rows. The reason for this varying performance is unknown to us at this point and will be the subject of further investigations. The training performance may improve if the architecture is optimized for each row, but this will require substantial effort and is probably more suitable for a follow-up study.