Response to Comments by Anonymous Referee #1

We thank the reviewer for taking the time to provide feedback and comments on this manuscript. There were many insightful comments concerning the machine learning aspect of this study. We agree that we can conduct more in-depth investigation on the sensitivity analyses and NN parameters. As these details could be of interest to many readers, we aim to take these comments into account and add descriptions of the machine learning into the manuscript along with some into the supplemental material. There are also some good suggestions of potential improvements in the future that we feel may take a lot more additional time to perform and would be more appropriate for future follow-up studies. The reviewer comments are highlighted in bold with our responses written below. We also state the changes that will be made in the revised manuscript.

1) **Line 194-195: How large is the sample size large enough? I think it would be good to have the number of trainable parameters in the DNNs reported here.**

To check sensitivity to sample size we performed a simple test in reducing the number of training samples by different amounts (e.g. 10%, 20%, 50%) and seeing how this affects test set prediction. While there was a noticeable effect after 50% reduction where the absolute error went up by around 0.3 km, when reducing by 10% the error stayed the same at around 1 km. The RMSE also goes up after dropping more than 10% of data. Given there's very little change when reducing number of samples by 10%, we could expect to see diminishing returns in retrieval quality with more samples.

% withheld	0	10	20	30	40	50
Mean Abs Difference	0.95	0.98	1.02	1.08	1.12	1.24
RMSE	1.46	1.45	1.62	1.69	1.79	2.00

We have added this discussion to the revised paper.

2) **Line 244-246: Have you tried to increase the number of neurons in each layer of the DNNs or increase the number of hidden layers in the architecture? A sensitivity analysis during a "fine-tuning" process is very helpful for the optimization of DNNs.**

Yes, we have tried to changing the number of neurons and the number of hidden layers and looking at sensitivities in the results. As expected, less complex NN structure led to more stability in the NN in the application to real data. Using larger amounts of neurons sometimes added unrealistic spikes in the results. The TROPOMI retrieval (Hedelt et. al, 2019) also had a two layer NN structure with 32 and 10 neurons. We have added a sensitivity analysis to the revised paper. In the future, we plan to attempt to further optimize and fine tune the NNs in the retrieval setup.

3) **Line 233-234: The tanh activation function is more frequently used in classification problems. For regressions, the rectified linear unit (ReLU) and parametric ReLU (PReLU) are more used.**

Thank you for pointing this out. We tested ReLU as an activation function, but the tanh seemed to perform better. We demonstrated this with an additional sensitivity analysis (below) which shows that tanh produces less error between the trained NN and test data set. The errors were determined by taking average of 5 training runs since training can slightly vary each time and the process was repeated for multiple datasets (OMI row based). In general, even though the ReLU does not reduce the performance by a lot, in our case using it does not improve the results. PReLU also did not improve the mean error. Note that for this analysis the random seeds and test data were held constant.

Mean absolute error (in km) between test set and predicted outputs by trained NN for different activation functions.

	Row 2	Row 9	Row 18	Row 27
tanh	0.968	0.943	1.035	1.036
ReLU	1.168	1.1135	1.154	1.174
PReLU	1.097	1.064	1.094	1.107

4) **Line 217: How do you control overfitting? Do you have a validation set? A validation set should be a small subset sampled randomly from the training set, over which the performance of the DNNs is checked after each epoch of the training. Sometimes the training loss gets reduced, but validation loss is not.**

- For overfitting, we chose to use L2 regularization. This has now been made clear in the revised manuscript.
- During the NN training the dataset is further split into a test and validation set with a 0.9/0.1 split. This validation set is different from the independent "test" set that is withheld from training. We added this information to the manuscript.
- Lastly, we did make sure that the training loss of the validation set decreased at the same rate as the training set and did not increase, in order to avoid overfitting. The training was stopped when validation loss was relatively constant for at least 30 epochs.

5) **Figure 3: Relative errors are much more meaningful here because the "truth" of volcanic SO2 layer heights is not a constant set. It is also clear that the predictability of volcanic SO2 layer heights has a strong correlation with SO2 column information. Since OMI is biased, have you tried to add SO2/O3 column information from other sources as predictors?**

This is an interesting idea and we have not attempted it thus far. The difficulty in using column amounts from other instruments is that other instruments (TROPOMI, OMPS, etc.) have different overpass times, spatial resolutions and number of cross track positions. The application phase retrievals are done on a row/pixel basis, meaning each input sample includes the radiances

+ parameters for that given pixel. The difference in resolutions and overpass times would make it tricky to be included with OMI data in terms of data processing. Secondly, there is a strong correlation between SO_2 height and SO_2 amount regardless of instrument as column amount algorithms typically require an assumed profile in the first place. It may be possible to add O_3 columns from an assimilated model dataset, however SO_2 columns from the models strongly depend on the input of emissions from volcanic eruptions, which in turn are often largely constrained by satellite observations.

6) **Line 229-231: If you scaled the parameters to be within the max/min range, plus the tanh activation function you used, the gradients of DNNs with respect to trainable parameters in DNNs would not be sensitive to predictors closer to the max/min values. The degraded performance of your DNNs shown in Figure 3b could be a result of this. Maybe try batch normalization?**

- Figure 3b shows degradation at high SZAs. Aside from NN parameters, this is also explained by worsening performance of the radiative transfer calculations at high SZAs (strongly reduced signals due to light absorption leading to much lower signal-to-noise ratios).
- Thanks for pointing out the possible effect of the max/min scaling. We agree this can pose a problem, although in our case satellite observations tend to have a much smaller range in the parameters for a given area. Therefore this is not expected to be a big problem when applying to OMI for the volcanic cases in the paper. We still hope to fix this issue while optimizing the algorithm in the future.

7) **Regarding the stability of DNNs, you could also consider to add skip connections. This is not technically hard. It would smooth the surface of the loss function and reduce the number of local minima. (arXiv:1712.09913)**

Thank you for the suggestion. As previously mentioned, we chose to use L2 regularization as the main method of improving stability and avoiding overfitting. This seemed to work adequately well. The "Dropout" technique was considered, however, it performed worse than L2 regularization for our problem. We will consider the skip connections in future work when we may attempt to optimize the algorithm but it does not seem like a simple implementation in the Python Keras module that we used for this particular study.

8) **It is difficult to evaluate the performance of your DNNs by comparisons shown in Figures 4, 5, 7 and 9. A heat map could be very helpful here, between the predicted volcanic SO2 layer heights and those retrieved from other satellites.**

We appreciate the suggestion. Comparing the satellite retrievals spatially present an issue since the overpass of OMI is not the same as the other instruments. Therefore the movement of the plume in between measurements needs to be taken into account. To get a general idea of the agreement we included the PDF plots as the comparison. It is also worth noting that the techniques for retrieving have some differences which can influence distribution of values within the plumes and that the information content and sensitivity of IASI IR retrievals differ from UV based retrievals, contain different physical parameters.

9) **Table 2 and 3 show statistics intercompared within the synthetic data set. I think it is meaningful to also have statistics compared with other independent satellite retrievals.**

For Table 2 and 3 the main goal was to show mainly the sensitivities within the NN training based on noise and restricting certain parameters. Comparing statistics between other retrievals in the application stage is also possible but with different metrics such as mean, median, quartiles etc. This is somewhat illustrated by the PDFs. We have implemented the suggestion to include another table with quantitative comparisons of the retrievals in the manuscript.

10) **If you do have access to other retrievals of volcanic SO2 layer heights, then I would suggest a multi-stage training. In stage 1, synthetic data sets are used. In state 2, you can keep training the model from stage 1, using a subset of real retrievals as the training set and the other as the testing set. If only synthetic data sets are used and the forward model generating these sets is not perfect, then the trained model must have a degraded performance compared to the forward model (due to errors from deep learning model itself). Moreover, during the forward model calculation, if the column data (O3 and SO2) are sampled within some range, then your DNNs would have a difficult time in predicting outliers and extrapolation should be very careful. Speed would be the only advantage then. The data-driven nature of DNNs should be taken advantage of.**

- Similar to using column amounts from other satellite data sets, introducing other height retrieval also would add error due to differences in overpass time.
- Another main challenge is the lack of fast and reliable retrieval of SO_2 layer height. Furthermore IASI (infrared) retrievals should not ideally be used together with OMI outside of a result comparison since some physical parameters are different between the two.
- It is true that there would be issues with extrapolating far outside the range of O_3 and SO_2 column. However, in volcanic eruptions the SO_2 column rarely exceeds 1000 DU and ozone column is also within the range used for forward RT calculations. If necessary, additional spectra can be calculated with an extended range of parameters and included in training.

11) **Bayesian neural networks are ideal for such prediction problem, when there are uncertainties associated with the outputs. This could be future work due to the technical complexity. But you could still alter the random seed to generate an ensemble of DNNs, such that you can provide a rough estimate of uncertainties on the predictions.**

- This is a very interesting idea. Ensembles of DNNs may improve the performance (accuracy) of single networks for solving remote sensing problems (Loyola, 2006), usage of ensembles for the SO_2 layer height retrieval will be investigated in the future.

- Thank you for this suggestion on altering random seed. We have attempted to change random seeds in the NN for initializing the weights and biases, but this did not impact SO_2 height results significantly with our current NN setup. We agree it may be useful to use the random seed variation to produce an estimate of random error within the NN. This idea has been implemented and tested on one of the OMI orbits to obtain an uncertainty measure of changing random seeds within the machine learning. This will be included as a figure in the revised paper and will be discussed further in the main text (in Section 4 discussion).

Minor comments
Line 61: replace VCD by vertical column density (VCD)

Changed in manuscript.

Fig 4: can you change the colorbar of (c) to the same scale as (a) and (b)?

We have replotted the GOME-2 figure (4c) to match the same color scales.