Response to Reviewer 2

Thank you for the review, we respond to your comments below. We have kept your original text in black, our responses are in blue and specific changes are in underlined blue.

Predictions of satellite retrieval failures of air quality using machine learning

Malina et al.

Summary

This paper investigates the usefulness of machine learning to streamline data processing for incoming satellite retrievals. They highlight the need for improvements in the computational time to go from level-1 to level-2 data with the ever increasing amount of data being created on a daily basis. The principle they explore in this paper is to use machine learning to remove retrieval failures before the processing stage, therefore reducing the amount of data needing to undergo time-consuming processing. They found the an extremely randomized tree model was the best fit for the task and trained the model on CrIS and AIRS-OMI data for ozone, CO and temperature profile.

Their model performs reasonably well with a few caveats and at a high speed, showing how this could be applied to future EO missions and data processing.

Major Comments

This is a well written paper and thorough study that fills an obvious niche. I have a few comments below.

Thank you for these general positive comments.

There is a lot of technical detail in the paper which makes its a long read. Most of it is needed but some sections (e.g. 2.3.1) could be shortened as they're not as relevant to the study.

Thank you for this comment, which is echoed by other reviewers. We have reduced the amount of text in section 2.3.1 and removed Appendix A, as these can be considered superfluous technical detail.

There is a little discussion about the cost/benefit at the end of the paper but there isn't much information/calculations on the actual benefit in terms of computing speeds. It would be good to expand on this point as that is the primary motivation for the paper.

We have cut out major discussion on benefit in computing speed, this is because the computing setup available to the TROPESS project will not be the same as that available to other teams/projects. We have added a new section 2.3.2 that describes in more detail how long retrievals take in the TROPESS/MUSES setup. Based on this information, we indicate through the paper how many retrievals are removed from the pipeline. From this information it is possible to identify how much time could be saved in a serial retrieval, which can then be scaled up for parallel retrieval setups.

I would like to see an expansion in the discussion about what the next steps would be to improve the model and what might be considered a good enough model to be implemented.

We have added the following/modified the following text into the discussion:

"There is a significant cost/benefit aspect to the ML model at this time, where significant processing speed-ups can be achieved, but potentially valuable information may be lost. At this time the ML models are sufficiently developed in order to be deployed in an operational sense, especially with a low threshold value which incurs minimal risk of the loss of valuable retrievals. However, there are clearly more improvement that could be made, for example, the cost/benefit might be improved with more and sophisticated training of the ML model, potentially to the point where there is very little cost in applying the ML model, which is a topic for further work and exploitation. For example, training could be undertaken per region, rather than globally, which may yield improved results. Further, more work can be performed on the QA values that the ML models are trained on. These are currently applied globally, but there could be some value in deriving QA values for distinct regions, and training the ML model on these regions.

As an alternative to regional models, the training data could be carefully constructed to ensure a similar frequency of retrieval failures geographically. Variations across time (night and day, different seasons, cloud coverage etc.) could be balanced in a similar fashion. In terms of ML, the classification performance may be improved by considering more classification methods and particularly more elaborate methods of dimensionality reduction that might be more suitable for spectral data."

Minor Comments

Line 28: is this not the case for all of the TROPOMI species, not just ozone?

Yes, correct, we have modified this sentence as follows:

"with many of the TROPOMI retrieval algorithms consisting…"

Line 48-50: These sentences don't scan well

We have changed this paragraph to read as follows.

"In this study, we investigate Machine Learning (ML) methods for predicting failed trace gas retrievals using measured satellite spectra prior to full retrieval. This research builds on previous studies, which focus on pre-selecting or filtering of trace gas retrievals (Mandrake et al., 2013; Mendonca et al., 2021), and offers a global solution."

Line 61: This doesn't scan well, should the 'allowing for multiple different products' be in brackets?

We have changed this paragraph to read as follows:

"The CrIS instrument was chosen for this analysis due to the high data volume and wide spectral range (allowing for multiple different products). CrIS products are currently a key component of TROPESS, where, for example, CrIS ozone retrievals have been used with reanalysis models to understand tropospheric ozone during COVID-19 lockdowns (Miyazaki et al., 2021)."

Line 63: I think these two sentences should be joined.

To us this would be a bit of a mouthful of a sentence. But we have made these two sentences flow together a bit better, given the changes we made above.

Line 124 (and elsewhere): There are some inconsistencies between 'L1b' and 'L1B' throughout the text.

Thank you, we have changed all instances to L1B.

Line 230: Should this be "number of **true** positives"?

Yes, thank you, changed.

Figure 10: The figures on the left appear to be repeated instead of for each species.

Thank you, this point was made by reviewer 1 as well and we repeat the answer here. These figures had so much data that all of the points were overlayed and thus showing more or less the same thing. We have now split out the pass and failures into two separate columns to give a better idea of the distribution of pass and failure flags. Both Figures 10 and 11 have been updated to reflect this.

Line 372: Wording doesn't make sense. Should this be a comma instead of a full stop?

Yes agreed, we have changed to a comma.

Section 5: There appears to be no reference to figure 10 although I think the text is actually meant to be referring to figure 10 but states figure 11. In which case there would be no reference to figure 11.

Good catch, we have changed this text to now reference both figures 10 and 11.