Response to Reviewer 1

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.

In their manuscript "Predictions of satellite retrieval failures of air quality using machine learning", the authors report on a study aiming at reducing the computational load of satellite data retrievals by identifying measurements for which the retrieval has a high probability of failing before even starting the retrieval. Their proposed method is based on a machine learning approach, trained on a set of satellite spectra and the corresponding error flags from the retrieval. The technique is demonstrated on MUSES retrievals of CO, temperature profiles and ozone using CrIS and AIRS/OMI radiances. The results show that a large fraction of unsuccessful retrievals can be avoided by applying the ML filtering, and that only a moderate number of successful retrievals is skipped.

The topic of this study is relevant as slow optimal estimation type algorithms can often not be applied to all measurements of modern satellite instruments due to a lack of computational resources. Reducing the number of unsuccessful retrievals, therefore helps in providing a larger number of results at the same cost. The proposed method is convincing, and the manuscript is clearly written and structured. However, I have several concerns and suggestions which should be considered before the manuscript can be accepted for publication.

Major comments

1) When using machine learning approaches in science and data retrieval, there is always the same concerns:

- Are all relevant situations covered appropriately in the training data set?

While we have tried to cover as much as possible the range of conditions the satellites will encounter, by covering as much as possible seasonal and spatial ranges. The reality is there are likely to be situations that won't have been covered by the training set. The results from this study show already good performance, which could be improved with constant updated training of the ML model to take into account situations not originally covered.

- Has the method generalised the information sufficiently to be applied to another data set?

That depends on the new data set. Section 5 presents an evaluation of the trained models on a completely new, so-far unused dataset from CrIS. The results of this evaluation show no significant drop in prediction performance on a new data set. This indicates that the model generalizes well to unseen data.

On the other hand, we have not tested the generalization on new data sets generated using different retrieval algorithms or featuring significantly different spatial or temporal distributions. It is reasonable to expect a drop in prediction performance when the test set has properties significantly different from the training set. This is alleviated by the simplicity and quick training time of the applied method, which makes it easy to adapt and train on such data sets.

- Did the algorithm learn the intended connections, or has it generalised correlations which exist by chance or are not cause-and-effect type links?

This is a challenging question. While some causes for poor quality retrievals are understandable, the application of machine learning allows models to exploit complex patters in the captured spectra that are connected to retrieval failure. Such complex patterns can often be difficult to interpret and analyse. Furthermore, the type of machine learning and model analysis we used in this work gives information only on correlations and connections. Analysing cause and effect is a much more difficult task that typically requires controlled experiments and more sophisticated approaches.

Some insights are given by our analysis of the feature importance of different wavelengths. However, more research is required for actionable insights on the failures of retrieval algorithms.

Nevertheless, we are confident that the models' predictions are not primarily based on spurious correlations. We back this claim with our evaluation procedure. We evaluated the performance in two stages – first, using cross validation (a standard and rigorous evaluation procedure in machine learning) and later using a completely new, so far unused data set. The relatively high predictive of our models indicates that they are capturing meaningful information.

- Is a bias of some form introduced in the results when applying the method?

Yes, some bias is inevitable introduced through the construction of the data set. This is true for all models trained using machine learning. To alleviate this, the trained models should not be applied in conditions significantly different from the training set.

Our analysis in Section 5 also reveals that the frequency of retrieval failure varies significantly geographically. This imbalance unfortunately leads to a bias in the trained models, which may produce more false positives in those regions. To alleviate this in future work, we propose the development of regional, specialized models. Alternatively, stronger global models could be obtained by constructing a geographically balanced training set, ensuring that the frequency of retrieval failure does not vary significantly in different geographical regions.

Some of these questions are discussed throughout the manuscript, for example, in the context of the erroneous flagging of high CO values. However, the manuscript would benefit from a specific section discussing the potential problems of the method and what the authors found in their tests.

We have included some of the text from the previous answers into the discussion to address the reviewers' request. As follows:

"Finally, as with all ML approaches there are challenges that could cause some problems with the results. For example, are the training datasets representative, or are biases introduced during training, or many other common issues not directly identified here. It is likely some issues are present in the current for of the ML model presented in this paper (for example biases). However, in order to increase confidence in the results, we evaluated the performance of our ML model in two stages – first, using cross validation (a standard and rigorous evaluation procedure in machine learning) and later using a completely new, so far unused data set. The relatively high predictive ability of our models indicates that they are capturing meaningful information and are effective. Therefore, although the performance of the ML model most likely can be improved, we are confident that they are effective."

2) As far as reported in the manuscript, the method was tested on a very limited data set. Surely, more MUSES retrievals are available to test fast ML filtering. Can a more robust test be performed using data from different seasons and different years?

This is an important point, our judgement with this paper was that it is already quite long through the description of the method involved, and further extending the paper through additional comparisons would make the paper exceptionally long. The current level of comparisons, which shows comparisons for two different satellites add weight to the validity of the methods. We therefore propose a follow on paper which future refines the ML tool and provides a much wider comparison would be more appropriate, than extending the current comparisons.

3) The discussion of the results for the individual flags is interesting but confusing to me. I do not understand why a new metric (the Cramer's V metric) is introduced instead of simply using the number of successful predictions and the number of false positives as quality criteria as in other parts of the manuscript. Maybe I just did not understand what the authors tried to achieve, but I do not see the benefit of this discussion.

The Cramer's V statistic allows us to understand the overall strength of association between the successful predictions under different thresholds and the various quality flags at each measurement location. We chose this statistic since it allows us to evaluate association between two categorical or nominal values, and it takes into account each measurement location having a prediction and various quality flags, then calculates the strength of association over all these locations.

4) In several places, the authors try to use the results of the ML filtering to identify spectral regions linked to certain error flags. This makes sense for cloud-related flags, as different parts of the spectrum contain different kinds of cloud information, and the ML algorithm may identify them. However, the formulations used in the text are sometimes unfortunate; for example, "Further, the SW CrIS band ... seems to have significant importance across most of the failure flags" suggests that a certain spectral region outside the fitting window is the source of a given retrieval problem, while in reality, one condition (such as broken clouds) can lead to effects in different regions. The ML filter does not necessarily hint at cause-and-effect relations but at correlations.

We definitely agree that cause-and-effect relations cannot be directly inferred by analysing the model. In the reviewer's example, it is indeed surprising that a spectral region outside the fitting window should be associated with retrieval failure. However, it is important to note that our analysis does not claim that region is the source of retrieval failure. Instead, it indicates that region is a useful feature for predicting the success of retrieval. Although the retrieval algorithm does not use that region as input, the measurement conditions that affect the retrieval can be visible in the region and used by the model for prediction, as the reviewer correctly surmised.

In other words, our analysis shows spectral regions that are important for model predictions. The best way to interpret this is to consider spectral regions of high importance as promising candidates for further investigations into retrieval failures that could potentially lead to identification of causes.

5) The OMI-AIRS ozone retrieval appears to be a very good example of the large benefits of ML-based data prefiltering. However, simple filtering using the OMI cloud product would be nearly as efficient in a real-world application without any additional machine learning effort. In general, filtering for known problematic or not interesting conditions could probably be a more transparent alternative to the ML filtering approach proposed here.

This is a highly valid comment, for which there is no clear answer at this point. It may turn out that the ML cloud based filter is more effective than using the cloud product, or vice versa, and we think this is an excellent topic for further investigation where a study directly compares the impacts of these tools. We include an additional element in the discussion to cover this point.

Detailed comments

- L7: duplication of "measurements"

Corrected, thank you. Changed to “multiple satellites”

- L12: applied to many EO satellites – applied to data from many EO satellites

Changed, thank you.

- Introduction: I do not see why the data rate of GEO instruments should be higher than that of LEO instruments. In practice, this might be the case, but this is more linked to GEO instruments being rather recent additions with better detectors.

This is true, we have removed text that have implied GEO instruments have a higher data rate than LEO instruments.

- Introduction: I think the main message of the authors is that more data is coming from the new generation of satellite instruments than can be analysed in NRT. For this simple statement, many references are used, which does not make sense to me. I suggest reducing them and focusing on those relevant to this study.

We have reduced the number of references as suggested, but we feel it’s important to keep a significant number here, as this highlights the wide range of methods and techniques currently under investigation to help with the ongoing problems of speeding up retrievals.

- Introduction: I also think that it should be mentioned that the problem addressed is mainly limited to Optimal Estimation type retrievals, while many other algorithms are fast enough to process the full volume of satellite data routinely.

We have highlighted that this paper largely targets optimal estimation retrieval algorithms.

- L67: “retrievals absorb” => “retrievals use absorption”

Corrected.

- Tables 1 & 2: I do not see the need for these tables

Thank you for this comment, we respectfully disagree as we think these tables provide relevant information about the instruments used in this paper and provide valuable context.

- Section 2.3.1: I think this can be shortened as it is not relevant to the manuscript

A similar comment was made by one of the other reviewers, and accordingly we have made this section shorter.

- Section 3.1.2: I'm not an ML expert, but I think it would be good to add a bit more information on the method of the "Extremely randomised trees" used here – are there no hyperparameters and other settings specific to the model you applied?

We added the following text to expand on the model details:

“In our experiments, we used the Scikit-learn implementation of extremely randomised trees with 100 trees in the ensemble, no depth limitation, Gini impurity as the measure of split quality, requiring at least 2 samples to split a node and at least 1 sample in leaf nodes. The rest of the hyperparameters were left at their default values.”

- L247: “only training is performed on” => “training is performed only on”

Corrected, thank you.

- L279: "These results suggest that non-fitted elements in the retrieval process have a significant impact on the overall quality of retrievals" – I'm not sure what the authors are trying to say and how this can be deduced from the fact that the ML algorithm is using information from outside of the fitting window to predict failure of the retrieval better. To me, this feels like a confusion of correlation and cause-and-effect relationship.

We have softened this statement to read as follows:

“These results suggest further investigation into non-fitted elements in the retrieval process, as these may be having an impact on the overall quality of retrievals, and potentially hint at some of the underlying reasons behind retrieval failure.”

- L308: As mentioned above, the ozone failure flag is special as it is linked to cloud cover in a simple and easy-to-predict way.

We are not sure how to respond to this comment.

- Section 4.3: I was surprised that the authors did not evaluate whether combining the prediction of individual flags would be better than training for the overall success flag.

We agree with the reviewer here, but as with the point above, this paper we consider more as a pilot study, where we have tested the viability of this method. We hope future studies will consider this point.

- Section 5: References to Fig. 11 should probably be to Fig. 10. Figure 11 is not discussed at all, as far as I can see.

This was caught by one of the other reviewers as well, references to both Figs 10 and 11 are now clear in section 5.

- Figure 10: Left column repeats the same figure three times, which I guess is a mistake.

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.

- Figure 11: Something is not quite right here – the right figures' colour scale does not seem correct.

As with the point above, apparent error was due to overlapping data points, splitting out the pass and fail flags into the new left and middle columns has helped highlight the correct distribution of flags, and related predicted failures. The updated Figures 10 and 11 answer this point.

- Figure 11 caption: "from a day in 2020" – which day?

Corrected to August 12th.

- L375: "do a good job in predicting the actual failures" – this is not clear from the current set of figures.

We have toned down this statement to "capable of predicting the actual failures"

- L418 and elsewhere: I find the percentage speedups difficult to understand. What is a 100% speedup? At least to me, it would be easier to understand if the reduction in computational time is given.

We understand the point made here, for us we are not sure how beneficial giving exact timings are, as our computational set will differ to those used by other teams, and therefore giving exact figures may not be so useful. What we now provide an additional section of text describing how long our retrievals take, to give a baseline of how we calculate speed ups. This is the new section 2.3.2.

"The TROPESS project has access to computational facilities that includes 100s of individual cores. This processing facility typically allows for the completion of trace gas retrievals in several minutes, with multiple retrievals occurring in parallel. The time for a retrieval depends on the instrument, with AIRS-OMI taking longer than CrIS. Based on the computational facilities available, and the processing times for retrievals, typically a test dataset of around 8000 retrievals takes roughly 2 days to create. Through this paper we refer to how the application of the ML tools allow for a speed-up in processing, this processing benchmark is what we base the speed-up off."

- L439: Again, I'm confused by the speedup given. If 74% of the data is removed, I would either see a speedup by a factor of 4 or a reduction in computational time by 74%.

We decided to remove most references to speed-up in the paper, given that processing set-ups will vary depending on the user, so explicit values are not so useful.

- Figure 15: it is clear from the figure that the filtering is mainly removing cloudy scenes and the right part of the OMI swath

Yes, agreed.

- L459: "This cost/benefit can be improved...". This might be the case, but the authors have not shown any indication of that

We have modified this statement as follows:

"This 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."

- L460: "This work represents the first step in understanding why and how retrievals fail" – I disagree. This is not what this work is about. If you are interested in finding the reasons for failing retrievals, the detailed error information from the OE retrieval will be more helpful.

Our aim with this statement is to identify future work paths, where ML could be used to help refine what causes retrievals to fail. We have modified this statement as follows:

"This work may represent the first step in understanding why and how retrievals fail, for example future work"

- L482: A different name is used here for your ML method than in the description in the text. Please make it consistent.

Corrected to "extremely randomized trees"

- L485: "which can be reduced..." – again, this has not been shown

Changed to "could be reduced"

- L486: "speedup of 66%" – I do not know how you compute these numbers. Only 67% of the original retrievals have to be performed, leading to a reduction of the computational time by 33%. The speedup would be by a factor of 1.47, but as discussed above, I think the computational time is much easier to understand.

As per the related comment above, we have removed direct references to speed-ups, as these specific figures are probably not so useful to other groups. We think references to how many retrievals are removed from the pipeline are more important than specific speed-up figures.

- Appendix A and B: I do not think that this is needed or adds anything to the manuscript

We have removed Appendix A, however we disagree with the removal of Appendix B (now Appendix A), we think it conveys relevant information about the retrieval windows used in this study.