Response to reviewers

Anonymous Referee #1

General comments:
This paper concerns the detection of unphysical outliers in satellite-based observations of chemical composition. While technical in nature, this topic is an important one for users of atmospheric composition data, and is little discussed in the literature. Therefore I find the subject matter appropriate for AMT.

It is clearly difficult to devise a consistent method which can be used for all atmospheric species at all latitudes/heights as the authors have attempted to do here. Also, the success of any method is quite difficult to quantify, as it is difficult to objectively determine which outliers are "unphysical", and the result of defective observations or retrievals, or just "natural" outliers, due to rare dynamical or chemical conditions.

I do however have serious concerns about the methods used by the authors, and the justifications for these methods. I doubt that the methods introduced here would be used by other satellite data users. I also fear that the authors' conclusion that the median absolute deviation (MAD) should not be used in outlier detection would actually hinder future attempts at outlier detection.

I strongly encourage the authors to revise the outlier detection method used, perhaps using some of the suggestions in the comments below, in order to try and find a simpler method. I also strongly encourage the authors to include more information and discussion about the potential sources of outliers in their data set, both natural and unphysical, in order to frame the difficulties and consequences of outlier detection.

We thank Anonymous Referee #1 for their comments and suggestions, which led to an update of the outlier detection method. All of these issues are addressed below

Specific comments:

I am not convinced of the quality of the outlier detection procedure described in this work. The distribution fitting procedure, step 1, is likely to run into problems when trace gas distributions are non-Gaussian, especially for instance at low concentrations when distributions are often log-normal.

Thank you for pointing that out. The distribution fitting is now done in log-space, and the fits have been greatly improved.

Secondly, as it is applied to the full data set (within latitude bands), it fails to detect outliers for species with strong seasonal variability, when those outliers lie within the variability of the full year.

Correct. That is the reason for implementing step 2.

Lastly, performing distribution fits for each species at each altitude seems a lot of computational expense for what actually amounts to little more than defining an "edge" of the distribution, especially when this edge is then arbitrary shifted by some amount based on the SD of the data.

The "shift" based on the SD is no longer used, and the computational expense of analyzing at each altitude level is not a concern (it takes under an hour to analyze the entire ACE-FTS dataset, including subsidiary isotopologues). Also, we chose to do the analysis at each altitude level in order to avoid the need to define altitude regions for each species.

It is also not clear if the first step provides any added value compared to the second step, i.e., does it detect any outliers that wouldn't be detected by the second step?

Because we need our detection method to be sensitive to the presence of physically realistic "outliers" (as discussed below), this first step is crucial to removing the extreme physically unrealistic outliers.

The second step, based on the standard deviation (SD) of data within a 15 day window, is problematic in its use of the SD rather than the MAD. Outlier detection based on the MAD is generally regarded as superior to methods based on the SD (Leys et al., 2013). The authors claim that using the SD and MAD leads to different thresholds, and that "using the standard deviation when there are extreme outliers can allow for the acceptance of data that should clearly be rejected" and that "using the MAD on asymmetrically distributed data can lead to the rejection of 'good' data". However, these conclusions are based on quite arbitrary thresholds, which are clearly different for the SD and MAD-based techniques since the distribution of the data they are being applied to is clearly non-Gaussian. If the distribution was different, different conclusions might be found. The authors go on to state that "Since the distributions of ACE-FTS data across regions (altitude/latitude/season/local time) tend to be asymmetric, the screening process does not make use of the median absolute deviation." Asymmetry of the distribution is clearly an obstacle, but not an unsurmountable one. For example, asymmetry can be dealt with by considering separately the median deviation on either side of the median (Rosenmai, 2013), using a so-called "double" or "two-sided" MAD.

We should have also stated that the distributions are also often multi-modal in addition to being asymmetrically distributed. When using the MAD on a multi-modal distribution, and one of the modes is sampled less frequently than the other (which is the case with our data), the MAD does a poor job of describing the variance of the entire data set. In this case, the MAD only describes the variance of the more prevalent mode and can screen out the less frequently observed mode as outliers. This is a common problem at high latitudes where observations can be vastly different depending on whether you are observing inside or outside of the polar vortex (and this problem cannot be solved with better binning parameters). This is often still a problem if the screening criteria are quite lax, even if you were to define outliers as those more than 10xMAD away from the median. Therefore, we require a screening method that is at least somewhat sensitive to the presence of outliers. This is now discussed in the text.

For step 2, instead of the mean and SD (which are highly sensitive to outliers), we are now using the median and the variance measure mean(|x_i-median(x)|). We are calling this measure the MeAD, and it is less sensitive to outliers than the SD, but more so than the MAD. The process no longer needs to be iterated, and any data that is 10xMeAD away from the median is considered an outlier.

Furthermore, the authors later introduce a method based on the SD of the monthly data, which should have the very same problem if the data distribution is asymmetric, in addition to being highly sensitive to the outliers (necessitating the use of an iterative method). I have the feeling that the use of such a "double MAD" could quite improve the outlier detection method used in this paper, and remove the need for the iterative SD-based method.

Again, the issue is more due to distributions being multi-modal, which should have been more explicit in the original text. The changes to step 2 have reduced the sensitivity to outliers (although is still more sensitive than the MAD, as required), and have eliminated the need for an iterative method.

The current title is very general, although the work involves a very specific instance, involving a single data set (ACE-FTS) and one (multistep) method. The title should be made more specific.

The title now refers to ACE-FTS data instead of satellite data.

The introduction requires an explanation of what "anomalous data" is. This is a subtle issue which requires some clear description – in fact, the use of the word "outliers" needs to be introduced, since "outliers" refers generally to points which lay far from the center of the distribution, but such "outliers" need not necessarily be "wrong", or faulty, just rare. As this paper seeks to remove outliers resulting from errors in the observations or retrievals, and not (I assume) outliers due to anomalous chemical or dynamical conditions in the atmosphere, some careful consideration and language is needed to differentiate natural vs unnatural outliers.

Correct. The manuscript now defines natural vs. unnatural outliers and is careful to distinguish between the two throughout.

The introduction is sparse in terms of referencing prior work concerning the use of probability distribution functions in regards to atmospheric measurements. I would suggest at least including citations to (Sparling, 2000), and (Lary and Lait, 2006).

Thank you for the suggestions; new references have been added.

Line 12: has been

Corrected.

Line 34: The technique doesn't assume that the distribution is symmetric, but application of the MAD as an outlier detector may run into problems if the distribution is asymmetric.

Correct. This has been rephrased.

Line 40: not clear why sampling bias is important. Sampling bias is an error in a mean due to insufficient sampling. The next sentence refers to the distribution may not be normal, but this is not clearly connected to sampling biases. More explanation is needed.

We think that this was most likely an editing error. We no longer refer to sampling bias.

Line 41: There is no reason the atmosphere need be normally distributed, and it's not clear why one would need to identify regions where it is.

We were not intending to state that the atmosphere should be normally distributed, it's just that the use of the MAD (for outliers) has an inherent assumption that the distribution of the data is symmetric. So use of the MAD would require binning the data such that the subsets are near-Gaussian. This has been clarified in the text.

Line 42: Very unclear what is meant here. Obviously, the atmosphere has some degree of predictability. But this has little to nothing to do with normality of distributions of atmospheric quantities.

Rare events can be observed that greatly change the state of the atmosphere in a particular region. Because of these, distributions can be either skewed or multi-modal. This has been re-worded in the text (now lines 47-51).

Line 44: The use of "normal" is confusing, as it's unclear whether it should mean "typical" or something connected to the "normal" or Gaussian distribution.

Typical was meant. This has been replaced.

Line 70: this sentence concerns "flagging" vs. the option of removing them from the data set I assume? This is not perfectly clear.

This has been clarified in the text as determining data to be potential unnatural outliers.

Line 71: Again, some clearer description of what "good" and "bad" means is needed.

We now state the good refers to "likely to be physically realistic", and bad refers to "likely to be physically unrealistic."

Line 78: This is repeating the discussion of the introduction, which, again, should not be a major hurdle to outlier detection.

The text was changed to now read, "Global satellite-based measurements of trace gases in the atmosphere typically are not symmetrically distributed and are often multimodal."

Line 80: This is a strong statement... of course if someone is interested in polar processes, they will focus on measurements there, but the fact that different processes are important in different locations does not mean one should partition the data.

We believe that it is logical to partition the data, to some degree, when trying detecting outliers. What is considered an outlier in the tropics may be perfectly typical in the polar regions and vice versa. Therefore it is desirable to partition the data into different regions that are each affected by similar processes. Making as sure as possible your subset of data are all exposed to the same local conditions is how you get a data set that is closest to uni-modal as possible.

Line 81: maybe point out that this is for all latitudes.

The text and caption now make clear that it is all data at that altitude.

Line 84: the use of "assuming a normal distribution" is somewhat misleading... 1.428*MAD will equal the SD in the case where the distribution is Gaussian, but you don't need to assume a Gaussian distribution to make use of 1.428*MAD for some purpose...
I know you are not saying otherwise, but some care is needed with the language in order not to confuse non-experts.

This has been clarified in the text.

Line 86: How does one know this data should be rejected?

The text has been reworded to read, "using the standard deviation when there are extreme outliers can allow for the acceptance of data that are most likely physically unrealistic."

Line 89: Clearly the temporal variability of H_2O (and other species) in the Antarctic vortex leads to an asymmetric VMR distribution. Later you use a 15-day running mean, but in theory one could use a MAD on each month of data rather than the full data set? In this case, the median and "rejection" thresholds shown in Fig 1 would have an annual cycle.

In many cases, using the MAD on one month of data is fine. However, in many cases, especially in the polar regions where observation locations can be either inside or outside the vortex, even one-month distributions can be asymmetric/multi-modal. In many of these cases, in order to keep "natural" outliers from being rejected, we need a way of defining the variance that is sensitive to outliers. This point is now discussed more thoroughly in the text.

Line 110: Defining a threshold in terms of the "expectation distribution" makes much less sense than doing so in terms of the relative pdf, since your threshold then depends on the number of measurements.

Since, in all cases, we have a sufficiently large number of measurements, we've chosen to pick the region where you're expected to have measurements, not a region that contains a high probability of having measurements.

Line 112: The meaning of "1" is hard to understand, and seems totally arbitrary. If the threshold was in terms of the relative pdf, at least you could place the threshold at a 99% confidence interval (or 99.99%, as stated), for example.

This was unclear in the original manuscript and has been improved in the revised text. We are determining the relative pdf from the measurements. Than we multiply it by the number of measurements. So, the integral of the distribution is equal to number of measurements (*N*). Where the integral from –infinity to *x* is equal to 1, you would expect to find 1 measurement in that region. Where the integral from –infinity to *x* is equal to 0.1, you would expect to find 0.1 measurements in that region. We were previously using a cut-off value of 0.0001. With the better fit in log-space, we are now using a limit of 0.025 (and N-0.025 as the upper limit). This means that any data point outside those limits have a 2.5% chance of being a natural inlier (97.5% chance of being an unnatural outlier.

Line 116: these statements must assume a typical value of N, unless I have totally misinterpreted the preceding description.

The improvements described above should address this issue.

Line 121: What is the justification for separating sunrise and sunset measurements?

This was done in order to separate observations into regions where local conditions (i.e. local solar time) are similar.

Line 125: What is the justification for using 3 Gaussians? In the cases shown, the method produces a good fit, but it doesn't need to always be so good. As there is no physical justification for the distributions to be a mix of Gaussians, this method seems arbitrary, and the quality of fits should, in principle, get better as more Gaussians are included. Plus, the quality of the fits will depend on the presence/absence of outliers, unless the fitting procedure is robust.

In this method, we are assuming that there is a maximum of three modes that influence the state of the atmosphere in any given region. This may not be the case, but we have not come across any instances in the ACE-FTS data set where the atmosphere exhibits a quad-modal distribution or more. As we now do the fit in log-space, it assumes that each of these modes are either normally or log-normally distributed, which, again, isn't necessarily the case, but we would be assuming that if we were to use the MAD.

To lessen the impact of fitting to outliers, first we remove any data points whose absolute values are 10,000x away from the median of the absolute values (as mentioned in the text). Then we use an ad hoc "Olympic" method, where we remove the lowest five and greatest five values before fitting the distribution. If they were inliers, then they will still lie on the fitted distribution and not be rejected. This revised method is now discussed in more detail in the text.

Line 130: The pre-screening of the data deserves some more attention, before the description of the outlier detection procedure would be best.

This discussion has been moved to earlier in the text, and further details have been provided—now lines 135-142.

Line 137-138: Which fit are these metrics for? The CH4 fit looks much worse than the NO2 fits.

It is an average of all fits, now explicitly stated. The CH_4 fits have improved with the new methodology.

Line 139: why are extreme events undersampled in ACE-FTS data?

Undersampled was perhaps a poor choice of words. We meant that ACE-FTS are going to observe fewer rare events, as it is a solar occultation instrument and simply makes fewer observations than other limb-viewing satellite instruments. However, this section has been deleted as we no longer are using the ad-hoc "shift" method.

Line 146: this is a disturbingly complicated method. It is not clear how this is different than simply adjusting the threshold value used. I would assume that use of a double MAD would be superior to this method.

If we were to "simply" adjust the threshold value for every species at every altitude/latitude/local time bin (that's adjusting the threshold for over 15,000 bins!), that would be no different than just going in by hand and rejecting the data where we think that threshold should be. This way we treat all the data in all the bins equally and let the statistics do the rejecting. As well, using a two-sided MAD would not address the issue of the data being multimodal.

Line 154: This is not really the fault of the expectation distribution, but actually of the method of applying the method to the full year of data at once. If one lumps all the data together, any method you use will only remove the outliers on the edge of the distribution, and if the species has strong cyclical variability, it will only find outliers at the extremes of the cycle.

We were not trying to imply that it is a fault. It is a hard limiting method to remove extreme outliers. Since our second step, which is applied on smaller timescales, is somewhat sensitive to outliers, it is preferred not to have those extreme outliers present in the analysis. This has been made clearer in the text—now lines 184-186.

Line 160: This iterative technique is exactly what the use of MAD can let one avoid. And the same problem MAD has with an asymmetric distribution is also affecting a SD-based outlier detection procedure, so it is not clear why the MAD is not used here. In fact, I would think that a temporally varying double-sided MAD filter could be all that is needed, making the fitting of Gaussians unnecessary.

Since much of the data is multi-modal, we need a method that is sensitive to natural outliers. This makes using the MAD less than ideal. To demonstrate this issue, a figure of what using 10xMAD produces for Antarctic H_2O data has been added (now Figure 7).

Line 165: Extreme outliers sure, but even moderate outliers can affect the standard deviation.

We are now using the MeAD, which is less sensitive to outliers than the SD (as described above).

Line 178: “Robust” has a specific meaning in statistics, meaning insensitive to outliers. Thus, the use of “robust” in this sentence is potentially confusing.

Robust is no longer used.

Line 179: “keep the data as inliers”: the outliers are data also. Some more careful wording required.

This has been changed to improve clarity.

Line 187: In addition to the identified instances, it appears as if the method is identifying a number of outliers for high altitude NO and CO at low concentrations which appear to be within the typical ranges of VMRs. The distribution of these species is almost certainly poorly approximated as Gaussian, perhaps log-normal would be a much better fit to the data distribution. It could be that the distribution fitting procedure is running into problems here.

This was simply an issue of scale. No likely realistic data was being rejected in these regions. However, with the updates to the methodology, this issue is reduced for CO, and we have therefore updated this section.

Line 202: Again, “outliers” should be further qualified – “unphysical outliers” or such.

This has been changed throughout the text.

Line 223: More description of “screening the data based on the percent error” is needed, as this has little or nothing to do with the material presented in the manuscript.

This is what was recommended in previously released versions to provide some basic data filtering guidance. For legacy reasons, this criteria was given a flag value. This is now mentioned explicitly in the text.

Anonymous Referee #2

This technical note describes the approach taken by the science team for the Atmospheric Chemistry Experiment Fourier Transform Spectrometer (ACE-FTS) instrument to identify "outliers" in their Level 2 dataset. The ultimate goal being to provide data quality flags and information to eventual users of the ACE-FTS data. If this is the approach that the ACE-FTS team have chosen to adopt (as implied by this paper), then I fully recognize that it should be documented, and the process by which it was arrived at should be described to the user community. A technical note in AMT is a very suitable way in which to do that.

However, I have serious concerns about the efficacy and suitability of the approach described. In addition, in my opinion, the paper lacks important and valuable detail on the background to the method. The second of these issues can and should be remedied by the authors before publication. The first is a larger issue that may motivate more work, that is perhaps left to a follow up analysis and paper.

We thank Anonymous Referee #2 for their comments and suggestions on our paper. The approach has been updated, as described in the response to the first reviewer. In addition, some more background has been added.

Fundamentally, this paper does not make it clear what phenomena give rise to these anomalous "outliers". Are they due to anomalous behavior within the ACE-FTS instrument? Are they due to poorly converged retrievals? Are they due to poorly modeled signal contamination from clouds or aerosols? One presumes that all three can contribute, individually or together, on a case by case basis to give such outliers. A discussion of these contributors should be included in the paper.

A brief discussion on the potential causes of outliers is now given in the text.

Furthermore, for each of these contributors, surely the outliers that arise are better diagnosed in a phenomena-specific manner, rather than by resorting to some rather ad-hoc examination of Level 2 data. Instrument anomalies would, presumably, be detectable in raw telemetry, interferrograms or radiance spectra.

A non-trivial number of occultations are excluded from the analysis due to known instrumental/processing errors. These issues (e.g. unrealistic N2O concentrations due to a convergence failure for occultations with low water levels, ice buildup on the detectors during early mission occultations, bad spectra used in the calibration, level 0-1 processing errors, etc.) were known before the analysis was performed, and all are listed on the ACE-FTS website. These occultations are flagged as 7s—known processing/instrument errors. This is now emphasized in the text.

Poorly converged retrievals could be identified through some radiance chi-square metric that the retrievals presumably (hopefully!) generate for each occultation scan (and/or even each spectrum). Cases of aerosol/cloud contamination would be identifiable as impacts to the "background" on observed spectra. Only when all avenues to screening for each phenomena in its "native landscape" are exhausted would I consider a screening based purely on Level 2 data with no reference to underlying anomaly mechanisms. Perhaps the team have already exhausted such approaches. If so, the manuscript makes no reference to that effort. If not, I would urge that the authors return to their dataset and explore a more "mechanism-aware" approach to the screening. Such approaches can really only be put together by a team with thorough end-to-end knowledge of the measurement system, and it is vital that this be done sooner rather than later, given the age of the ACE-FTS instrument, and the limited funding that can be expected for it in future years. If the team is genuinely confined to approaches that rely purely on searching a dataset for outliers, without any reference to characterizable underlying mechanisms, then the approach described is probably the optimal one for the input dataset.

Both of these (chi-square and background impacts) are potentially possible for pre-screening the level 2 data; however, implementing these is not something that is currently feasible. Hopefully, in the future these methods can be implemented (we are investigating the possibility of making chi-square values available for the next version release), but in the meantime, this approach gives our data users quality flags to help them appropriately screen the data for our current and historical data sets.

However, all users need to be fully aware of the classes of potentially genuine atmospheric science phenomena that may be flagged as spurious by this approach. The authors give some examples of these. However, I can think of others, not least the unprecedented enhancements seen in lower stratospheric species in response to the "Black Saturday" fires in Australia in February 2009 (e.g., Siddaway 2011, Pumphrey 2011) and the episodic injections of moist air into the summer time northern mid-latitude lower stratosphere reported by Anderson et al. (Science 2012) and as seen in satellite observations (Schwartz et al., GRL 2013).

Since we analyze both the final inlying data along with the data we reject, we can say with confidence that we are not rejecting genuine atmospheric phenomena (now discussed in the text). Still, the whole approach is based on not flagging rare phenomena as spurious. As such, we do not reject biomass burning events, we do not reject NOx intrusions from the upper atmosphere into the stratosphere, we do not reject H_2O increases in the summer mid-latitude lower stratosphere, we do not reject anomalously low HCl in the Arctic spring lower stratosphere (Manney et al., 2011, Nature). We have included H_2O and HCl examples in the manuscript. Unfortunately, ACE observations were not in the immediate vicinity of Australia anytime near Black Saturday, so we can’t show any extreme changes due to that event.

The approach described in this paper is, mercifully, not likely to reject large scale events out of hand (as happened, it seems, during the 1980s era of TOMS data processing). However, that episode has left many in the satellite atmospheric remote sensing community with understandable trepidation about all data screening methods of this type. To summarize my "big picture" questions, the manuscript should be modified to describe the origins of these outliers, and why other approaches to performing the screening in a more "mechanism aware" manner are not viable (or at least are a "next step"). If they are viable, then my choice would have been to start with them rather than the approach documented here. However, having some screening is preferable to having none, so I do not necessarily want to hold up publication of this paper.

As noted above, there is now a brief discussion on the origin of outliers, and we discuss the types of issues that affect occultations that have been pre-screened and removed from the analysis.

The manuscript is generally very well written, and the figure are clear (if a little small, but perhaps that's an AMTD vs. AMT issue).

More minor comments.

The title implies a far broader subject matter than is covered in the text. If the goal is to provide needed direction to the ACE-FTS data users, then it is critical that ACE-FTS be included in the title. Also, the "detecting outliers" title is unclear: are we supposed to think of these as spurious data to be rejected or geophysically real data of scientific interest (e.g., Black Saturday, H2O injection etc.). In point of fact, the abstract and the manuscript are also a bit vague on this in places also. It's only because the abstract talks about "quality flags" that one can draw the conclusion that it's the former. For the title, I would suggest something like: "Detecting [screening for?] spurious outliers in ACE-FTS atmospheric composition observations"

The title has been changed to say ACE-FTS instead of satellite-based.

Abstract Line 14: People might think "fitting error" means radiance fitting error, but that's not the topic of this paper (even though I feel that would be the preferred method). It would be good to clarify that this is a statistical fitting of the Level 2 data to some distribution.

This has been clarified and now reads "retrieval statistical fitting error".

Page 8395

Line 1: "This method is much less sensitive to extreme outliers" is not as clearly explained as it could be. Is that a good thing (it gets a better picture of what is "normal" because of that lack of sensitivity, so can more easily identify fliers), or is it a bad thing (it misses fliers).

It can be either a good thing or a bad thing depending on the data being analyzed. The text now reads, "This method is much less sensitive to extreme outliers, as the presence of outliers typically has an insignificant effect on the median value. It can be used as an efficient tool in detecting outliers for data that is normally distributed. However, using this value as a method of detecting outliers can be ineffective if the data being analyzed are multimodal and/or are asymmetrically distributed about the median."

Line 14: Add some more examples (e.g., those listed above).

Examples (biomass burning and presence of PSCs) have been added:

Line 18: Add "spectral" between "high-resolution", just to be explicit?

This has been added.

Page 8396

Line 1: Perhaps "are" -> "have been" just to be completely clear.

This has been corrected.

Line 11: Is this the grid that the level 2 data are distributed on? If not, will users have to interpolate level 2 data to this new grid to use the flags? Also, if not, why not put the flags on the same grid as the data?

This is the grid of the level 2 interpolated-grid data (referred to as the “1-km” or “scientific” grid). A level 2 version is also released where the actual tangent altitudes are used (not on a constant grid), however this makes it more complicated when binning the data, and the vast majority of data users use the interpolated-grid data.

Line 13: "As well, it was..." -> "It was also..."

This has been changed.

Page 8397

Line 1: "consistent" is unclear in this context. Consistent with what? Between the mean/median approaches? Between the different products?

This now reads, “1.428 is the scale factor for the MAD to equal the σ assuming a normal distribution”

Line 4: I would take issue with "clearly", again for the reasons discussed above (except in the case of strongly negative vmrs, I grant).

“Data that clearly should be rejected” now reads, “data that are most likely physically unrealistic”

Equation 2: This seems strongly at odds with the traditional definition of both E(x) and of pdfs, for which one performs an integral to compute the probability that x is within a given range. I suggest you modify your nomenclature to avoid confusion and this break from tradition. If the references given below establish a new "tradition" then it might makes sense to stick with these unconventional definitions, but you should be more explicit about that.

We gave an incorrect explanation in the original manuscript. The manuscript now refers to the expectation density function (edf).

Page 8398 Line 9: Just to be sure this is 10ˆ-4 on the scale where less than unity gave you pause before, yes?

Correct. The new wording (lines 143-157) should make this clearer in the text.

Page 8399

Line 8: "... for in the fit. For each subset ..." is a bit unclear, is the discussion starting with the "For each subset ..." here the description of the "ad hoc" method introduced above, or is this a new topic?

This section is no longer included, as the “ad hoc” method is no longer necessary.

Table 2

The meaning of the "percent error is not within 0.01-100%" cases was not made clear in the text. Is this the ratio of a Level 2 error bar to the retrieved value, or is it something else. In either case, more detail is needed. Apologies if it’s there and I missed it.

It’s the retrieval statistical fitting error divided by the retrieved value. This is now made clear in the text.

Figure 2

Would be good to label the plots (or at least the x-axes) with O3, H2O.

Titles have been added to the figures.

Figures 3 and 4

These could be merged into one figure (granted the Figure 4 ones get followed up with further analyses, so perhaps it’s OK as is). However, I also don’t see why you don’t include the green fitted histograms on figure 4.

We decided to keep the figures separate and have added the fitted Gaussians to Figure 4.