Overview

We thank the reviewer for his/her comments, which will result in an improvement of the manuscript.

The paper compares the H2O and HDO data retrieved from the ACE-FTS solar occultation instrument and from two different retrieval algorithms applied to the MIPAS limb-emission instrument. This is an update on similar work performed by Hogberg and Lossow in 2019 but using reprocessed ACE-FTS and MIPAS-ESA data. However, it is difficult to know what conclusions can be reached, and how or if these have changed with the new data.

In addition to the new ACE-FTS and MIPAS-ESA versions, the MIPAS-IMK data are a new version and employ a new processor approach, too. We say this explicitly in line 96 and specify which version we use in line 118. In the revised paper, we will make clearer which version was used in previous papers, and which version we use here.

This seems a rather 'mechanical' paper - mostly just reproducing earlier results, the only new aspect being the updated datasets. Indeed, the sort of paper one can imagine being generated, in a few years if not already, by some of the more advanced AI machines.

It might have been acceptable if the original authors wanted to update their paper with new results, in which case I would expect a narrative focusing on the algorithm changes, and the expected and observed impacts on the intercomparisons with respect to their earlier results rather than, as here, analysing the results in absolute terms as if they were being presented for the first time. But if it's a paper with new lead authors it also needs some significant new insight or analysis. I also have some doubts about the methodology.

It is critical to intercompare and validate each new satellite dataset as it is produced. As we have new data versions and methodologies, the data users need to understand changes and updates to these H_2O and HDO data products. This validation work should not only fall to the data producers but should be taken up by other members of the community.

In case of MIPAS-IMK (versus ACE-FTS), we refer several times to previous results (e.g., line 315/316; line 335-337; line 373-377; 436-438; 458-461; 492/493; 501-503; 513-515). We will make clearer in the revised version what changes in the retrieval set-up were made and will discuss the expected versus observed changes between the data versions.

Furthermore, we would like to point out that it was by intention that we followed the methodology of Högberg et al. so closely. By using the same methods, we ensure that the new results of this paper are directly comparable to the results presented in Högberg.

General Comments and Suggestions

1) Colocations

The comparison has been performed on the two versions of MIPAS data as if these were independent satellites. A more satisfactory approach would have been, first, to apply the respective filters to these MIPAS datasets and take just the common profiles, and then compare this with ACE-FTS. This would not only ensure no time/space mismatch between the two MIPAS datasets but also ensure exactly the same time/space mismatch between ACE-FTS and either of the MIPAS datasets.

On this topic, Fig 1 looks very odd. It seems most unlikely that in a 3 day period the best ACE-IMK coincidences are in a different latitude to the best ACE-ESA coincidences - I would expect them to be mostly the same locations. The averaging of all MIPAS profiles within the coincidence criteria also seems undesirable. The averaging will reduce the random noise in the MIPAS profiles so the contribution to the overall SD is no longer straightforward. Better to take just the closest profile. Also, noting the difference in time/space would allow subsequent analysis as to whether the chosen margins are adequate or, more ambitiously, allow the colocation error to be quantified eg when switching to grid boxes or zonal averages.

This was addressed in the response to reviewer G. Toon as follows. "In the revised version of the manuscript, when multiple MIPAS profiles are spatially coincident with an ACE-FTS profile, the MIPAS profile closest in time is selected. In addition, there must be both MIPAS IMK and MIPAS ESA processed data available for this coincident profile."

2) Algorithm descriptions

In 2.1.1/2.1.2 the descriptions of the two retrievals seem to be taken directly from the source papers using their own terminology (possibly via the SPARC papers), with little effort to standardise the information let alone provide some interpretation in terms of possible impact on the comparisons.

For example, 'non-linear least-squares global-fitting technique with Tikhonov regularisation' (for IMK) and 'least-squares global fitting, using the Gauss-Newton approach modified with the Levenberg Marquardt method' plus 'a posteriori regularisation' (for ESA). The reader has to work quite hard to understand whether or not these are essentially the same thing and hence whether significant differences may arise from these aspects of the algorithm.

Even on a more basic level:

- Does ESA-MIPAS retrieve H2O and HDO as log(VMR) or VMR (which affects how you should evalulate biases)?
- Does the IMK-MIPAS account for horizontal inhomogeneities in the line-of-sight direction and/or assume LTE?
- Do both use microwindows in the same spectral region?

Similarly with ACE-FTS, orbital altitude and inclination are measured but these are not given for Envisat. Spectral bands are given for ACE-FTS but not for MIPAS.

We will improve the descriptions of the algorithms and retrieval set-ups and make them more consistent to each other. In each case, we will also describe the changes applied since the previous data version.

3) Retrieval diagnostics

There is no reference here to the retrieval characteristics such as expected random noise, systematic errors or averaging kernels, at least typical values - these are not likely to change much except (for MIPAS) towards the poles where the atmospheric temperature is low.

The SD of the bias, for example, could be put in the context of the retrieval random errors, and the bias itself in terms of the overall systematic errors. Meanwhile the averaging kernel describes the ability of the retrieval to follow 'real' atmospheric variations, which has an impact on the correlations as well as the SD of the comparisons. The lack of an HDO time signature in the ESA data might be explained in terms of the averaging kernel.

We will add all available information to the revised version.

4) CH_4 consistency
Given the large discrepancy between ACE and the two MIPAS profiles in the stratosphere a simply self-consistency check would be to see how these compare with their equivalent CH_4 retrievals, on the basis that (H_2O + $2CH_4$) should be conserved.

While this is an interesting idea, it is beyond the scope of this paper and will be considered by the authors for future work.

5) The authors should be aware of the difference in LaTeX between a minus sign ($-$) and a dash (--) indicating a range of numbers. Where both positive and negative values are possible it would also help if '+' were added in front of the positive numbers to further distinguish them from the --. Thus, in L24: $-$8.6--+10.6 Incidentally, assuming this is taken from Table 2, the actual number in the table is -8.7. Further on this particular point, negative and positive biases don't have any particular meaning in this sentence, so I would just have said 'biases of up to 10%'.

Thank you very much for this suggestion. The changes that the referee proposes will be considered. By one hand and according to the AMT style, we will use the word "to" for indicating a range and en dashes (–) for numerical purposes. By the other hand, the biases will be referred in total percentage, when appropriate, along the text in the revised version of the manuscript.

6) There are numerous references to 'global' averages whereas I would suggest 'dataset' averages, or something similar, unless you really claim that your intercomparison dataset represents some sort of uniform sampling of the globe.

The term "global" will be clarified in the revised manuscript.

7) It is unclear whether or not the MIPAS-IMK product has been updated (these authors refer to 'V5H' and 'V5R' whereas Hogberg (2019) referred to V20 as being the newer version.

See our comment above. As a result of the analyses by Hoegberg et al., 2019 and Lossow et al., 2020, a new retrieval approach for HDO has been developed for MIPAS-IMK, which we present here. We will clarify which different versions have been used for the H_2O, HDO and derived delta-D results for all instruments.

8) 'Standard Deviation' is already defined as the spread around some mean value, so 'debiased standard deviation' is just 'standard deviation' since we are already talking about mean differences (or 'the standard deviation of the relative bias' which is how it is described in the caption for Fig.4). Perhaps you thought SD might be confused with root-mean-square-difference? Also in Tables 1 and 2 this has become '1 \sigma Bias' which sounds like a different thing altogether, but I assume it's the same.

We will make sure to use the same technical terms throughout the paper. We have used the term "de-biased standard deviation" to make clear that it is the standard deviation around the mean difference (spread around the bias) and not the spread of the data sets themselves. What is called "1-sigma bias" in Tables 1 and 2 is meant to be the de-biased standard deviation. This will be updated, too.

9) The time series plots could be enhanced by subtracting out the mean profile and then also removing the average annual cycle to show interannual variability. The latter may have some QBO correlation which could be investigated.

Again, we are intrigued by this suggestion for additional analyses. However, they are beyond the focus of this current paper.

10) Details of bias determination (3.1.3, 3.2) are, firstly, confusing because of the repeated use of the same 4 coordinates for each parameter, secondly difficult to read because of the small fonts and, finally, quite standard. Even then, there are a few problems here Eq 1 presents the bias as 4-coordinate average of b_i which are themselves 4-coordinate quantities. However, it seems unlikely that any two measurements are exactly matching in latitude or time (it's not clear what 'period' means here) so I assume these coordinates are what is being averaged over in i=1, n so should not appear on the l.h.s. as well. And b_i is presumably also a function of longitude, also averaged.

We agree with the referee in the confusing use of 4 coordinates for each parameter. In the revised version of the manuscript the notation is simplified using the formalism of Dupuy et al ACP, 9, 287–343 (2009).

sigma_x1 and sigma_x2 in Eq (7) are undefined,
Will be fixed.

σ_b has a bar over the b here but not in Fig 4.
Will be fixed.

Fig 4 shows SD as a percentage (of what?) but Eq 5 defines this in absolute terms.

The debiased standard deviation is calculated to the mean relative bias, therefore the unit is percentage. Eq (5) is defined in relative terms, it will be clarified in the revised version of the manuscript.

When considering relative bias, if the two datasets were retrieved as log(VMR), the geometric rather than arithmetic mean of the two values would be more appropriate as the reference value.

Only one of the datasets is retrieved in log(VMR) so the arithmetic mean is applied in all cases.

I appreciate that this sort of thing was included in the previous papers (and had I reviewed those I would have said the same thing) but that's even less reason to include it again here. Everyone knows what you mean (and they can look up Pearson Correlation on Wikipedia) so no need to drag the reader through all the small print. On this (rare) occasion, it really is simpler to explain in words rather than equations.

For clarity on how the correlation calculation is performed, we have chosen to include this equation.

11) You could save a lot of wordage simply referring to these data as 'IMK', 'ESA' and 'ACE' (at least in the text, perhaps not in the figures)

Thank you for the idea. Since these abbreviations are not commonly used in the previous literature, we would prefer to maintain the references to the datasets as they are.

Minor/Typographical comments

Abstract

L139: Use a regular reference for the SPARC special issue to avoid the typesetting difficulties caused by putting the URL (http:...) in the text.

This is the standard method for referring to this special issue.

L185: 'from 1 to 70 km'

The referee is right! The word “since” will be changed by “from”.

L189: Presumably the two MIPAS datasets are automatically colocated, so this just refers to colocating ACE-FTS with MIPAS.

See methodology response above.

Fig 2: I assume the variation in low altitudes is due to cloud-screening but what causes the reduction in MIPAS-ESA comparison data at high altitude?

The reduced number of coincidences for the ESA profiles above 50 km is due to the fact that profiles from different observation modes are used, in particular measurements from the UTLS observation mode, which are about 8% of the total MIPAS observations and are performed mainly in the period August 2004/August 2005, are characterised by the altitude range 8.5 - 52 km.

However, as stated in the first comment, in the revised version of the manuscript we compare the same number of coincident profiles for the three databases. The number of coincidences in the stratosphere is 15263 and it decreases going to the lowest altitudes up to 4078 at 10 km.

L201: It would be useful to have at least an approximate figure as to what percentage of MIPAS profiles fail the quality control tests.

For MIPAS-IMK about 0.16% of all started retrievals of HDO did not converge. About 0.05% of all started retrievals failed or encountered corrupted spectra. In total 2,314,011 profiles were processed. This means that 99.79% of all profiles are considered healthy.

The two flags that we provide with the data (visibility flag needs to be 1, and averaging kernel diagonal needs to be > 0.03) are meant to be applied to single points in the vertical profile, i.e., these flags reduce the altitude coverage of one profile, but leave the other parts of the profile valid. Please note that we always provide the profiles from 0 to 120 km. The flags define the valid altitude range.

MIPAS-ESA uses a different approach for filtering out bad profiles.
The quality of the retrieved profiles is judged “good” when three requirements are met: the retrieved profile adequately reproduces the measurements (i.e. the chi-square value at the final step of the iterative procedure is smaller than a pre-defined mode and species-dependent threshold), there are no outliers in the retrieval error (i.e. the maximum value of the retrieval error profile is smaller than a pre-defined mode and species-dependent threshold) and the iterative retrieval procedure successfully converges.
When at least one of the previous requirements is not verified, the whole retrieved profile is flagged as bad in the output file (post_quality_flag=1) and it is not used as either profile of an interfering species or initial guess in subsequent retrievals. Otherwise, if all previous conditions are verified, the post_quality_flag is set to 0 so that the retrieved profile is considered “good”, and it can be used for subsequent retrievals. If the retrieved temperature is flagged as bad, no VMR retrieval is performed, since a proper temperature profile is fundamental for the retrieval of the trace species.

Each retrieved profile is properly and fully characterised on the full retrieval range provided in the output files by the corresponding CM and AKM. Altitude regions with poor information on the retrieved target can be identified by the low values of the diagonal elements of the AKM and/or the large values of the diagonal elements of the CM. Since the AKM and the CM are calculated considering the retrieval on the full vertical range, even if some of the retrieved values are discarded by the user, we recommend to use the full profile along with its full CM and AKM.

2.54 million HDO profiles are available in the products. The percentage of ESA good profiles is reported in Fig. 6 of Dinelli et al., 2021 paper for all retrieved trace species. 8% of HDO retrieval procedures fails.
Part of this information will be provided in the revised version of the paper.

L208: In L187 the grid is from 1-70km. Assuming the IMK grid is at 1km intervals how do you get 58 levels?

The MIPAS-IMK grid is not strictly a 1-km grid. It is a 1-km grid from 0 to 44 km, followed by a 2-km step width from 46 to 70 km. These are 58 levels. We will correct this description in the revised version.

L304: I would not refer to these as 'error bars', just 'bars'.

Will be changed. Thank you.

L319: Describing these values as 'global' minima is misleading, they're actually the minima in the intercomparison dataset.

The term "global" will be clarified in the revised manuscript.

L325: Similarly, 'global mean' profiles.

Also, here.

Fig 3. These plots would be more informative (and take less space) if all three datasets were combined on the same plot, allowing MIPAS-IMK and MIPAS-ESA to be directly compared. Use eg dashed lines to mark the 1sigma variation for each (Also, I wouldn't call these 'error bars' as in the current caption).

We agree with the reviewer suggestion, the three curves will be combined in the same plot. We also improve the figure 3 including the MIPAS-IMK to MIPAS-ESA comparison. Regarding the "error bar", we think that the referee is right, they are "bars".

L365: It is perhaps worth mentioning that the low SD (Fig 4(c)) between ESA-ACE and coupled with the high correlation (Fig 4(d)) are consistent with the MIPAS-IMK retrievals being less sensitive to actual atmospheric variations in H2O, and conversely for IMK-ACE for HDO.

We agree with this statement for altitudes above ~25 and ~15 km for H2O and HDO, respectively, and will add it to the revised version.

Fig 4(g) mis-labelled(?) as '\sigma_b Bias' (and similarly Table 1)

No, this notation is correct as it is the de-biased standard deviation.

Table 1 & 2: I assume these figures are a summary of the plots shown in Fig 4, in which case make it explicit in the Table caption (and similarly in the Fig 4 caption refer to these tables). Rather than have rows labelled eg 'MIPAS-IMK vs 'ACE-FTS' I suggest replacing 'vs' with $-$ to make it absolutely clear which way around you are defining the sign of the bias.

These changes will be made in the revised paper.

L393: Is there such a thing as a 1\sigma standard deviation? I though the SD was, by definition, 1 \sigma.

We agree with the referee. Our sentence is redundant. Will be changed.

Fig 5 caption refers to 'meridional cross-sections' which suggests a single slice through a particular longitude. 'Zonal mean' or 'Zonally averaged' distributions are the usual terminology for such plots. And rather than use 'summer' and 'winter' - which differ from north to south hemispheres - give the actual months averaged (as in Fig 6).

These updates will be made.

L405: I'm assuming that the labelling on Fig 5 is correct (I would have expected ACE-FTS to be missing the high latitude measurements during the local winter months but, on the other hand, they may also specifically target as high latitude as possible during polar winter months). However, the very low ACE values for both H2O and HDO in the winter polar vortex (compared to the two MIPAS datasets) are worthy of comment.

Yes, this figure is correct. ACE-FTS is in an orbit that targets high latitude measurements, more than 50% are at latitudes higher than 60 degrees, to investigate polar ozone chemistry. Note, that the local winter mean from ACE-FTS does not include data from all of these months because of the requirement for sunlight for its measurements. This requirement leads to the ACE values sampling only the later part of this season (in austral winter, mainly Aug. in JJA at the highest latitudes) than the two MIPAS datasets sampling all months. It is likely this sampling difference that leads to ACE-FTS showing more dehydration than MIPAS in these zonal mean plots.

Also, why is the IMK cloud filtering any different for HDO than for H_2O?

Lossow et al., 2020 have made sensitivity tests regarding the retrieval of HDO from MIPAS data. They found that for the HDO retrieval, the upward error propagation was pronounced at the lower end of the profiles in previous data versions. I.e., incorrectly retrieved vmrs due to, e.g. unidentified cloud contamination trigger retrieval errors in the vmrs above this altitude. To be on the safe side, Lossow et al., 2020 recommended to discard the retrieved values in the altitude range of the lowest two (V5H) to three (V5R) tangent altitudes. They found that the propagated error fades out sufficiently above this level so that data from levels above can be used.

L450: 'interannual variability' implies what's left after removing the annual cycle. From Figs 7a,b,c alone it is not clear to me that ACE-IMK have the greatest similarity.

The referee is right, we are not analyzing the “interannual variability” (seasonal cycle subtracted from the data sets) but the time series themselves. This sentence will be changed.

The referee is also right again, while the tape-recording effect is clearly seen in the MIPAS-IMK HDO time series, this is less evident in both MIPAS-ESA and ACE time series.

L465: 'shows' rather than 'shown'.

Will be changed.

L467: 'while the ACE-FTS instrument can measure with higher sensitivity ...' What is the basis for this statement?

ACE-FTS has a higher sensitivity due to the combination of the long-path length through the atmosphere in limb-view and the solar occultation measurement technique used. This makes the ACE-FTS measurements less susceptible to perturbations due to thin cirrus clouds in the UTLS.

L485: 'vertical behavior' (or 'behaviour') presumably just refers to the small mean profile bias averaged over the whole dataset, shown by the red-lines in Fig 4a,b. 'Behaviour' suggests that the two track each other well in other respects such as low SD and high correlation, which they don't (at least not above 20km).

This sentence will be clarified in the revised manuscript.

L487: 'according to the uncertainties' Which uncertainties are these? There was no reference to estimated uncertainties in the datasets in the main part of the paper.

Given that the two MIPAS profiles are 1ppmv higher than ACE after averaging over several thousand profiles suggests rather a significant discrepancy to me.

As mentioned above, we will provide all available information on uncertainties in the revised paper.

L495: 'quantitaty'?

Will be fixed.

L525: 'MIPAS-IMK dataset provides a more realistic signal for the entire stratosphere' This is a bold statement and needs some qualification otherwise it could get quoted out of context. Is this for H2O, HDO, delta D or all three? Is it in terms of the time-evolution or mean profile? Is it because it has the best correlation or lowest SD compared to both other datasets?

“More realistic than MIPAS-ESA” was to be meant and below the 30 km of altitude. However, we agree with the reviewer and even with all the context the statement can be a bold given the existing data. We reword this paragraph in the revised version of the manuscript.