Comment on amt-2022-114
Anonymous Referee #3


Review of amt-2022-114

The manuscript provides a validation overview of the MIPAS ESA operational products using measurements from a balloon version of MIPAS. Overall I found the manuscript is well-written, easy to follow, and provides information that is relevant to users of the operational MIPAS data. I would recommend publication after a few issues are considered. I have a few minor corrections and a larger concern about how the error analysis is performed, but I don't think accounting for any of my suggestions will be particularly difficult for the authors.

General Comments:

My main concern is how the combined error (Eq. 3) is calculated when comparing the two instruments. One issue is that this seems to neglect any potentially correlated error source between the two instruments. The classic example is a spectroscopic error, where if both retrievals use the same spectroscopic database this equation will overestimate the combined error, but there also may be correlated effects from non-LTE errors or other effects absent in both forward models. I understand that these errors may not be perfectly correlated between the two instruments due to differences in retrieval methods (I see different microwindows were mostly used), spectral resolution, etc., but if the dominant source of estimated systematic error between the two measurements is a potentially correlated error like a spectroscopic error it draws into question some of the conclusions made. It is not easy to correctly account for these correlations, but I would suggest at
least stating what the dominant source of systematic error is for each case and analyzing if it is potentially correlated between the two instruments. Since some of the main conclusions of the paper are to recommend caution in areas where the observed differences between MIPAS-E and MIPAS-B is larger than the estimated systematic error it is critical that the estimated systematic error is interpreted correctly.

A full validation of every species measured by MIPAS is a monumental task, and the MIPAS-E to MIPAS-B comparisons done by the authors is one piece of that puzzle. This is fine, I don't think the authors need to include more data or analysis, but as a naive MIPAS-E data user some of the results are hard to interpret on their own. The main takeaway that I get is that I should go read the MIPAS product quality document instead (of which a version of this manuscript serves as input to). Once again, this is not a problem by itself, but the manuscript could use some further explanation on how this work fits into the larger body of MIPAS validation efforts.

Specific Comments

Section 2.1 l. 68: "hereinafter also referred to as MIPAS-E..."

There are some places where simply MIPAS (mostly the remainder of this section) is used to refer to MIPAS-E. I would recommend always using MIPAS-E when referring to MIPAS on ENVISAT for clarity.

Section 2.1 l. 78: "in steps of 3 km below 45 km"

What about above 45 km?

Section 2.1 l. 94: "... an equivalent improvement in the vertical and horizontal (along-track) sampling"
I understand details of the sampling for the FR/OR modes of MIPAS can be found elsewhere, but the horizontal/vertical sampling of each mode should be stated in this section. Especially since the change in MOPD is stated.

Section 2.1 l. 118: "All molecules except HDO ..."

Is there a fundamental reason why HDO could not be validated as well?

Section 2.2:

Has there been any validation of the MIPAS-B data products separate from MIPAS-E that could be mentioned here?

Section 2.2 l. 125: "... MIPAS-B performance is superior, in terms of NESR ..."

Is the improvement in NESR from the averaging spectra or is there an instrumental difference that provides better NESR?

Section 2.3 l. 200: "A bias between both instruments is considered significant if the SEM is smaller than the bias itself."

Should twice the SEM be used here instead to be at the ~95% confidence interval?
Section 2.3 l. 204: "Since the vertical resolution of the atmospheric parameter profiles of both instruments is of comparable magnitude, a smoothing by averaging kernels has not been applied to the observed profiles"

I assume that the error estimates for both instruments also do not include the classical "smoothing error"? If it is then this would cause the error estimates to be inflated since both instruments have a similar vertical resolution.

Section 3 l. 226: "Trajectory matches are based on diabatic 2-day forward and backward trajectories with a collocation criterion of 1 h and 500 km as described in section 2."

Is it possible to demonstrate how well the trajectory matching is working? If I understand correctly there are conditions where the measurement locations are collocated enough that trajectory matching is not necessary, maybe this can be used to show the effectiveness of the trajectory matching.

Section 3.1 l. 245: "... although the standard deviations exceed the expected precision ...

Could this be because the trajectory matching introduces some variance into the comparisons as well? Even if the trajectory matching is perfect the collocation is still only within 500 km and 1 hr, which would contain some atmospheric variability.

Section 3.1 l. 252: "A possible reason for this difference between both MIPAS sensors could be an inaccuracy in the altitude assignment..."

Presumably this error is included in the error budgets of the instruments, you should be able to quantify if this could actually be the case.
Section 3.2 l. 259: "FR and OR mode comparisons show different vertical shapes of the differences between MIPAS-E and MIPAS-B"

Is there a significant difference between the retrieved vertical resolution of H2O in the FR and OR modes? Particularly with the strong altitude gradient of H2O a small change in vertical resolution could cause a large observed difference. In general for every species I wonder how much of the difference between the two modes can be explained from the changing averaging kernel.

Section 3.6:

Are there any estimates of how much error could be introduced due to an imperfect photochemical correction? I'm wondering if there could be some effect where the balloon flights tend to occur at a similar time each day and so you don't average over an ensemble of random SZA differences, but I'm just throwing things out there.

Technical Corrections

Section 3.1 l. 250: Is MIPAS here referring to MIPAS-B?