We thank the reviewers for their careful and constructive comments on the paper. Please find our answers below. While the comments are given in standard font, our answers are given in italic. Added text is shown in red for clarity.

Anonymous Referee #1

The paper by Engel et. al. provides a coherent validation of four trace gases (CH4,N2O, CFC-11 and CFC-12) from the MIPAS satellite experiment on Envisat with 7 flights of the ballon-born BONBON instrument. BONBON provides precise in-situ measurements of the above species. The validation method is clearly described and adds an important piece of information on the quality of the operational MIPAS (v6.0) for the mentioned trace gases.

Some detailed comments:
p 7460 line 11: Mission started in July 2002, not March 2004 (end of first phase) line 18: due to problems with the scanning mirror. Replace with due to problems with Interferometer Drive Unit. The scanning mirror is a different part of the optics, adjusting the LOS of the instrument.

This will be changed to changes to "Interferometer drive unit"

p 7462 paragraph 2.4 MIPAS-E has a 3 km vertical spacing and therefore vertical resolution in the measurements (for the considered altitude range), but trajectories are calculated every 200m or 20-50 m, respectively. So you probably have interpolated MIPAS-E to the higher resolution for calculating the trajectory starting points. Please give more details here for full insight in the method.

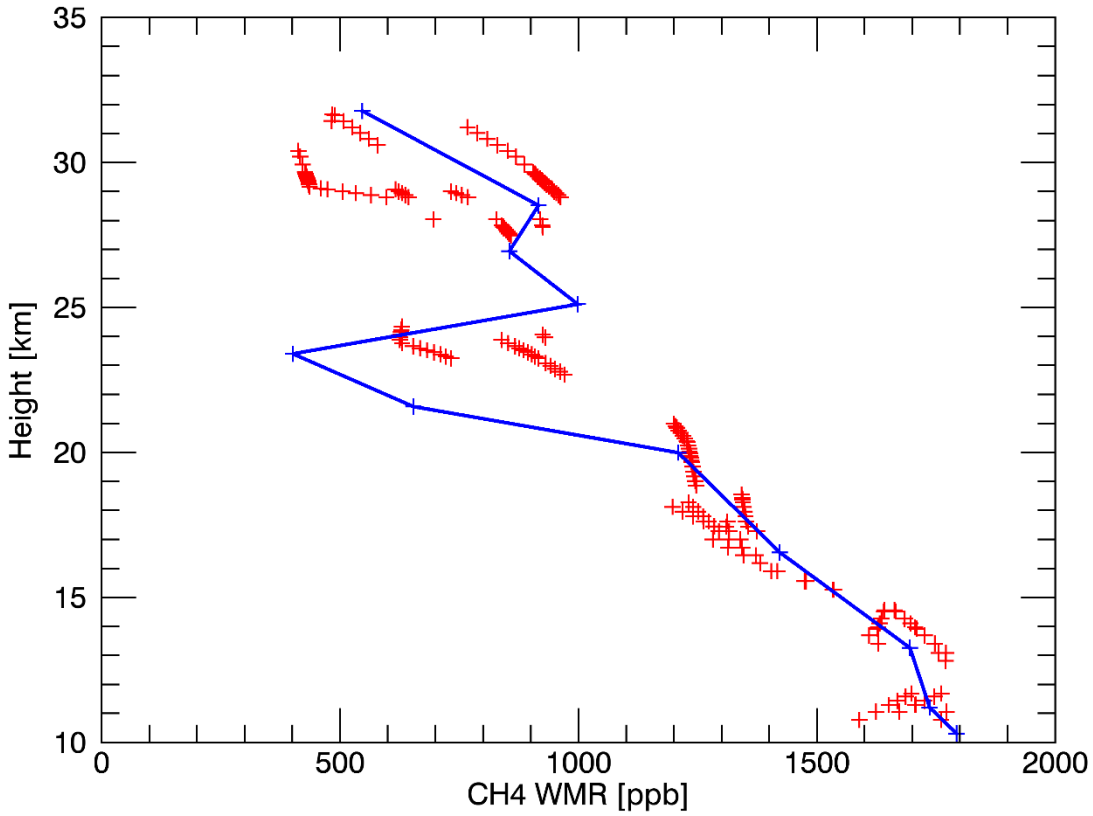
The reviewer is correct in pointing out that the trajectories have a higher resolution than the MIPAS data in some cases also higher than the cryosampler data. We will change the text in section 2.4. as follows:

"Trajectories are initiated about every 200m in altitude (for flights prior to 2009) or every 10 s, i.e. about every 20–50m (for flight in 2009 and 2011) along the flight path of the balloon and run forward and backward in time for five days (Grunow, 2009). MIPAS-E data are interpolated to the respective potential temperature of the trajectory at the point of the match."

p 7464 paragraph 3.1
Here you discuss the sharp structure in the BONBON profile, not visible in the MIPASE mean profile of the matches. You state, a few of the matches indeed show this structure. Can you provide a plot with the individual matches to give an impression? Is the structure better visible for near-by matches with short trajectories? This would support your guess, that too long trajectories cannot follow such small and probably short-living structures.

In order to test this we have repeated the validation with more stringent match criteria. When doing this, there is indeed an indication of the structure, as mentioned before. However, as is obvious from the direct intercomparison plot, the sharp gradients and the very restricted number of matches under these criteria (maximum of 2 days trajectory run time;

200 km distance and 1 hr time mismatch) make it impossible to use these data for validation. The following Figure will be added to the paper (as Figure 5).



We will add the following text to the manuscript at the end of section 3.1. (p 7465):

"This rather sharp feature is a good example of the restriction of applying the match method to satellite validation, especially fora limited amount of validation data. In order to demonstrate this, we show the direct validation results (interpolated MIPAS data) for all trajectories which fulfill more stringent validation criteria (maximum difference 200 km and maximum trajectory travel time of 2 days). Figure 5 shows that in this case the meteorological feature is indicated also in the satellite data, but that the resulting number of matches is rather low and therefore a quantitative validation is not possible forthese more stringent match criteria. Hence,the validation results for the other species will also be given disregarding the altitude region between 20 and 25 km during flight B45. "

Figure 6./10./14. :
Please use :As left panel of Fig. 2., ...

done

Anonymous Referee #2

This paper provides an important evaluation/validation for key trace gas products fromthe MIPAS-E satellite measurements, specifically the operational ESA retrieval versionML2PP/6.0. It is essentially a comparison study of data from BONBON, a balloonbornecryosampler instrument, and the MIPAS-E operational retrieval products CH4, N2O, CFC-12, and CFC-11. The comparison makes use of a trajectory matching techniqueto increase the number of coincidences between the balloon and satellite data.

A total 7 balloon flights are used for comparison.

The BONBON data are of high quality and can be considered a useful standard forvalidation of the satellite data. Thus, the paper represents an important contribution tothe scientific community that uses MIPAS data in their analyses. This paper may besuitable for publication in AMT, provided that the major comments below are adequatelyaddressed.

Major Comments

1. There are a number possible uncertainties that can creep in when comparing in situand satellite vertical profiles. First, even in the case of a direct satellite overpass, oneusually has to consider the impact of vertical averaging kernels in comparing satelliteand balloon vertical profiles. This is particularly important for species with large verticalgradients, such as CFC-11. There is no discussion of the effect of MIPAS averagingkernels on the comparison. Second, the impact of trajectory errors on the comparisonshould be explored. These may turn out be small, but they should be quantified. Forexample, even at 1.25 degrees , it is unlikely that the initialization point exactly lines upwith the balloon location, so how do the results differ if one chooses the next-nearestECMWF gridpoint for initializing the back- and forward-trajectories? In the vertical, it ispossible that using climatological heating rates could lead to under- or over-estimationof vertical displacements. Although the manuscript states that these deviations aresmall, for gases such as CFC-11 with large vertical gradients, even small vertical displacementscould have a large impact on the matched CFC-11. Finally, the choiceof 500 km for spatial coincidence should be justified. Presumably, it is a compromisebetween the quantity of matches and trying to sample the same airmass. The paper should discuss what kind of results were obtained using other coincidence criteria (e.g.250 km, 1,000 km, etc.) to give an indication of how that impacted the number ofmatches and whether or not that changed the mean differences.

The reviewer touches some critical and important points here, by pointing out that uncertainties in the method should be discussed in more detail. The major points mentioned are trajectory mismatches and vertical resolution.
A)match criteria.
The match criteria have indeed been chosen to provide a compromise between having a sufficient number of matches and yet still having comparable air masses. As shown by the example of flight B45, the current values of 500 km and 1 hr are already not sufficiently stringent to completely exclude wrong matches. By reducing the criterion to 250 km only very few matches were found in some cases, so that a systematic validation was not possible anymore. These considerations led to the choice of 500 km. Due to the low number of matches when using more stringent criteria, a comparison of the results is not possible.This would need to be done in a more systematic approach focusing on the matching technique

itself, which is not the emphasis of this paper. We will add the following to the paper (p. 7462, l 17).:
"These values were found to give the best compromise between finding a sufficient amount of matches and ensuring that differences in air mass were sufficiently small."

B)trajectory uncertainties.
The influence of the vertical uncertainties in the 5 day trajectories due to the use of different heating rates is on average about 120 m in the tropics and about 170 m in high latitudes. For singles trajectories,vertical displacements can be as high as 500 to 700 m. We have tested the influence of using actual and climatological heating rates and found no significant differences for the validation of N2O and CH4. We will add the following to the manuscript:
"The choice of using climatological heating rates or heating rates calculated specifically for the trajectories was found to give no significant differences, as tested for the validation of N2O and CH4 (Grunow, 2009). When using isentropic trajectories vertical differences to trajectories using explicitly calculated heating rates were on the average 120 m in the tropics and 170 m in high latitudes during winter (Grunow, 2009). "

C)averaging kernels.
The reviewer is correct in pointing out that averaging kernels should be taken into account in satellite validation studies. The cryosampler data and the satellite data have different vertical resolution. In the case of the cryosampler this resolution varies from sample to sample and in addition no information is available between the samples. A degradation of the vertical resolution of the cryosampler to those of the satellite measurements by applying the averaging kernels of the satellite retrieval is thus not possible, as no information is available between the discontinuous samples. We have therefore restricted our analysis to the "geophysical validation", i.e. without taking into account sampling issues of the instruments. In this respect we suggest to also change the title to "Geophysical long term validation …. " . We will add the following to the manuscript (p 7463, l 9):
"The two data set have different vertical resolution. The vertical resolution of the satellite is given by the averaging kernels. In the case of the cryosampler, the vertical resolution varies from sample to sample based on the vertical speed at which the balloon was travelling and the sample integration time. As no informationis available for the cryosampler measurements in between the samples (discontinuous sampling) a degradation of the vertical resolution of the cryosampler to those of MIPAS-Envisat was not possible. This paper thus only presents a validation of the geophysical results of the satellite retrieval."

2. The paper does a good job at separating the comparison between the high spectral resolution, "older" dataset, and the lower resolution "younger" dataset. But given thedifferences in latitude of the 7 balloon flights, and the fact that the two tropical flights(B42 and B43) occurred during the younger MIPAS phase, it is possible that some ofthe changes seen between the older and younger data comparisons may in fact bedue to latitude effects. The analysis of B42-B46 could be split into low- latitude andhigh-latitude comparisons. It may not be necessary to show results in a figure, but itwould allow the authors to comment on any differences with respect to latitude, whichcould be significant given the very large contrast between low- and high-latitude verticalprofiles.

The reviewer is correct in pointing this out. Unfortunately, the amount of data do not allow for a significant validation in different latitude regions. As we found the most significant

changes to occur with time, we have chosen to present the data in this way. We will add the following to section 3. (p.7463, l.25) of the paper to make this clearer.

"The flights were also conducted in different latitude regions. However, we found the variations with time (low spectral resolution vs. high spectral resolution) to be the dominant factor of variability. As the total number of flights was not sufficient to derive systematic differences also for different latitudes (and seasons), we have chosen to restrict the comparison to differentiate between different spectral resolution."

3 Units for differences: The significant differences between MIPAS and BONBON arereported in mixing ratios (ppb or ppt), but since all of these gases have steep vertical gradients, it would also be useful to report any significant differences in percent values such as (MIPAS-BONBON)/BONBONx100%. Presumably the mean % differencescould be calculated in a straightforward way. It may not be necessary to includeadditional figures, but the major points in the conclusion ought to include percentages.For example, the underestimation of N2O by 20 ppb around 15 km and of CFC-12 by50 ppt at 10 km in the younger MIPAS data should also be reported as mean % differences.Also for CFC- 11, it would be good to note how the percent difference growrapidly from 10-20% around 15 km to a few hundred percent or larger at 25 km.

We thank the reviewer for pointing this out. Indeed, a difference of x ppb will be less significant at higher mixing ratios than at lower mixing ratios. Percentage differences will be added where a quantitative comparison is made.

4. Comparison to other retrieval algorithms: As noted in the manuscript, there are a number of different analysis algorithms for retrieving CH4, N2O, and the two majorCFCs from MIPAS data. This can lead to confusion and uncertainty in applying MIPASproducts to science investigations. If there has been an intercomparison study betweenretrieved products, then it would be good to include a reference and make some generalstatements about the implications of the results in this paper to the other retrievals. ForCFC-11 in particular, since this study does not recommend its use in scientific studies,it would be useful to know whether that applies to the other retrievals as well.

This is in principle a good suggestion. However, dedicated papers to the validation of the "scientific" retrievals at KIT Karlsruhe are being published for Methane (Laeng et al.,2015) and for CFCs 11 and 12 (Eckert et al., 2015). A comparison between the different algorithms is beyond the scope of this paper. We will add the following at the end of section 1 (p. 7458, l. 28)

"The validation of other retrieval algorithms of the MIPAS data for CH4 and CFCs 11 and 12 is subject to further studies (Laeng et al., 2015 and Eckert et al., 2015)."

Minor Comments

1. section 2.3, line 16: "where" should be "were"
done

2. section 2.3, line 25: “winter” should be “winter/spring” for 9 march and 1 april
done
3. section 2.4, lines 17-20: The sentence is unclear. Should the “matched altitude interval of the satellite data be *less* than 1.5 km?
Indeed, this was unclear. We have rephrased as follows for more clarity:
“In addition only data have been used for validation where matcheswere obtained for at least 4 trajectories covering an altitude of at least 1.5 km. In this way we ensured that a match cannot be obtained from a single trajectory as single trajectories may show significant uncertainties. (Schwarzenberger,2014).”

4. section 3, line 27: “...in the in situ *data* is sufficiently small...”
Has been added
5. section 3.4, lines 27-28: It appears from Figure 14 that that for the younger data, the CFC-11 differences below 20 km can be explained with the measurement uncertainty, so this statement seems to be too general.

The reviewer is correct. We have rephrased as follows:

“The differences between satellite and balloon mixing ratio cannot be explained within the measurement uncertainty for the older data with high spectral resolution. While the CFC-11 mixing ratios of the MIPAS derived with this retrieval for the younger low spectral resolution data still overestimates the balloon data, the agreement is within the uncertainties below 20 km altitude.”

Anonymous Referee #3

This paper is fairly well written, scientifically sound, and presents material that is appropriatefor Atmospheric Measurement Techniques. However, one major issue, concerningvertical resolution matching between the instruments, does need to be addressedbefore it should be considered for publication. There are a few other minor details,listed below, that also need to be addressed.It seems that no effort was made to match the vertical resolutions of the two instruments,which, in some regions, will undoubtedly have a significant impact on the comparisonresults. The cryosampler data needs to be smoothed either by the MIPASaveraging kernels or by convolving with the MIPAS vertical resolutions for accuratecomparisons. This would most likely diminish the effect of the dynamical feature in flight B45, and help explain why some comparison results are better in the “reducedresolution” measurements (as the reduced spectral resolution naturally led to improvedvertical resolution).The discussion of comparison results tends to be solely qualitative. Throughout theresults discussion, phrases like “excellent agreement” and “large variance” should bequantified (e.g. within $\pm$1 ppb, or up to 10 ppt, etc.)

The reviewer brings forward an important question, which is the different vertical resolution of the two instruments. Unfortunately, the application of MIPAS averaging kernels to the cryosampler data is not possible due to two reasons. Firstly,the cryosampler data themselves present averages over a certain altitude region which may be up to 1.5 km and this resolution varies from sample to sample. Secondly, the sampling of the cryosampler is irregular and discontinuous, which means that we have no information on the vertical profile in between the samples. A degradation of the vertical resolution of the cryosampler to those of the

satellite measurements by applying the averaging kernels of the satellite retrieval is thus not possible. We have therefore restricted our analysis to the "geophysical validation", i.e. without taking into account sampling issues of the instruments. In this respect, we suggest to also change the title to "Geophysical long term validation …. " . As noted in the reply to reviewer 2, we have added the following at the end of section 2.4 to the manuscript:

"The two data set have different vertical resolution. The vertical resolution of the satellite is given by the averaging kernels. In the case of the cryosampler, the vertical resolution varies from sample to sample based on the vertical speed at which the balloon was travelling and the sample integration time. As no information is available for the cryosampler measurements in between the samples (discontinuous sampling) a degradation of the vertical resolution of the cryosampler to those of MIPAS-Envisat was not possible. This paper thus only presents a validation of the geophysical results of the satellite retrieval."

The second issue mentioned here is the use of terms like excellent agreements and large variance. We will rework the manuscript to quantify such statements and give deviations in numbers (both absolute and relative).

Technical issues

Throughout, dashes are used after instrument names that aren't needed, e.g. "MIPASE-profiles or CH4-validation" should just be "MIPAS-E profiles or CH4 validation".

done

Throughout, terms "older" and "younger" data are used to refer to data prior to early2005 and after this time, respectively. These should probably be changed to "earlier"and "more recent".

As noted below, we will use the term low spectral resolution and high spectral resolution. Where we use a temporal notation we will change to earlier and more recent.

Similarly, this data is referred to as high resolution and low resolution. It should bemade clear throughout that this refers to spectral resolution, as the degradation yieldedmeasurements with a finer vertical resolution (which as previously mentioned needs tobe discussed and properly dealt with). Alternatively, these could be called originalresolution and optimized resolution.

We thank the reviewer for this suggestion. We have changed the notation to specify spectral resolution throughout the manuscript.

The name of the university should be constant throughout. Sometimes it is referred toas the University of Frankfurt, and at other times University Frankfurt.

Changed to University of Frankfurt

Throughout the almost equal sign (≈) is used when it should probably be the approximately sign (∼).

changed

P7457 lines 4-5: “MIPAS-Envisat” should probably be moved to between “the” and“operational”.

P7457 line 11: “as of” could be “in”

P7457 lines 11-15: Please put in specific results

P7457 line 19: “traces” should be “trace”
changed

P7457 lines 20-22: the different types of studies should all have references, not just relative lifetimes.

We have added references for the use of tracers for the other studies mentioned.

P7457 lines 22-23: what is meant by “relative changes”? And does “gradients” mean vertical gradients?

Relative changes means that no absolute numbers are required. Gradients can be both vertical and horizontal. As this statement is not really necessary but adds confusion we have deleted it and only state that high accuracy and precision are needed for tracer studies and budgets. We rephrase as follows:

“Such studies using tracers require high precision and accuracy in order to derive quantitative information.”

P7458 line 10: “should” instead of “can”

Has been changed

P7458 lines 11-12: “: : :one single vertical profile.” This assumes that the satellite onlygets one profile in the time coincident with the in situ measurements. It should probablybe something like “However, this typically can only be achieved for a single verticalprofile from a given satellite data set.”

We have followed the reviewers suggestion and rephrased accordingly.

P7458 line 14: May want to define, or be more specific than, “trajectories”

We have specified “backward and forward air mass trajectories”

P7458 line 16: MIPAS has yet to be defined

Thank you, we have added the explanation of the acronym in parenthesis.

P7458 line 24: “that spans nearly” instead of “for about”

changed

P7458 line 27: insert “MIPAS” between “the” and “data”

done

P7459 line 2: insert “comparison” between “the results”

We have added “of the comparison” behind results: “… we present the results of the comparison for …”

P7459 line 6: MIPAS should have already been defined.

Yes, we have eliminated the full name and only used MIPAS here.

P7459 line 12: “to combine” should be “for both”
changed

P7460 line 14: “scales” instead of “scale”

done

Figure 1 caption: “corresponding” might be better than “related”

Thank you, changed.

Figure 4 caption: The first sentence needs to be clearer, i.e. state what it is that iswithout this data. Also, it would be helpful to give the altitude range of the feature.

The reviewer is correct; the first part of the figure caption is unclear. We have changed the first sentence as follows:

“Deviations between MIPAS and cryosampler data after elimination of meteorological feature in profile B45 between 20 and 25 km altitude.”

Figures 5-15: should start with “Same as”

done

Figures 6, 10, 14: “Same as left panel of: : :”

done