We would like to thank the reviewer for his/her constructive comments and suggestions. Below we reply to the raised issues point by point.
Figure numbers refer to the discussion version of manuscript and supplement.

Major comments:
1. The inherent question in any STS algorithm is whether features in the V* field belong in the stratosphere or troposphere. With its relatively smooth stratosphere (evident in Fig 3), STREAM correctly identifies some structures in unpolluted regions as tropospheric (e.g. tropospheric transport events). But it also folds some synoptic-scale stratospheric features into the troposphere. To clarify such ambiguities, it is essential to test an STS algorithm against truth by applying it to synthetic data. The authors perform such a test and show the effective retrieval errors. However given its importance, this section of the paper seems a bit brief. It would be useful to see how the weighting factors and convolution kernel sizes can be tuned to improve the retrieval. The section on parameter sensitivities indicates the magnitudes of the effects of parameter adjustments, but it could go further in determining which are actually better and why the authors believe STREAM's method is optimal. For example, could an increase in wcld (Fig. S9) help counter the low bias in T* seen in Fig S 18?

Reply: The correct identification and removal of synoptic scale stratospheric features is indeed quite challenging for all algorithms based on modified RSM. We have performed the stratospheric estimate on synthetic data in order to quantify this effect and other possible shortcomings, like the tropospheric background, and generally the accuracy of STREAM. We agree that the idea of optimizing the weighting factors for the synthetic data is quite tempting. As stated by the reviewer, such an approach would indeed favor a larger w_{cld}, which would also increase the tropospheric background in the Pacific for OMI (see Fig. S11). However, we refrain from this approach, as the results for SCIAMACHY and GOME-1/2 would become worse due to the different dependencies of total columns on cloud fraction. The chosen definitions for weighting factors therefore remain somewhat arbitrary. However, as demonstrated, the impact of changes of these definitions on resulting TR is small.
In the revised manuscript, we have extended section 4.3 according to detailed comments of reviewer 2. In addition, we conclude the section with a new paragraph:
"The application of STREAM to synthetic data thus provides a valuable estimate of the algorithm's accuracy. One might think of using the synthetic data for optimizing the definition of weighting factors being the next step forward. However, we refrain from doing so due to some contradictory results for different instruments. Concretely, the remaining bias in TR for synthetic data of about 0.1 CDU could be further reduced by increasing w_{cld}. This, however, has adverse effects on SCIAMACHY and GOME results (see sections 5.3 and 5.4)."

2. Stratospheric NO2 shows significant diurnal variation across an orbital swath. This is evident in the daily OMI DOMINO and NASA stratospheres when plotted orbit by orbit. From the description of STREAM, it is not apparent to me that the algorithm accounts for this, and therefore some of the stratospheric variation in each orbit will be aliased into the troposphere (this is also a problem in the RSM and in some of the other MRSM approaches). Please give an estimate of the magnitude of this effect and discuss the consequences of ignoring it, as well as the handling of orbital overlap – especially at mid- and high latitudes. A somewhat related question is how STREAM treats OMI pixels near the swath edges. Are these weighted differently?

Reply: As stated in section 2.3, each satellite pixel is assigned to one grid pixel (i,j) according to its center coordinates.

The satellite pixels are all treated equally, independent of viewing angle or pixel size. However, since the center of OMI's swath contains many more satellite pixels per grid pixel than the swath edge, the smaller center pixels dominate the weighted convolution procedure. In case of orbital overlap, as for OMI, a grid pixel might contain satellite pixels with different local time. This effect is neglected in STREAM. According to the reviewers request, we have estimated the magnitude of this effect by calculating the mean total NO2 VCD separately for the different OMI viewing zenith angles (directly related to local time). We refer to this aspect in section 2.4 (Data processing) of the revised manuscript, and have added this investigation to section S 2.4 of the Supplement:

Section 2.4:
"STREAM estimates stratospheric fields and tropospheric residues for individual orbits, using NO_2 measurements of the dayside of the orbit. Note that the effect of changes of local time on stratospheric NO_2 across orbit is generally low (see section S2.4 in the Supplement) and is thus neglected within STREAM."

"*S 2.4 Data Processing*
Diurnal variation of stratospheric NO_2

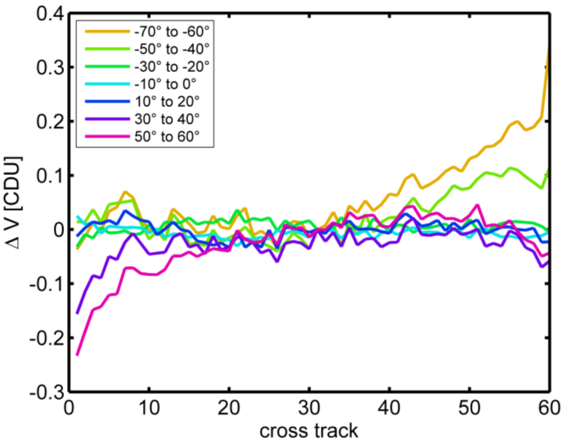

Fig. A: Dependency of OMI total VCD V on cross-track position (compared to nadir) at different latitudes in January 2005.*

In case of orbital overlap, as for OMI, a grid pixel might contain satellite pixels with different local time. This effect is neglected in STREAM. We have estimated the magnitude of this effect by calculating the mean total NO_2 VCD separately for the different OMI viewing zenith angles, which are directly related to local time (LT) (see Fig. A).
For low latitudes, the effect is negligible. Only for high latitudes (>50°), the effect can exceed ±0.2 CDU at the swath edges. Consequently, if individual orbits are considered, a small cross-track dependency of TR could be observed for high latitudes, which is actually caused by the LT dependency of stratospheric NO_2. For gridded data, however, where OMI orbits significantly overlap, the effect (and its impact on STREAM performance) is generally negligible."

3. In section 4.2 the authors describe a number of sensitivity studies. But sensitivity to the pollution weights should also be included here since it seems likely these weights would have a large effect on the retrieval.
Reply: We have performed a sensitivity study on the impact of w_{pol}:

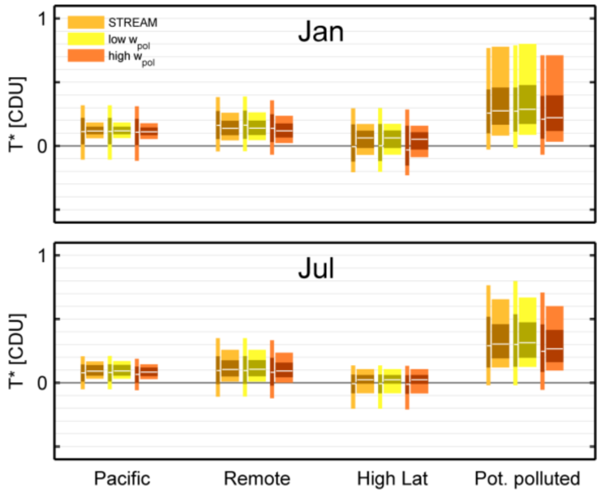

Fig. B: Regional statistics of OMI TR from STREAM for different settings for w_{pol} for January (top) and July (bottom) 2005.

The figure is included in the revised supplement (section S4.2.6).
In the manuscript, we have added

"4.2.6 Impact of pollution weight
The impact of pollution weight is investigated by multiplying w_{pol} (where different from 1, compare Fig. 2(a)) by 0.1 ("low w_{pol}") or 10 ("high w_{pol}"). In the first case, the resulting pollution weight over most continents is below 0.01, while in the second case, it is increased to 1 (meaning that w_{pol} is switched off) except for industrialized pollution hotspots.
In remote regions, the change of w_{pol} has almost no impact. In potentially polluted regions, the impact is only moderate as well. Low w_{pol} does not differ much from the baseline, as the latter already assigns rather low weighting factors to potentially polluted pixels; a further decrease by factor 0.1 thus does not change much.
Only for high w_{pol}, a significant change of TR can be seen; in this case, the inclusion of more partly polluted observations causes a high bias in the stratospheric estimate and the resulting TR are biased low by almost 0.1 CDU in winter."

Additional comments:
1. (pp 5-6) I think the algorithm description needs a more detailed outline than the 2-step version given at the top of page 5. Please list each step explicitly, including the names of the three weighting factors, their product and the iteration involving the TR weights. This would improve clarity, since the order of presentation in the paper is not exactly the order of data processing.

Reply: We have slightly extended the algorithm description in section 2:
"STREAM consists basically of two steps:
1. ~~a~~ A set of weighting factors ~~are~~ is calculated for each satellite pixel: a "pollution weight" that reduces the contribution of potentially polluted pixels, a "cloud weight" that increases the contribution of cloudy observations, and a "tropospheric residue weight" that adjusts the total weight in case of exceptionally large or negative tropospheric residues. The product of these weighting factors determin~~ing~~es to what extent the associated NO_2 total columns contribute to the estimated stratospheric field (Sect. 2.2). ~~, and~~
2. ~~g~~Global maps of stratospheric NO_2 are determined by applying weighted convolution (Sect. 2.3)."

This introductory paragraph is meant as high-level overview and thus does not provide more details, but refers to the respective subsections.

2. (p 8) In section 2.2.3, setting-description #3 says that the minimum size of a region for considering a TR weight is a grid pixel and the immediately surrounding pixels. Is

this also the maximum size?
Reply: Within the TR iteration, the criteria 1-3 are checked for each grid pixel ij. I.e., the size of the region that will be modified by w_{TR} is not restricted and can be quite large, as long as the 2nd criterion (|T*|>0.5 CDU) is fulfilled as well.
We have modified the formulation of criterium 3 in the revised manuscript to make the motivation for this additional condition more clear:
"*3. w_{TR} is only applied for grid pixels where the adjacent grid pixels exceed the threshold as well. By this additional condition it is guaranteed that a single outlier in the satellite measurements cannot trigger w_{TR}, as any satellite measurement is assigned to one grid pixel (see section 2.3).*"

3. (p 12) In section 3.1.1, the current version of the OMNO2 data product should be v2.1, not v3.
Reply: The data files we have downloaded from http://mirador.gsfc.nasa.gov/ all include "v003" in the filename, and on http://mirador.gsfc.nasa.gov/collections/OMNO2__003.shtml, it is explicitly stated that the dataset version is 003. We have thus modified "*v3*" to "*v003*".

4. (p 12) Also in 3.1.1, the authors claim that new SCDs from van Geffen et al. (2015) and Marchenko et al. (2015) would not affect the TR. This is not true: the smaller SCDs will also decrease the retrieved tropospheric amounts.
Reply: Note that we did not claim that the updated SCDs will not affect the TR, but the performance of STREAM.
We have modified the section to
"*However, such an overall bias will be interpreted as stratospheric feature by STREAM and thus does not affect its ~~the~~ performance ~~of STREAM with respect to tropospheric residues~~. (the same holds for the operational NASA and TEMIS STS algorithms). Still, the resulting TRs are expected to decrease slightly as the bias decreases for larger SCDs (see Marchenko et al. (2015), Figure 3 therein).*"

5. (p 15) As described in section 4.1, Fig. 3 shows V* for two days along with the STREAM stratosphere. Please also show also the stratospheres for RSM, OMI DOMINO and OMI NASA on the same days, since these would help indicate how much of the V* structure is assigned to the stratosphere in each algorithm. It is difficult to determine this by looking only at the highly structured daily TR maps for the respective algorithms (see major comment 1)
Reply: Following the reviewer's suggestion, we have added maps of stratospheric NO2 from RSM, OMI NASA, OMI DOMINO, and EMAC to Figure 3 for the two selected days.

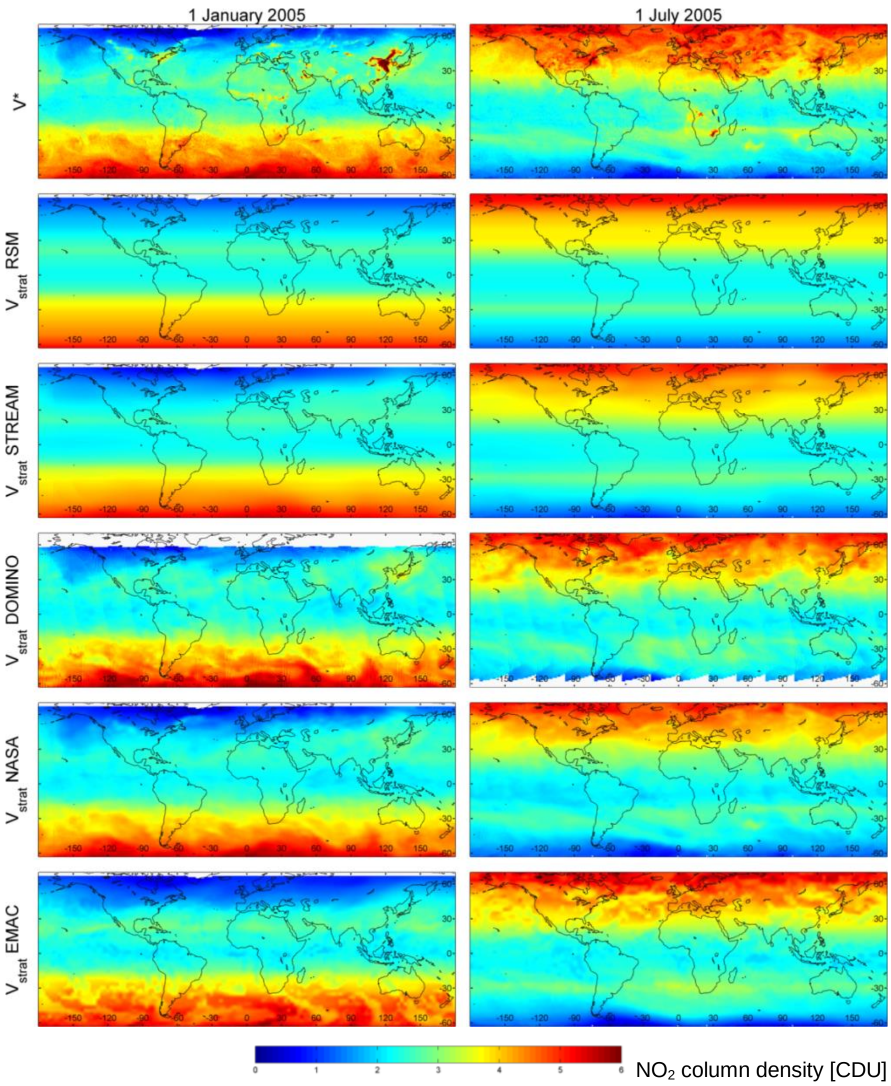

Fig. 3: Total OMI VCD V (top) and the resulting stratospheric estimate V_{strat} from RSM and STREAM for 1 January (left) and 1 July (right) 2005. Stratospheric estimates from other algorithms are included as well for comparison (see section 5.1).*

6. (p 16) In the discussion of cloud weight (4.2.1, part (b)), why not use a larger wcld, which lowers Vstrat and increases the magnitude and standard deviation of T* – i.e. why does the algorithm have a lower wcld as its baseline?

Reply: As explained in the reply to the major comment 1, a higher w_{cld} would indeed be preferable for OMI. However, this would cause artefacts for SCIAMACHY and particularly for GOME due to the different dependencies of total columns on cloud properties.

We have added a discussion on this to the end of section 4.2.1:

"Following the argument that cloudy observations provide a direct measurement of the stratospheric column, a higher cloud weight would be expected to be more favorable, and to result in higher tropospheric background over the Pacific. This is indeed observed for OMI. For other satellite instruments, however, results are somewhat contradictory (see sections 5.3, 5.4 and 6). Thus, the definition of w_{cld} in Eq.6 is kept as compromise in order to have common algorithm settings across different satellite platforms."

7. (p 19) I suggest a wording change at the bottom of p19 (section 5.1.2): "...interprets the difference between the full total column and the (small) tropospheric model as stratospheric column..."

Reply: We have seized the reviewer's suggestion and slightly changed the sentence to *"...interprets the difference between the total column and the (small) modelled tropospheric column as stratospheric column..."*

8. (p 20) Regarding section 5.1.2, bullet point 2, the way the stratosphere is estimated need not affect NO_2 retrievals by cloud slicing (see Choi et al., ACP 2014).
Reply: Choi et al. (2014) do not use the operational tropospheric product, but instead calculate dedicated clouded AMFs and derive a stratospheric estimate by themselves.
But if any user would investigate the dependency of the official tropospheric NO_2 column on cloud height, results would probably be quite different.

9. (p 20) In section 5.1.3, please state why an additive rather than multiplicative offset is applied to EMAC?
Reply: We have applied an additive offset correction as simplest procedure. If instead a multiplicative adjustment is performed, results are hardly affected. We note this in the revised manuscript.

10. (p 22) Section 5.2, first sentence at the top of the page, a misspelling: "...extraordinarily high..."
Corrected.

11. (p 23) Section 5.3, minor wording changes: "...overall still works well...", "...similar to GOME-2...", "...similar to OMI or GOME-2..."
Reply: We have revised the manuscript accordingly.

12. (p 26) Section 5.7 states that STS errors cannot be directly quantified since the "true" stratosphere is not known. As the authors have already shown, this can be known to some extent using synthetic data (major comment 1). Please include a comment here to this effect.
Reply: We agree and have extended the section to
"*The uncertainty of STS can often not be directly quantified, as the "true" stratospheric 4D concentration fields are not known. One approach to assess the STS performance is the usage of synthetic data, as in section 4.3.*
~~But still~~In addition, the TR can be used to evaluate the plausibility of the stratospheric estimate and to derive realistic uncertainties:"
In addition, we refer to the results of the synthetic study when discussing standard deviations and typical differences between different algorithms:
"Overall, the standard deviation of TR from different STS is low (typically <0.1 CDU, and below <0.2 CDU for most parts of the world). It is thus consistent with the uncertainty estimates of stratospheric columns given in literature (Boersma et al. (2011): 0.15-0.25 CDU (SCD); Valks et al. (2011): 0.15-0.3 CDU (VCD); Bucsela et al. (2013): 0.2 CDU (VCD)) and with the magnitude of systematic deviations found in the study on synthetic data (section 4.3)"

13. (p 26) Section 5.7, last sentence, a misspelling: "...focusing..."
Reply: Corrected.

14. (p 27) Section 6, Conclusions, 2nd paragraph, a misspelling: "...with a high weighting factor."
Reply: Corrected.