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.

• The number 1E14 molec/cm2is mentioned several times as tropospheric background NO2 columns over the remote Pacific Ocean, and cite a publication by Valks et al for reference.
However, it should be noted that other studies (Martin et al., doi:10.1029/2001JD001027, Fig. 8; Hilboll et al., 2013, doi:10.5194/amt-6-565-2013, Fig. 5) derive significantly higher background values over the Pacific. It would be good if the authors could acknowledge that the number they use is at the lower end of a range of values proposed by previous studies.
Reply: We agree that the discussion of the tropospheric background was too simplified in the original manuscript. Thus we have revised the manuscript accordingly:
a) in the introduction, we have modified line 22 to
"Neglecting the tropospheric background results in tropospheric columns that are biased low by some 10^{14} *molec* cm^{-2} *(Martin et al., 2002; Valks et al., 2011; Hilboll et al., 2013)"*
b) in section 5.6, we have extended the discussion respectively:
"As (M)RSMs usually estimate the stratospheric column based on total column measurements over clean regions, they generally miss the (small) tropospheric background of the order of some 0.1 CDU. Several (M)RSMs explicitly correct for this effect based on a-priori tropospheric background columns (Martin et al., 2002; Valks et al., 2011; Bucsela et al., 2013). In case of STREAM, however, cloudy pixels, which allow a direct measurement of the actual stratospheric column (except for the small tropospheric column above the cloud), are emphasized. Thus, an additional tropospheric background correction ~~is not required~~ should be unnecessary.
Accordingly, in case of OMI, TR from STREAM are about 0.1 CDU over clean regions, similar as for TR from DOMINO and NASA. This is close to the a-priori value chosen by Valks et al. (2011), but below the values given in Martin et al. (2002) (about 0.15-0.3 CDU, assuming a tropospheric AMF of 2) and Hilboll et al. (2013) (0.1 up to >0.6 CDU; note, however, that the high values are only reported at higher latitudes in winter, when the ratio AMF_{strat}/AMF_{trop} *is almost 2 (Fig. S1); thus the large discrepancy is at least partly resolved if the TR is transferred in a TVCD via equation 4.)"*

• p. 7, l. 4-5: The mean spatial distribution does not reflect the pollution probability, as claimed by the authors. E.g., the same mean value can be caused by a single extreme pollution event in an otherwise clean region, or by moderate, constant in time, pollution levels. So the notion of probability should not be used in this context.
Reply: We agree that the term "probability" is misleading here.
We have revised the respective sentence to
"We thus define a pollution weight w_{pol} *based on our a-priori knowledge about the mean spatial distribution of tropospheric* NO_2*, reflecting ~~the tropospheric pollution probability~~ potentially polluted regions."*

• p. 7, l. 5-6: The authors should specify if the multi-annual mean trop. NO2 column the use does include the seasonal cycle, i.e., if they have one "multi-annual mean" per month. If not, the authors should clarify how the seasonal cycle is being considered in the weight calculation.
Reply: The basis of the calculation of P is the mean NO_2 column density from SCIAMACHY. So far, no seasonality is accounted for.

This might be modified in a future implementation.
However, given the weak dependency of the results on the definition of w_{pol} (see reply to major comment #3 of reviewer #1), the impact of a seasonally varying pollution weight is expected to be negligible.
We have added this discussion to section S2.2.1 of the Supplement:
"So far, no seasonality of NO_2 is considered in the definition of P. This could be added to a future version, but the impact on STREAM is expected to be low (compare section 4.2.6)."

• p. 7, l. 10: It would help the reader if the authors could give a range for the pollution proxy P. Otherwise, it is impossible to grasp how large wpol is in comparison to the other weights.
Reply: The pollution proxy P is displayed in Fig. S2d. Values are about 1 CDU for most remote continental regions up to >3 CDU for central Europe and the US Eastcoast and >6 for China. According to Eq. 5, the resulting pollution weights for these numbers are 0.1, 0.0037, and 0.0005 (compare Fig. 1a).
In the revised manuscript, we have added
"Details on the definition of P are given in the supplement (sect. S2.2.1), and P is displayed in Fig. S2d."
to the end of page 7, line 9.

• p. 8, l. 12: The authors should explain why measurements where the strat. contribution has been overestimated should contribute more strongly to the strat. estimate.
Reply: In cases of negative TR, the estimated stratospheric column is larger than the total column measurement itself. This overestimation of the stratospheric column is caused by neighboring observations via weighted convolution. By increasing the weight of the local measurement, its impact on the stratospheric estimate is enlarged, resulting in a lower stratospheric estimate, and a higher (ideally>0) TR.
We have clarified this in the revised manuscript:
"As negative columns are nonphysical, $T^*<0$ indicates that the stratospheric field has been overestimated. This happens if the weighted convolution with neighboring pixels with high total columns causes the estimated stratosphere to be even higher than the local total columns. In order to avoid this effect, the respective local total columns should be assigned a higher weighting factor such that they contribute more strongly to the stratospheric estimate."

• p. 9, l. 2: The reference to "S4.2.3" is wrong.
Reply: We have corrected this to "*S4.2.5*".

• p. 9, l. 3-6: The example of pixels over U.S., Europe, central Africa, and China leading to low wTR is not helping, since without further information, the reader has to assume that these regions already have low weight due to wpol. As the differentiation between the pollution and the trop. res. weights is not immediately clear to the reader anyways, it might be a good idea to find an example of unusually high polluted regions, which would not have been assigned low weights by using wpol alone.
Reply: The listed regions indeed already have a low w_{pol}. However, the additional use of w_{TR} further lowers the net weight by orders of magnitude. While in some regions with $w_{pol}<1$, w_{cld} lifts the net weight >1 (e.g. over Northern Scotland), w_{TR} keeps the net weight low for

actually polluted pixels. In particular over Eastern China, the measurements with high TVCDs are essentially ignored during weighted convolution (even in case of clouds).
Examples for low w_{TR} outside the potentially polluted regions (w_{pol}=0) cannot be found in Fig. 2(c), as they are excluded intentionally. This was a late modification of the algorithm which was accidentally not appropriately updated in the manuscript. So we would like to thank the reviewer for digging deeper in this respect.
The motivation for applying w_{TR} only for pixels in potentially polluted regions was the finding that stratospheric structures regularly result in high TR in remote regions, see e.g. the stripe of enhanced TR in the Indian ocean on 1st of July in Fig. 5. If w_{TR} would be applied to this stripe, the respective measurements would be weighted even further down, though they actually represent the stratosphere, and the systematic bias would even be amplified.
In the revised manuscript, we have clarified the procedure:
Page 8 line 6:
"*A high value of $T*$ ~~generally~~ likely indicates tropospheric pollution, in particular over potentially polluted regions.*"
Page 9 line 3:
"*4. A $w_{TR}<1$, which is meant to decrease the weight of polluted pixels, is only applied over potentially polluted regions with $w_{pol}<1$. This restriction was introduced to avoid the amplification of stratospheric patterns which are still present in the TR, like the stripe of enhanced TR in the Indian ocean on 1st of July in Fig. 5.*"
Page 9 line 5:
"*Observations over these regions are already associated with a low pollution weight. However, due to the additional application of w_{TR}, the net weight is lowered further by orders of magnitude, and the respective satellite pixels will hardly contribute to the stratospheric estimate in the next iteration, even in case of high w_{cld}.*"

• p. 10, l. 1: If the authors set Wij = 0 in case of measurement gaps, then Vij as defined by the authors is not defined. Shouldn't it be enough to set Cij= 0? Otherwise, please amend Eq. 11 so that it yields a well-defined Vijeverywhere.
Reply: In case of measurement gaps, Vij (the mean VCD of all available obervations) is just not defined. In the revised manuscript, we have added a respective note directly after equation 11.
For the procedure of weighted convolution (equations 12 and 13), however, it is essential to set both Cij and Wij to zero in empty grid boxes.

• p. 10, l. 4: The authors should clarify if the 2D Gaussian they use as CK is defined in degree-space or in kilometer-space. If it is defined in degree-space, they should justify the resulting inconsistency depending on latitude.
Reply: The first part of the section is meant to describe the general procedure of weighted convolution, with a yet unspecified CK, while the details of the choice of CK are given later. To make this clear, we have replaced *"(~~e.g. a 2D Gaussian~~)"* by *"(see below)"*.
In the definition of G (eq. 15), it is evident that a lat/lon grid is chosen. We have added *"Note that the difference of the CKs, which are defined on a regular degree grid, is even more drastic in kilometer space."* to equation 15.

• p. 13, l. 31: There's a spurious "see" in the refernce to Jöckel et al.
Reply: Corrected.

• p. 14, l. 6: The authors should specify how the EMAC model determines the tropopause height, i.e., thermal, dynamical, ... criterion?

Reply: We have revised the sentence to
"Stratospheric VCDs were calculated by vertical integration of the modelled NO_2 mixing ratios between the tropopause height (as diagnosed according to the WMO definition based on lapse rate equatorwards of 30° North/South, and as iso-surface of 3.5 PVU potential vorticity poleward of 30° latitude) and the top of the atmosphere."

• p. 17, l. 5: Currently, the manuscript states that "the final Vstrat [...] as weighted mean of both [CKs]". This is not really precise, it should rather say that the final Vstrat is the weighted mean of Vstrat calculated with both CKs.
Reply: We have modified this sentence to
"In STREAM, two different CK are applied, yielding two stratospheric estimates, and the final V_{strat} is calculated as weighted mean of both (see section 2.3 and equations 15 and 16)."

• p. 18, l. 6: The authors write of "the" small-scale structures of strat. NO2 in EMAC. It would be good to elaborate a bit on the "the", i.e., in which regions, in which months, ...
I suspect that these structures are mostly there at low latitudes, but it would be good if the authors could be more explicit about this.
Reply: To illustrate this, we have added a figure showing the synthetic stratospheric and tropospheric NO2 fields to the supplement:

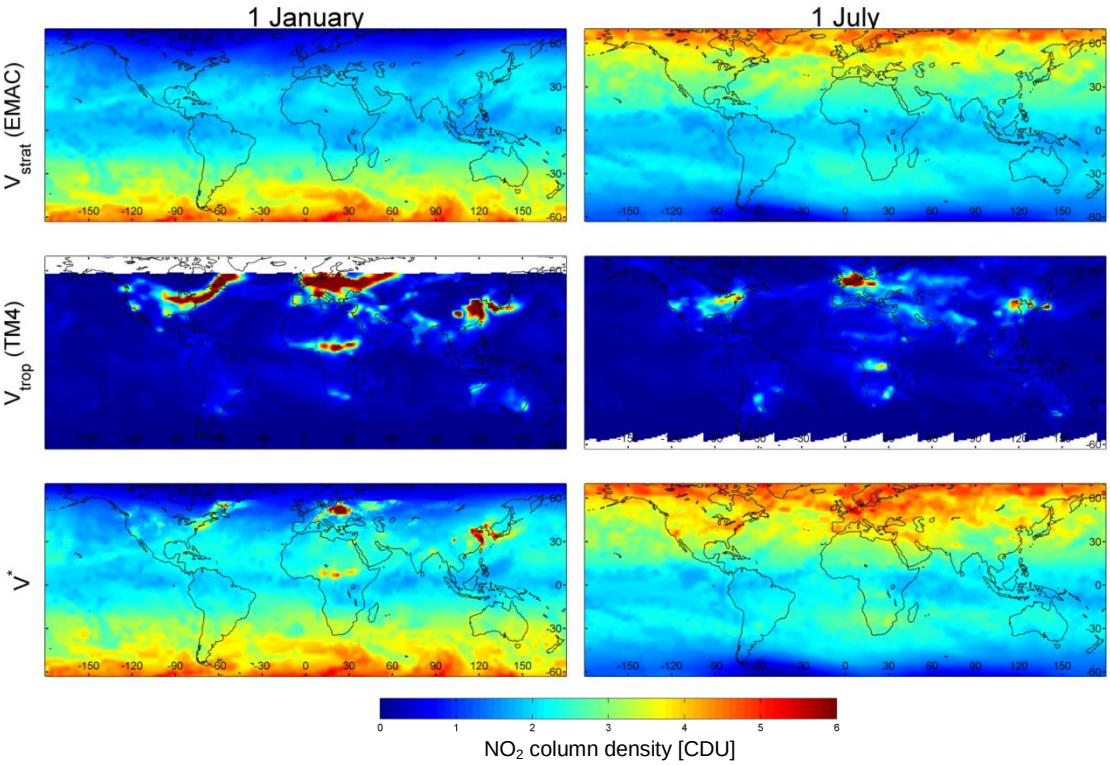

Figure C : Synthetic NO_2 column densities of stratosphere (EMAC, top), troposphere (TM4, center) and total V (bottom) for 1 January (left) and 1 July (right) 2005.*

We have extended the discussion accordingly:
"This is mainly caused by small-scale structures of stratospheric NO_2 in EMAC over the Pacific, in particular at southern latitudes, which are resolved by neither STREAM nor RSM (see Figure C)."

• p. 18, l. 10-11: The authors should clarify if the "remaining biases" are low or

high biases.
Reply: We have specified this sentence to
"Remaining systematic biases are about -0.1 CDU over polluted regions, i.e. resulting TRs are slightly underestimated, as expected due to the general approach of using total column measurements as proxy for the stratospheric estimation."

• p. 18, l. 26: "meaningful" has an extra "l"
Reply: Corrected.

• p. 19, l. 11-12: The authors should comment on whether they expect higher variability of T∗in DOMINO or STREAM.
Reply: STREAM uses convolution to derive stratospheric fields, with large-scale convolution kernels at low latitudes. Thus, any variability of stratospheric NO2 on smaller spatial scales will result in increased variability of T*, even if the mean is appropriately estimated. DOMINO, on the other hand, can generally resolve gradients on smaller scales. However, daily maps of TR from DOMINO (Fig. S19) reveal patchy residuals over the ocean. This might be related to a misrepresentation of stratospheric gradients in the model, but this is not our expertise and not the focus of the manuscript.

• p. 19, l. 17: underestimation "by", not "of" DOMINO
Reply: Corrected.

• p. 21, l. 9-11: Might the longitudinal dependency in STSEMAC partly be caused by the temporal sampling of the EMAC model? The LT difference between East and West end of the OMI swatch is rather large, and using a fixed EMAC time might introduce longitudinally varying biases
Reply: The effect of different local time at the swath edges is generally small, as shown in detail within the reply to the major comment #2 of reviewer 1.
What is meant here is the large scale longitudinal pattern of stratospheric NO2: over the Pacific, STS_{EMAC} is by construction matching to OMI. But the longitudinal dependency at northern latitudes is quite different for STREAM and EMAC (compare the updated Fig. 3, which now includes EMAC stratosphere as well as proposed by reviewer 1). Consequently, STREAM results in low background TR over Siberia (Fig. 5), while STS_{EMAC} underestimates the stratospheric column over Siberia and results in high biased TR (similar as RSM).
We have revised the manuscript accordingly:
"*The systematic deviations North from 35°N (1.&2.) are caused by the longitudinal dependency of stratospheric NO_2 from EMAC which differs from the pattern in total column (see Fig. ~~S23~~3). In detail, stratospheric NO_2 over Siberia is quite low in EMAC, resulting in high biased TR (similar as for RSM) and indicating that the mean longitudinal dependency of stratospheric NO_2 is not fully reproduced by EMAC.*"

• p. 21, l. 28-29: GOME/SCIAMACHY/GOME-2 might have systematic differences in the CTP products, which might in turn influence the STS results. Furthermore, the spatial resolution of the different instruments should lead to different pdfs for the cloud fraction, again possibly influencing the STS results. It would be good if the authors could comment on this issue.
Reply: We have modified the sentence to
"This might be related to the differences of cloud statistics due to pixel size, in particular a lower number of fully clouded pixels for GOME-2, as well as differences in local time, cloud products, or systematic spectral interferences caused by clouds in either retrieval algorithm."

• p. 22, l. 24: "MOZART" should be replaced with "MOZART-2"
Reply: Corrected.

• p. 24, l. 20: Also, the smaller pixel size will lead to a higher fraction of purely clouded pixels. The authors should briefly comment on the implications.
Reply: This is already stated in line 20 of the manuscript, with the implication that more sampling points will become available over potentially polluted regions (line 21). See also the reply to the comment above, which already picks up this aspect in section 5.2.

• Sect. 5.5.2: Sentinel-4 is the name of the satellite, not the instrument. The authors should instead write something like "UVN onboard Sentinel-4" or the like.
Reply: Sentinel-4 is the name of the mission, which will include a UVN instrument to be launched on a MTG-S (Meteosat Third Generation Sounder) satellite.
We have tried to be more precise on this without bothering the reader with details irrelevant for this study by changing lines 25-26 to
"Over Europe, Sentinel 4 (S4, Ingmann et al., 2012) will be the first ~~geostationary~~ mission providing a spectral resolving UV/vis instrument on a geostationary satellite."
For simplicity, we still refer to "S4 measurements" in the remaining section.

• p. 26, l. 8: Please add "on average" before "negative T∗, because single negative T∗are expected, as the authors already noticed elsewhere.
Reply: Done.

• p. 26, l. 31-32: The authors should note that applying STREAM to other trace gases, where the bulk of the profile cannot be expected to lie within the boundary layer but rather in the same altitude ranges as clouds is at least challenging.
Reply: The intended message here was that the concept of weighted convolution might be useful for background estimates in other contexts as well. Of course, the reviewer is right in pointing out that the application of w_{cld} is not appropriate if the background in question is of tropospheric nature, or any kind of algorithm bias.
We have thus revised the last sentence to
"Thus, the concept of weighted convolution could be used within the satellite retrievals of e.g. SO2, BrO, HCHO, or CHOCHO, with appropriately chosen and optimized weighting factors."

• p. 27, l. 21-24: The authors should think about ways in which the LT of measurement being in the morning could hinder STREAM for GOME/SCIAMACHY/GOME-2.
Reply: As demonstrated, the algorithm generally successfully worked for GOME1/2 and SCIAMACHY, except for the tropospheric background, which turned out to be related to different dependency of total columns on cloud conditions as compared to OMI.
We thought and discussed among the co-authors about the possible impact of LT on the observed differences in the response to cloudy pixels. However, we could not think of any mechanism how the differences in LT could explain our findings. Still, we have modified the respective sentence to
"*The emphasis of clouded observations, which provide a direct measurement of the stratospheric rather than the total column, should supersede an additional correction for the tropospheric background, which successfully worked for OMI, but less for GOME and SCIAMACHY. This might be related to differences in pixel size or local overpass time, both*

potentially affecting cloud statistics, or differences in the cloud algorithms. However, the detailed reasons are not yet fully understood and require further investigations."

• Fig. 8: I cannot understand how the 90% percentile T∗over polluted regions can still be lower than 1 CDU. Is this really correct? Furthermore, I have trouble deriving the minimum in the T∗difference over China/Japan in Fig. 9 from the statistics given in this Fig. 8.
After seeing the definition of "polluted" in Fig. S7, this becomes clear; but it points to the misleading label "polluted" in this context.
Reply: We agree that the term "*polluted*" is misleading and changed it to *"potentially polluted"* in all respective diagrams.

• Figs 9, 10, 11, S19, S20, S23, S24, S26, S27b, S30-S31, S33 should have ΔT∗ as colorbar label instead of T∗.
Reply: Figures 9-11 display differences of T* from different algorithms, and thus the reviewer is right. We have corrected the labels accordingly.
The listed figures in the supplement, however, actually display T*.

• Fig. S9: Isn't increasing the cloud weight by a factor of 10 equivalent to increasing the total weight by a factor of 10, due to the definition of the total weight in Eq. 8? Why should this Figure then say something about the cloud weight in particular?
Reply: The reviewer is absolutely right: a simple factor of 10 would just raise the total weight everywhere. The description is not exact.
What we have done in the "high cloud" scenario is switching Equation 6(a) from $10^{(2 w_c w_p)}$ to $10^{(3 w_c w_p)}$. I.e., in case of full cloud cover around 500 hPa ($w_c w_p=1$), w_{cld} is indeed raised by a factor of 10, but for low cloud fraction or for high/low clouds ($w_c w_p=0$), w_{cld} is still 1. We have clarified this in the revised manuscript in section 4.2.1 (b).

• The authors should introduce the OMI instrument before making reference to it from p. 8 onwards. In the current manuscript, OMI is only introduced later, in Sect. 3.1.
Reply: In the revised manuscript, we now introduce SCIAMACHY, GOME-2, and OMI in the first sentence of the introduction:
"Beginning with the launch of the Global Ozone Monitoring Experiment (GOME) on the ERS-2 satellite in 1995 (Burrows et al., 1999), several instruments (SCIAMACHY, OMI, GOME-2; see Table 1 for acronyms and references) perform spectrally resolved measurements of sunlight reflected by the Earth's surface and atmosphere."