Referee (#1)

The authors would like to thank Referee #1 for his/her thoughtful and helpful comments and suggestions. Below are the comments by Referee #1 in blue and answers in black. Any modification made to the text has been highlighted within a green box. The line numbers correspond to the version of the manuscript available for online discussion.

Line 132: "It is desirable to use as much information as possible retrieved from the satellite instrument itself." I didn't see much in the model descriptions with regards to this statement. What information from the satellite is used, is it the same for all the models, etc.

Answer: In the AMF calculation, usually as much information as possible is used from the satellite itself (in this case measurements from OMI). These are the viewing geometry, surface albedo (from the OMI LER climatology), cloud fraction and cloud pressure (from the OMI O2-O2 algorithm). The external parameters are terrain height (from a database) and a priori profile shape (from chemistry transport model simulations). This is the case for most of the retrievals that participated in the comparison. The sources for the different parameters in each of the retrievals are summarized in Table 3.

We have slightly modified the sentence:
"It is desirable to use as many forward model parameters as possible retrieved from the satellite instrument itself."

Line 275: "The agreement in this study is better than previous RTM comparisons:" Any idea why?

Answer: In this study we make a very detailed comparison of TOA reflectances in order to have a quantitative and specific number for the agreement between the RTMs. The comparison by Stammes (2001) only included two RTMs (DAK and MODTRAN), and Wagner et al. (2007) did not go as much in detail as we do here to get a TOA reflectance structural uncertainty value. Both studies give a more general value for TOA reflectance differences "up to or within 5%". Another general reason might be that RTMs have improved over the last 10-15 years.

Line 357 (or 337): The differences between online and LUT AMFs would depend on the resolution of the LUT. Since the Castellanos et al (2015) study indicated an 8% error (which is larger than the error due to RTM treatments you found), it would be worthwhile to have a brief discussion on the LUTs used in the different models and possible differences that may arise.

Answer: In Sect. 3.3.1 we obtain differences in AMF (using the same a priori information) of 6.5% for polluted areas and 2.5% for unpolluted areas. These numbers represent the differences introduced in the AMF calculation by (1) the use of different RTMs and (2) interpolation errors due to the use of LUTs. Other studies like Castellanos et al. (2015) found an average difference of 1% (and less than 8% for individual measurements) between interpolated LUT and online AMFs. Lin et al. (2014) found 1 to 5% differences in VCDs retrieved with interpolated LUTs and with online radiative transfer over China. These numbers are of the same magnitude as the

differences we found in our comparison. Furthermore, in our study interpolation errors are likely much less of a concern than in Castellanos et al. (2015) and Lin et al. (2014) because of the coarse grid used in their LUTs (this of course does not make those errors go away e.g. in DOMINO-2).

Based on this we conclude that the differences that may arise from using different LUTs by the different retrievals are not larger than 6.5%. Following the suggestion from the reviewer we add some discussion on this topic:

L376: "6.5% represents an upper limit value for the differences that using different RMTs and LUTs may introduce in the final AMF calculation."
L436: "Different groups use different LUTs for their AMF calculations, and POMINO uses pixel-by-pixel online radiative transfer calculations. The LUTs are different in several aspects: the RTMs used to create them and the number of reference points for each dimension. All these differences affect the AMF structural uncertainty. Based on the discussion in previous sections we consider that the use of different LUTs introduces a structural uncertainty of the order of 6.5%."

Section 3.3.2: Which model is used to evaluate the cloud corrections?

Answer: The AMFs used to evaluate the cloud corrections in Sect. 3.3.2 are those from KNMI/WUR. We add a sentence to clarify this on the text (**line 390**):

''To quantify the differences between the two approaches, we compare here tropospheric NO_2 AMFs calculated by WUR (see Table 3) with the IPA and CM approach for two complete days of OMI measurements (02 February 2005 and 16 August 2005)''

Line 467-470: Characterization of the sensitivity of AMF to albedo and to a priori profiles is inconsistent. Surface albedo is said to "explain some" of the difference, while the AMFs are "highly sensitive" to the a priori, even though their correlations in Table S3 are very similar (0.21 & 0.50 compared to 0.19 & 0.55). Also, line 492 calls the a priori profiles the "main cause" of the differences. From the information given, AMFs seem equally sensitive to both a priori and albedo, however the text suggests otherwise and should be rephrased.

Answer: Our argumentation is based on the fact that in our ensemble, the number of pixels where albedo differences (ΔA_s) co-vary with AMF differences (ΔAMF) is considerably smaller than the number of pixels where modelled NO_2 vertical column differences (ΔNO_2) co-vary with AMF differences (5382 vs. 15142 in 16 August and 1876 vs. 6483 in 02 February).

However, we agree with the reviewer that the albedo differences are important for the AMF differences and that the original text might suggest that the influence of surface albedo is not really important. Therefore, we have rephrased the sentences in lines 467-473 to make it clearer.
"…and that surface albedo differences explain WUR and BIRA AMF differences especially in winter, when NO_2 is found close to the surface…"
"In our ensemble, the WUR-BIRA AMF differences are highly sensitive to the differences between the a priori NO_2 profiles used, especially in summer."

Line 492: "In the previous section, we found that differences between a priori NO_2 profiles and the surface albedo are the main cause for AMF structural uncertainty when cloud parameters are identical."

Line 490: It's not clear how the use of an averaging kernel will reduce the effect of the a priori. Averaging kernels are most frequently used for making comparisons between two models, or between a model and a retrieved observation, in order to reduce errors that may arise when two models are based on different a priori profiles. From my understanding this does not reduce the retrieval's sensitivity to the a priori itself. This statement needs further explanation.

When averaging kernels are being applied, for instance when comparing retrieved NO_2 columns with modelled NO_2 distributions or with observed NO_2 profiles (aircraft, balloon), the comparison will become self-consistent in terms of using a priori information. Using the averaging kernel reduces systematic and random differences between modelled and satellite-observed columns because the representativeness of the modelled state for the observed state improves (e.g. Boersma et al., 2016). We agree that the retrieval of NO_2 columns will stay sensitive to the choice of the a priori profile, but using the averaging kernel provides a data user with the means to improve the consistency associated with the a priori profiles in interpreting the satellite data.

As both reviewers have raised their concern in this particular statement, we have tried to make it clearer:

It is worth to note that using averaging kernels in satellite applications (e.g. when comparing retrieved NO_2 columns with modelled NO_2 distributions or observed NO_2 profiles) will reduce the representativeness errors in the comparisons associated with the a priori trace gas profile used in the retrieval scheme (e.g. Boersma et al., 2016).

Technical notes:

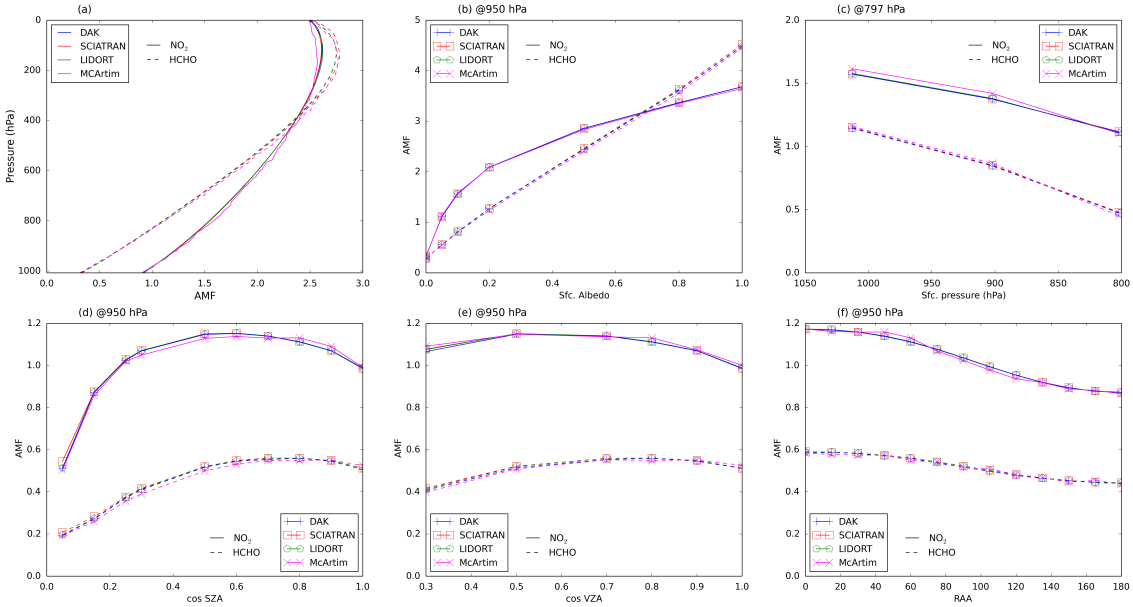
Line 43: "20-50% from typical VCDs uncertainties of 40-60%" is ambiguous. Is it 20-50% of the typical uncertainty, or is it 20-50 percentage points. Consider rewording.

Answer: We have now rephrased it:

Previous studies indicated that AMF calculation is the largest source of uncertainty (contributing up to half of the typical VCD uncertainties of 40-60%.) in the NO_2 and HCHO retrievals…"

Figure 4: The green and red lines are hard to distinguish from the others in parts b-f. I realize that this is because they are overlapping, but (for example) in 4b are the green and red lines under the blue one, or under the pink one? It would be good to find a way to make this clearer.

Answer: We have changed the figure, particularly we have now applied a different marker for each RTM. For example in Fig. 4(b) even if the pink line is still in front, by looking at the markers we can see that green and red lines are under the blue line and not under the pink line.



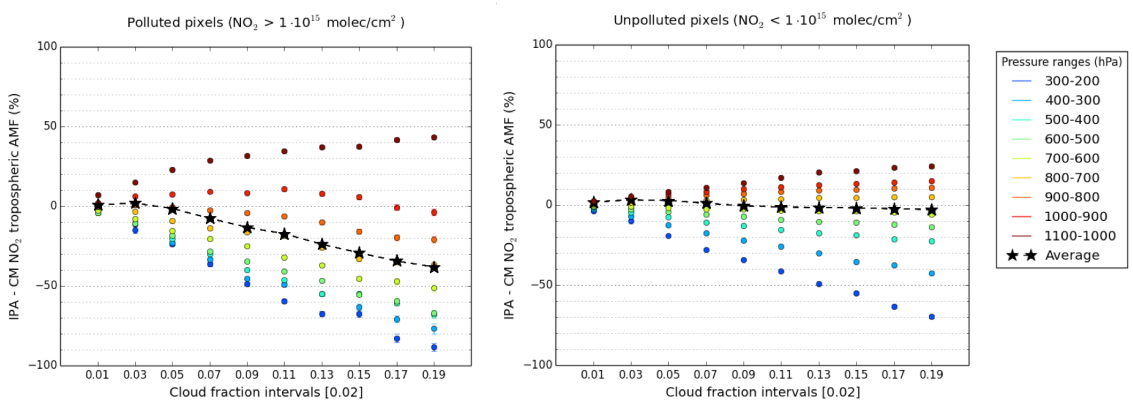
Line 426: "when cloud fraction is less than 0.1…" Should this be "greater than 0.1"? "Less than" seems to contradict the discussion in Section 3.3.2.

Answer: Yes, indeed. IUP-UB applied IPA cloud correction when cloud fraction is greater than 0.1. We have modified the text:

"IUP-UB and BIRA now apply IPA only when cloud fraction exceeds 0.1 and 0.2 respectively"

Lines 390-405/Figure 7: The text discusses relative differences in AMFs, and mentions that differences are small in unpolluted situations with larger differences in polluted areas. Figure 7 shows absolute differences in AMF, and the polluted and unpolluted plots have a similar vertical range. I would suggest including a plot of the relative differences to better illustrate the conclusions made in the text.

Answer: Thanks for this comment. Following the reviewer's suggestion, we have substituted Fig. 7 with the relative differences in the y-axis (see figure below).



References

Boersma, K. F., Vinken, G. C. M., and Eskes, H. J.: Representativeness errors in comparing chemistry transport and chemistry climate models with satellite UV-Vis tropospheric column retrievals, Geosci. Mod. Dev., 9, 875–898, doi:10.5194/gmd-9-875-2016, 2016.

Castellanos, P., Boersma, K. F., Torres, O., and de Haan, J. F.: OMI tropospheric NO_2 air mass factors over South America: effects of biomass burning aerosols, Atmos. Meas. Tech., 8, 3831–3849, doi:10.5194/amt-8-3831-2015, 2015.

Lin, J.-T., Martin, R. V., Boersma, K. F., Sneep, M., Stammes, P., Spurr, R., Wang, P., Van Roozendael, M., Clémer, K., and Irie, H.: Retrieving tropospheric nitrogen dioxide from the Ozone Monitoring Instrument: effects of aerosols, surface reflectance anisotropy, and 725 vertical profile of nitrogen dioxide, Atmos. Chem. Phys., 14, 1441–1461, doi:10.5194/acp-14-1441-2014, 2014.

Stammes, P.: Spectral radiance modeling in the UV-Visible range, IRS2000: in: Current problems in atmospheric radiation, edited by: Smith, W. L. and Timofeyev, Y. J.„ pp. 385–388, 2001.

Wagner, T., Burrows, J. P., Deutschmann, T., Dix, B., von Friedeburg, C., Frieß, U., Hendrick, F., Heue, K.-P., Irie, H., Iwabuchi, H., Kanaya, Y., Keller, J., McLinden, C. A., Oetjen, H., Palazzi, E., Petritoli, A., Platt, U., Postylyakov, O., Pukite, J., Richter, A., van Roozendael, M., Rozanov, A., Rozanov, V., Sinreich, R., Sanghavi, S., and Wittrock, F.: Comparison of box-air-mass-factors and radiances for Multiple-Axis Differential Optical Absorption Spectroscopy (MAX-DOAS) geometries calculated from different UV/visible radiative transfer models, Atmos. Chem. Phys., 7, 1809–1833, doi:10.5194/acp-7-1809-2007, 2007.