The replies to the reviewer comments are marked in blue

Wagner et al., 2018 address a very important topic of the need of scaling factor to bring MAX-DOAS measured differential slant column densities (dSCD) of oxygen collision complex (O4) retrieved from 352 – 387 nm in agreement with the radiative transfer modeled dSCD at 360 nm. An extensive and very thorough evaluation of the error sources in the DOAS analysis and RT modeling is presented. The authors analyzed data from two time periods (18 June and 8 July 2013) during MADCAT campaign in Mainz, Germany, when time and location coincident MAX-DOAS, aerosol (AERONET, Ceilometer) profile measurements were conducted with a support of additional surface observations (PM2.5, PM10, temperature, pressure and relative humidity). They identified "standard" cases for DOAS fitting and for RT model simulations, and a number of potential scenarios deviating from the standard cases. The authors concluded that the agreement between the measured and modeled O4 dAMF is almost perfect 1.01 (±0.16) on 18 June 2018. On the other hand the "measured" O4 dAMF had to be scaled by 0.71 (±0.12) to bring in agreement with the modeled absorption for standard case DOAS fitting and RT modeling scenarios. The cause of the discrepancy was not identified.
This work is very important and is well suited for AMT publication. However, I think the article will benefit from some reorganization.

We thank the reviewer for the positive assessment of our paper and for the good suggestions. We addressed most of them as described in detail below.

Major comments:

I think that there are two main topics that the authors are trying to address (I would say each of them is worth a separate publication):
(1) Is a scaling factor required to obtain closure between measured and modeled atmospheric O4 absorptions – Part A: identifying best-case scenarios based on auxiliary measurements and best practices.
In this part the best case DOAS fitting scenario and best case RT modeling scenario should be identified based on the best available data to describe atmospheric conditions during the selected periods. Potential sources of errors for ***these particular*** scenarios should be evaluated. For example, for RT modeling:
- Mie scattering phase functions using AERONET inversion data results for size distribution and refractive index real and imaginary parts extrapolated to 360 nm from longer wavelengths (440, 675 nm). Evaluating errors associated with these particular inputs to the RT (e.g using 440 nm inversion results directly?). Please also note that the AERONET level 2.0 inversions are not available during some of the selected periods, potentially due to presence of clouds. Available dates/time are listed below:
6/18/13 07:24:51
6/18/13 15:34:32
6/18/13 16:12:07
7/8/13 05:16:20
7/8/13 05:48:33
7/8/13 06:54:34
7/8/13 07:32:12
7/8/13 15:38:04
7/8/13 16:12:13

7/8/13 17:18:13
7/8/13 17:50:24

- Ceilometer backscatter profiles corrected by AERONET CIMEL AOD, and their errors (backscatter to aerosol extinction coefficient profiles conversion, wavelength differences, extrapolation to the surface)
- Radiosonde temperature, pressure and relative humidity measured profiles at fine grid with ECMWF ERA-Interim reanalysis above and their errors (e.g. different groups extraction of the data, usage of MERRA-2 profiles available at better than 1 km resolution near ground and every 3 hours)
- Accounting for polarization and RRS in the RT calculations and their errors (e.g. different models)
- If we consider O4 cross section by Thalman and Volkamer (2013) accurate at all temperatures use T-dependent O4 cross sections for RT calculations.
- Surface albedo from satellite measurement or AERONET inversion at 440 nm (which varies from 2.7 to 4% during the selected times).
- Effect of instrument FOV and pointing error, especially under shallow aerosol layer presence (the fact that measured dSCD at several low VEA are close to each other does not exclude potential error in pointing that has to be accounted for in modeling). DOAS fitting scenario selected for the standard case can be considered best practice. The only things I would probably recommend changing is the offset from polynomial order 2 to 1 and not applying polynomial at all to the O4 cross section due to its broad band wavelength dependency. In calculating the errors due to the fitting, I would not go to the extreme case of no offset. At low elevation angles the effective O4 temperature is around 270K, I would suggest using O4 cross section at 273K as one of the sensitivity cases.

There is another change I would recommend here – what quantity is actually compared. Since the actual measurements are ground-based hyperspectral sky radiances the derived variable directly from the measurements without any assumptions about the atmosphere (accept for species effective temperatures) is the differential slant column density (dSCD).

There are no passive measurements at the bottom of atmosphere that do not contain O4 absorption, including the reference used in this study (zenith direction). From Beer's law, ignoring wavelength shift, offset and other corrections:

$$\left(\frac{\ln\left(I_{90}^{measured} - I_{VEA}^{measured}\right)}{\sigma_{O4}(T)}\right)_{\lambda\ window} = dSCD_{VEA}^{measured} =$$

$$= \underbrace{SCD_{VEA}^{total} - SCD_{90^o}}_{individual\ components\ are\ not\ measured\ directly}$$

$$dAMF_{VEA} = \frac{dSCD_{VEA}^{measured}}{VCD} = \frac{SCD_{VEA}^{total} - SCD_{90^o}}{VCD} = AMF_{VEA} - AMF_{90^o}$$

From the above discussion AMF and dAMF are quantities derived based on the assumptions made about AMF90 and VCD:

$$AMF_{VEA} = dAMF_{VEA} + AMF_{90^o} = \frac{dSCD_{VEA}^{measured}}{VCD} + AMF_{90^o}$$

I believe the paper will benefit if dSCD are compared directly with the RT modeled dSCD in the first section of the paper.
At the end of this section the reader should clearly see based on the best DOAS fitting and relevant to it errors and best atmosphere modeling (with its relevant errors) whether the measured and modeled dSCDs agree and to what extent.
(2) Is a scaling factor required to obtain closure between measured and modeled atmospheric O4 absorptions – Part B: error analysis to explain potential causes of SF (varying the parameters outside of (1).
This section can include all the other cases for (d)AMF comparisons. Its main purpose could be to make recommendations and identifying problems with using less realistic atmospheric scenarios in the MAX-DOAS data inversions and DOAS fitting limitations.

We thank the reviewer for this suggestion. We understand the intention, but we decided not to split the paper into two parts. The main reason is that both suggested parts are closely linked and it would thus be difficult for the readers to follow them when split into separate papers. In addition, the suggested part 2 would be rather short and mostly speculative, because the reason for a scaling factor is still not known.
Thus we addressed the suggestion of the reviewer by including a new section (section 5.2 'Which conditions would be needed to bring measurements and simulations on 8 July into agreement?'). In that section changes of the measurement conditions are discussed which could bring measurements and simulations into agreement.

The detailed suggestions of the reviewer given above (for part 1) are addressed below:

In this part the best case DOAS fitting scenario and best case RT modeling scenario should be identified based on the best available data to describe atmospheric conditions during the selected periods. Potential sources of errors for ***these particular*** scenarios should be evaluated. For example, for RT modeling:
- Mie scattering phase functions using AERONET inversion data results for size distribution and refractive index real and imaginary parts extrapolated to 360 nm from longer wavelengths (440, 675 nm). Evaluating errors associated with these particular inputs to the RT (e.g using 440 nm inversion results directly?).

In our opinion, we already selected scenarios for the quantitative comparison which are (at least close to) the optimum choice. On both days we selected periods around noon, for which the measured intensities are high and the variation of the SZA is small. Moreover, during the selected periods, the variation of the ceilometer profiles is relatively small compared to before and after. We added this information to section 3.2.

Many thanks for the information about the available phase functions! We performed sensitivity studies to quantify the effect of the extrapolation of the phase functions. We found that the O4 (d)AMFs hardly change (<1%) if either the phase functions at 440 nm or extrapolated to 360 nm are used. Similar small changes are found if the phase functions before or after the selected periods are used.

- Ceilometer backscatter profiles corrected by AERONET CIMEL AOD, and their errors (backscatter to aerosol extinction coefficient profiles conversion, wavelength differences, extrapolation to the surface)

As stated above, the variation of the ceilometer backscatter profiles was relatively small during the selected periods.

- Radiosonde temperature, pressure and relative humidity measured profiles at fine grid with ECMWF ERA-Interim reanalysis above and their errors (e.g. different groups extraction of the data, usage of MERRA-2 profiles available at better than 1 km resolution near ground and every 3 hours)

In principle one could use fine grid ECMWF ERA-Interim reanalysis data, but since the uncertainties related to the temperature and pressure profiles are rather small compared to other uncertainties, we did not use additional meteorological data.

- Accounting for polarization and RRS in the RT calculations and their errors (e.g. different models)

As shown in our study, the effects of polarization in the RT are negligible. RRS was taken into account for the synthetic spectra and almost perfect agreement with the simulated O4 (d)AMFs was found. Thus we conclude that the effects of polarization and RRS can be neglected.

- If we consider O4 cross section by Thalman and Volkamer (2013) accurate at all temperatures use T-dependent O4 cross sections for RT calculations.

This is in principle a good idea. However, the effect is probably very small as indicated by the very good agreement of the results from the synthetic spectra and the simulated O4 (d)AMFs.

- Surface albedo from satellite measurement or AERONET inversion at 440 nm (which varies from 2.7 to 4% during the selected times).

The variation of the surface albedo could also be taken into account, especially if it deviates strongly from the ‘standard settings’. However, as shown in our study, the influence of small changes (e.g. from 5% to 3%) on the O4 (d)AMFs is rather small (below 1%).

- Effect of instrument FOV and pointing error, especially under shallow aerosol layer presence (the fact that measured dSCD at several low VEA are close to each other does not exclude potential error in pointing that has to be accounted for in modeling).

We agree with the reviewer and performed additional sensitivity studies varying the FOV and also systematically distorting the elevation calibration by ±0.5°. The changes of the O4 (d)AMFs were below 1%. We added this information to the text (section 3.2).

DOAS fitting scenario selected for the standard case can be considered best practice. The only things I would probably recommend changing is the offset from polynomial order 2 to 1 and not applying polynomial at all to the O4 cross section due to its broad band wavelength dependency.

In our opinion there might be good reasons for increasing the degree of the fitted intensity offset. For example, the relative contribution of spectral stray light could cause an intensity offset in the measured spectra, which changes non-linearly with wavelength. Thus we think it is difficult to give a clear recommendation on the degree of the intensity offset.
We added the following text in section 4.3.2: 'Higher order intensity offsets might compensate for wavelength dependent offsets (e.g. spectral straylight), which can be important for real measurements, while the synthetic spectra do not contain such contributions.'

Concerning the application of the polynomial, there might be a misunderstanding. We included the O4 cross section without any previous high or low pass filtering. Concerning the degree of the DOAS polynomial we see good reaons to use such a polynomial, e.g. that the broad band wavelength dependence of the measured spectra are different for the different elevation angles, and also change with time. The very good agreement between the results of the synthetic spectra and the simulated O4 (d)AMFs indicates that the chosen polynomial degree is not problematic.

In calculating the errors due to the fitting, I would not go to the extreme case of no offset.

We agree. Note that the case without intensity offset was already ignored for calculating the errors in the discussion version of our paper.

At low elevation angles the effective O4 temperature is around 270K, I would suggest using O4 cross section at 273K as one of the sensitivity cases.

In principle we agree with the reviewer here. However, we did not change the O4 cross section because of two reasons:
a) the effect of such small temperature changes is rather small.
b) in most existing studies, O4 cross sections for room temperature were used. Thus we prefer to stay consistent with those studies.

There is another change I would recommend here – what quantity is actually compared. Since the actual measurements are ground-based hyperspectral sky radiances the derived variable directly from the measurements without any assumptions about the atmosphere (accept for species effective temperatures) is the differential slant column density (dSCD).
There are no passive measurements at the bottom of atmosphere that do not contain O4 absorption, including the reference used in this study (zenith direction). From Beer's law, ignoring wavelength shift, offset and other corrections:

$$\left(\frac{\ln\left(I_{90}^{measured} - I_{VEA}^{measured}\right)}{\sigma_{O4}(T)}\right)_{\lambda\ window} = dSCD_{VEA}^{measured} =$$

$$= \underbrace{SCD_{VEA}^{total} - SCD_{90^o}}_{individual\ components\ are\ not\ measured\ directly}$$

$$dAMF_{VEA} = \frac{dSCD_{VEA}^{measured}}{VCD} = \frac{SCD_{VEA}^{total} - SCD_{90^o}}{VCD} = AMF_{VEA} - AMF_{90^o}$$

From the above discussion AMF and dAMF are quantities derived based on the assumptions made about AMF90 and VCD:

$$AMF_{VEA} = dAMF_{VEA} + AMF_{90^o} = \frac{dSCD_{VEA}^{measured}}{VCD} + AMF_{90^o}$$

I believe the paper will benefit if dSCD are compared directly with the RT modeled dSCD in the first section of the paper.

In our opinion, the only difference to your suggestion is that we divide the O4 (d)SCDs by the O4 VCD. Both choices are equivalent. To make the interpretation of the results in units of (d)SCDs easier, we added second y-axes in Figures 2 and 3 in (d)SCD units.

Minor comments:

1. The paper is very long and difficult to read due to constant references to the appendices and main body figures and tables. Some of the figures and tables can be consolidated or eliminated.

We understand this concern. However, one important part of the study deals with the quantification of the uncertainties of the spectral analysis and radiative tarnsfer simulations. For readers with interest in the details of the sensitivity studies the figures and tables in the appendix will be important. In contrast, for readers who are mostly interested in the general findings the figures and tables in the main part should be sufficient. We therefore decided not to remove any figures or tables.

2. Clear days are probably more appropriate to call cloud-free?

Changed

3. L 49 … agree within 1% with the corresponding radiative transfer simulations at 360 nm

‘at 360 nm’ was added

4. L246: which version of LIDORT is used in this study?

Version 3.3. The version is 3.3. This information was added to the text.

5. L277: rephrase to make clear that the comparison is done between hyperspectral fitting DOAS analysis vs. singe wavelength

We added the following text: ‘at one wavelength (here: 360 nm)’

6. What is the source of extraterrestrial irradiance used for synthetic spectra simulation?

We used the high resolution solar spectrum from Chance and Kurucz (2010). We added this information and the corresponding reference in section 2.4.

7. L293: Level 2 data are available now. It will be good to comment how it compares to level 1.5.

Many thanks for this hint! The Level-2 data are exactly the same as the Level-1.5 data. We removed the corresponding sentence about the Level 1.5 data from the text.

8. L306: Link from the pdf does not work, URL is valid.

Many thanks for his hint! This link should work after the final copy-editing.

9. Abstract refers to the campaign MAD-CAT, other places MADCAT

Now consistently ‘MAD-CAT’ is used.

10. L348: Intensity Offset polynomial of order 2 is quite large. Can you please explain why it was chosen?

The following text was added in section 4.3.2:

‘Higher order intensity offsets might compensate for wavelength dependent offsets (e.g. spectral straylight), which can be important for real measurements, while the synthetic spectra do not contain such contributions.’

11. L903: Can you please explain how 1.01±0.16 and 0.71±0.12 are calculated? Is this for the entire two days and all observation geometries?

The information was added that the ratio was calculated for the middle period of that day.

Time scale on Fig. 1 for the top panel (A) is unclear.

The corresponding labels were added.