Response to Referee #1

We would like to thank the reviewer for carefully reading the manuscript and for providing helpful comments, remarks and suggestions. You can find below our responses in red after each individual comment:

Due to the reordering of Sect. 3 (Results) and adding new figures, all references to text and figures of the manuscript that are mentioned below refer to the revised version, while in square brackets [] they refer to the preprint.

e.g., Figure 6 [8] refers to Figure 6 of the revised version and to Figure 8 of the preprint.

General comments

Karagkiozidis et al. present a comprehensive comparison and validation study of two MAX-DOAS profiling algorithms. The algorithms retrieve trace gas and aerosol profiles from MAX-DOAS observations over Thessaloniki, Greece.

The manuscript is well written, the analyses have been performed thoroughly and the conclusions are interesting.

However, while reading this document, I was wondering, what is the aim of this study? From the title, I expected a characterization of the temporal and spatial distribution of NO2, HCHO and aerosols over Thessaloniki. But the authors focused mainly on the comparison and validation of two profiling algorithms. Algorithms which have already been validated in other studies! In my viewpoint, the authors should change the manuscript slightly in order to go more in the direction of either a pure algorithm validation/verification paper or a characterization paper of Thessaloniki's trace gas/aerosol distribution.

If the authors decide for case 1, I would expect a detailed comparison of vertical profiles. If validation is not possible due to sparse measurements of ancillary instruments, please add a comparison of temporal/spatial mean profiles of both algorithms. I was also wondering if MAPA retrieves concentrations in higher altitudes compared to MMF? On the other hand, MMF does not retrieve small VCD's even though the correlation with in situ data is high. Does the a priori SH of 1km leads to this constrain? If the manuscript is modified based on these suggestions, please change the title accordingly.

In case the authors decide for a characterization paper of the tropospheric composition over Thessaloniki, I would expect a discussion of weekday to weekend variations. I would also expect diurnal variation plots. Furthermore, in this case, the analysis of HCHO is insufficient. Even though validation is not possible we learn nothing about the spatial distribution of HCHO from your study. You neither show HCHO profiles nor do you talk about possible sources (for all species).

In both cases, I would love to see some contour plots of seasonal mean profiles for all species.

The point made by the reviewer about focusing on the intercomparison and validation of two profiling algorithms in this study, instead of characterizing the aerosol and trace gas distribution at Thessaloniki is very reasonable. The main motivation for conducting this study was the absence of MAX-DOAS derived information on the vertical distribution of aerosols and trace gases in the troposphere at Thessaloniki, even though MAX-DOAS measurements are regularly performed since 2014. As the reviewer states,

MMF and MAPA have been tested in other studies and based on their performance they were accepted and adopted by the FRM$_4$DOAS project. By using two well-performing algorithms, we expect to retrieve profiles (and subsequently VCDs/AODs, surface concentrations, etc.) of higher quality and we tried to validate them with ancillary, reference data when available. In addition, we wanted to evaluate the ability to retrieve vertical profiles with a recently installed MAX-DOAS system that has superior characteristics than the instruments used in the past (e.g., wavelength range, spectral resolution, signal to noise ratio, field of view). We agree that additional analysis of the data (i.e., characterizing the aerosol and trace gas distribution at Thessaloniki) would be scientifically meaningful. However, in our belief, this paper already contains a lot of information and including an additional geophysical analysis would significantly increase the size of the paper and make it difficult for the reader to follow. Such analysis could be part of another paper in the future. Thus, we have followed the first suggestion of the reviewer and focused to an intercomparison/validation study and we have revised the title of the paper to "Retrieval of tropospheric aerosol, NO_2 and HCHO vertical profiles from MAX-DOAS observations over Thessaloniki, Greece: Intercomparison and validation of two inversion algorithms".

The reviewer correctly pointed out that according to the title, vertical aerosol and trace gas profiles should be presented, yet they are not shown in any of the figures. As the reviewer suggests, we included an additional subsection in the results section were we discuss seasonal mean profiles for all species retrieved by the two algorithms. In this subsection the vertical profiles are intercompared, the variability of each algorithm is discussed along with possible sources for each species. Similar comments about the absence of the vertical profiles in the plots and insufficient information about HCHO were made also by reviewer #2. Along with the seasonal mean vertical profiles of HCHO we included an extra appendix showing the comparison of the VCDs retrieved from the integration of the profiles with the VCDs that are obtained using the geometrical approximation in order to make a link with previous datasets.

Please also add the following points:

1. Even though the applied flags have been applied elsewhere, please add a table of flags for each algorithm in the appendix. I guess that flagging thresholds might differ for UV and vis?

As the reviewer states, flags and thresholds of the algorithms have already been discussed in detail in other studies (e.g., Beirle et al., 2019). However, it is true that a short discussion about flagging should be present in the manuscript. Instead of adding an extra table that contains information about the flags and thresholds in the appendix, for better readability we included a short discussion for each algorithm in the relevant sections (2.4 for MMF and 2.5 for MAPA). The default flagging thresholds in MAPA are the same for the UV and VIS spectral range. The user can change the thresholds and even set different values for the UV and VIS. Yet, in this study we use MAPA with the default configuration.

2. What is the conclusion of your flagging scheme discussion? I would consider your results as unclear. Maybe there is no clear conclusion to be made?

It is true that no clear conclusion could be achieved based on this study about which flagging algorithm (MMF's or MAPA's) performs overall better (the text has been revised). However, the flagging

algorithms have been evaluated during the comparison of MAX-DOAS derived products (AODs and NO_2 surface concentrations) with reference instruments. For the AODs in the visible range, higher correlation coefficient values are achieved between the MAX-DOAS and the CIMEL/Brewer when both MMF and MAPA contribute to the flagging; yet, this is not the case for AODs in the UV, where MAPA's flagging algorithm leads to better agreement. For NO_2 surface concentrations, the effect of the flagging scheme is not as strong as for aerosols. The combined flagging leads to slightly better agreement with the in situ data. Individually, the MMF's flagging performs slightly better than MAPA's since it results in higher correlation with the in situ data even though a much smaller fraction of the data is flagged as invalid (Table 3).

3. Please add a short discussion of NO2 retrieved in the UV. You have mentioned that HCHO and aerosols in the UV might be negatively affected by increased spectral noise. Is there a similar conclusion for UV NO2?

The reason for this suggestion of the reviewer is fully understood. However, since this instrument allows for the retrieval of NO_2 dSCDs in the visible range (425 – 490 nm), we don't normally retrieve NO_2 in the UV range, where larger fitting errors are expected due to the increased spectral noise in this region. Even though NO_2 dSCDs can still be retrieved in the UV, a lot of computational effort is required in order to retrieve its vertical profiles for approximately 1 year of data. Besides, including an "additional" species in the analysis would increase the size the paper without providing much extra valuable information (except for the differences between the UV and VIS).

4. Please add a short discussion of possible issues of your aerosol retrieval due to the inaccurate Henyey-Greenstein phase function at the proper sections in your manuscript.

For most viewing directions, the respective choice of the aerosol phase function has only a minor effect, because the main effect of aerosols is that they reduce the visibility of the atmosphere. However, for small scattering angles also the forward scattering properties of aerosols can become important. Thus for small scattering angles (< about 10° to 20°) the uncertainties caused by the incorrect description of the phase function can also become important, and the results for such viewing geometries should be treated with caution. Additional investigations are needed to quantify the respective uncertainties for small scattering angles. The discussion is included in the revised version of the manuscript (P14, L332-334).

5. If I understand the authors correctly, the instrument measures in an altitude of 80m. How is this "elevated" position treated by the algorithms? What is the meaning of the lowermost retrieval grid point in this context?

The retrieved profiles for both algorithms are provided for the altitude range 0 – 4 km with 200 m vertical resolution refer to altitudes above the instrument. The term 'surface concentration' does not refer to the concentration directly at ground level, but to the mean near-surface concentrations in the first 200 m above the instrument. This is done because the profile parameterization used within MAPA allows for the

retrieval of lifted trace gas layers for a shape parameter (s) greater than 1. In such cases, the surface concentration of the parameterized profile would be zero and the comparison with the in-situ measurements would lead to low-biased MAPA results. Therefore there was no need for extrapolation to obtain the 'surface concentrations'. This approach is also used in other studies (e.g., Tirpitz et al., 2021). The altitude of all the remote sensing instruments is 60 m (and not 80 m, as was wrongly stated) and is given relative to the sea level. The in situ measurements of NO_2 are performed at a site that is located 174 m above sea level, i.e., 114 m above the MAX-DOAS instrument, and is well within the first 200 m layer retrieved by the MAX-DOAS.

Specific comments

P2, L43: "can lead to or") can lead to ... or deteriorate ...

Done.

P3, L81 - L84: You mention that the data is also analyzed regularly within the FRM4DOAS project. Please name the specific differences in retrieval settings between your study and the regularly submitted data and the reason for specific changes of settings. It would also be interesting to compare FRM4DOAS data with your new settings.

Done: A discussion is included in the paper.

Fig. 2: Please add other instruments if not measured at the same location (e.g. in situ).

Done: In-situ measurement site location is now also included.

P10, L211: "by assuming a correlation length") "by assuming a Gaussian function with correlation length of...". Note that a correlation length of 50m was used in the cited publication!

Done.

P10, L224: What is the lowermost retrieval altitude for each algorithm? Surface values were extrapolated?

Please see our response for the 5th point above.

P12, L285: Why did you use hourly mean values? You could also average all in situ values for the corresponding MAX-DOAS elevation scan cycles.

The suggestion made by the reviewer is very reasonable. However, this would not be possible in our case because hourly mean NO_2 concentration is the highest temporal resolution that is available for the in-situ measurements from the air-quality network stations.

P13, L288 - L292: I don't understand your reasoning here. I guess that tracffic emissions contribute strongly to the MAX-DOAS signal but then an in situ site should not be a background site. How far away is the next site in viewing direction of the telescope?

The municipal network of air-quality stations that are distributed around the city center are installed very close to the ground (sampling inlet at ~3 m) and are strongly affected by local traffic emissions. Furthermore, they are located closer to the sea level and are therefore about 30 – 40 m below the MAX-DOAS. Due to its higher altitude, the MAX-DOAS system measurements are considered more representative of the NO_2 concentrations in the local boundary layer and less influenced by the local traffic of the city center. The selected for the comparison air-quality site, which is located 114 m above the MAX-DOAS system, probes air that is also more representative of the local boundary layer.

P13, L304: 5° is already quite small, especially when using Henyey-Greenstein. Have you tried different values? I would expect that 10° improves data quality significantly but might decrease the number of data points (maybe too much?).

At this point, we wanted to exclude from the analysis the elevation scans that have been performed very close to the solar azimuth angle due to potential errors of RTM simulations. The limit of 5° was selected arbitrarily in order to keep a balance: The elevation scans close to the solar position are rejected and in the meantime there is no major loss of the data. As Reviewer 1 already mentions, using 10^o or more as relative azimuth angle filtering would lead to a significant reduction of the data points number.

P13, L308-L309: I am wondering how negative columns can make it through any flagging step? Also 8.5% is a really large fraction of invalid profiles. Is there any reason known why MAPA produces so many unrealistic profiles?

Due to the low SNR in the UV, larger scatter of the retrieved dSCDs (especially for HCHO) is expected and consequently also of the corresponding profiles. Retrievals of negative columns are not necessarily flagged as invalid by default in order to keep the means unbiased. By doing so, in cases where no tropospheric trace gas is present the retrieved profiles would then contain pure noise. Otherwise, the resulting mean would be high-biased (greater than zero).

P14, Table 3: When looking at the HCHO fraction (also aerosols in UV) of valid profiles for MAPA, I am really worried about the general performance of MAPA in the UV. Is there any particular reason for this bad performance? There was already a BIAS found for MAPA's HCHO results in Tirpitz et al. 2021 so I don't think that noisy data can explain everything!

Since in MAPA retrievals no a priori constraints are used, more strict flagging needs to be applied for retrieved dSCDs that are characterized by large uncertainties (e.g., due to larger fit error or the effects of clouds). As a result, a smaller fraction of the data is flagged as valid. Especially for HCHO, the apparent worse performance of MAPA could be explained by the lower SNR in the UV, along with the higher HCHO profile height compared to NO_2 and the decreasing sensitivity towards higher altitudes. Moreover, the trace gas retrievals depend also on the aerosol retrievals, so the respective uncertainties for high AODs would be larger than for low AOD.

P15, Figure 5: For reach row, MAPA shows values close to zero, except for NO2. I am not sure if it is a good thing, that MMF doesn't show small values at all or that MAPA cannot find them only for NO2. Could you please say something about that? And again, I would be interested in MAPAS flagging thresholds and if they differ in the visible and UV spectral range.

The sensitivity of the MAX-DOAS decreases with altitude and it is very limited at altitudes above 2.5 km for the species measured in the VIS spectral range or even lower (1.5 km) for the species in the UV (Figure 6 [8]). For NO_2, this is generally not a problem since the total column is dominated by the concentration in the lower layers of the troposphere (see also Figure 9 and discussion 3.5 of the revised version of the manuscript). However, HCHO can be vertically extended at higher altitudes, where the sensitivity of the MAX-DOAS is low. In the case of HCHO, MMF is more prone to result in the a priori profile, while MAPA retrievals become more unstable. Thus, the vertical profiles of MAPA are expected to have greater variability. For aerosols, the discrepancies between MMF and MAPA retrievals are also due to an additional factor: The variable O_4 scaling factor that is included in MAPA (O_4 SF = var), while no scaling factor is applied to MMF retrievals (O_4 SF = 1).

P15, Figure 6: I think this figure tells us that MMF has a positiv Bias for low elevation angles (reddish dots more often over black line) which would also explain why we don't see small values in Fig. 5 for MMF. It seems that the algorithm has problems in retrieving accurate profiles for small dSCD, especially in the UV. This could be explained by more noise but the MAPA results seem to be unaffected. Do you have any explanation for the different LOS depending performance of both algorithms?

We assume that this comment of the reviewer is associated with the O_4 dSCDs (especially in the UV) rather than the trace gas dSCDs. The positive bias for low elevation angle that the reviewer mentions is not apparent for NO_2 and neither for HCHO, where the data is much noisier. It is true that MAPA results seem to be less affected. The main driver for this behavior would be the different O_4 scaling factors applied to the retrievals of MMF and MAPA. While MMF does not include a scaling factor, MAPA fits an optimum O_4 scaling factor in order to bring measured and simulated dSCDs into better agreement.

P18, L382 - L384: Concentrations for the lowermost layer rather than conc. at ground? Do you mean the lowermost layer with concentrations larger than zero? If not, please explain!

The output grid of both MMF and MAPA ranges from the ground up to 4 km with 200 m vertical resolution. The term “surface concentration” in Sect. 3.3 does not refer to the concentration at ground level and also it does not refer to the lowermost layer with concentrations larger than zero. It refers to the average concentration in the lowest 200 m above the instrument, as retrieved for the MAX-DOAS first profile layer (i.e., 0 – 200 m). The text has been revised.

P18, Figure 7: Again, MMF doesn't show HCHO values close to zero which means that the main HCHO concentration is found in higher altitudes. MAPA seems to retrieve HCHO closer to the surface. However, P18, L382 - L384 tells us that this conclusion might be wrong. So I am wondering if you could show a similar figure with surface concentrations only? I have to admit that I am confused by the sentence P18, L382 - L384 and the fact that MAPA finds HCHO concentrations close to zero!

The non-zero near-surface MMF HCHO results are probably a result of the non-zero surface concentrations in the a priori profile. They are not a direct consequence of possibly enhanced HCHO concentrations at higher altitudes (note that both retrieval results use the same input measurements; thus the information content is the same for MAPA and MMF).

Concerning the statement in [P18, L382 - L384]: It is true that the surface values are also influenced by HCHO concentrations at higher altitudes. But for the retrieval results close to the surface, the sensitivity of the MAX-DOAS measurements clearly peaks for HCHO close to the surface. Thus the influence of HCHO at high altitudes (> 500m) is usually rather small.

P21, L439 - L440: I am not sure if I understand scheme #3 correctly. In this line, you write about warning flags while you use "erroneous" in Table 4. Please describe this scheme more detailed.

Corrected: From “Data that are not flagged erroneous neither by MMF nor by MAPA are considered valid” to “Data that are flagged as warning by either MMF or MAPA are also considered valid”. Both MMF and MAPA flag the data as either valid, warning or error. Scheme #3 rejects the error flagged data but treats the warning flagged data as valid.

P24, L493 - L494: "Aerosol layers between 2 and 4 km are "invisible"...". This is not correct! An elevated layer will for sure be identified as elevated layer in these altitude regions if aerosols below are negligible. MAX-DOAS might not find the correct altitude but the elevated layer will be identified for sure showing a small but existing sensitivity.

This statement was used in order to explain the difference in the profiles of the MAX-DOAS and the lidar between 2 and 4 km for this particular case scenario of 05-Jun, but perhaps it could be interpreted as a more generic statement by a reader. We revised this statement to: “Aerosol layers between 2 and 4 km that are detected by the lidar cannot be well retrieved by the MAX-DOAS, due to its limited sensitivity at these altitudes”.

Figure 12: It is hard to say which profile is the best, especially for the 21. of July. Could you please add a subplot showing the modelled and measured dSCDs at each elevation angle for all scenarios and both algorithms? Maybe this helps to assess better the performance here.

The figure below shows the measured and simulated O_4 dSCDs as a function of the elevation angle (left) and the correlation plots between measured and modeled dSCDs for MMF and MAPA (right). The performance of the forward models is similar with no clear conclusion of which simulates the true state (measured) better. The aerosol extinction profiles that are shown in Figure 13 [12] are retrieved by the MAX-DOAS in the visible range. As seen in Figure 9 (Sect. 3.5 of the revised version) larger discrepancies are found in the profiles of MMF and MAPA during summer. Since these plots are not very helpful in resolving this issue, we have not included them in the revised manuscript.

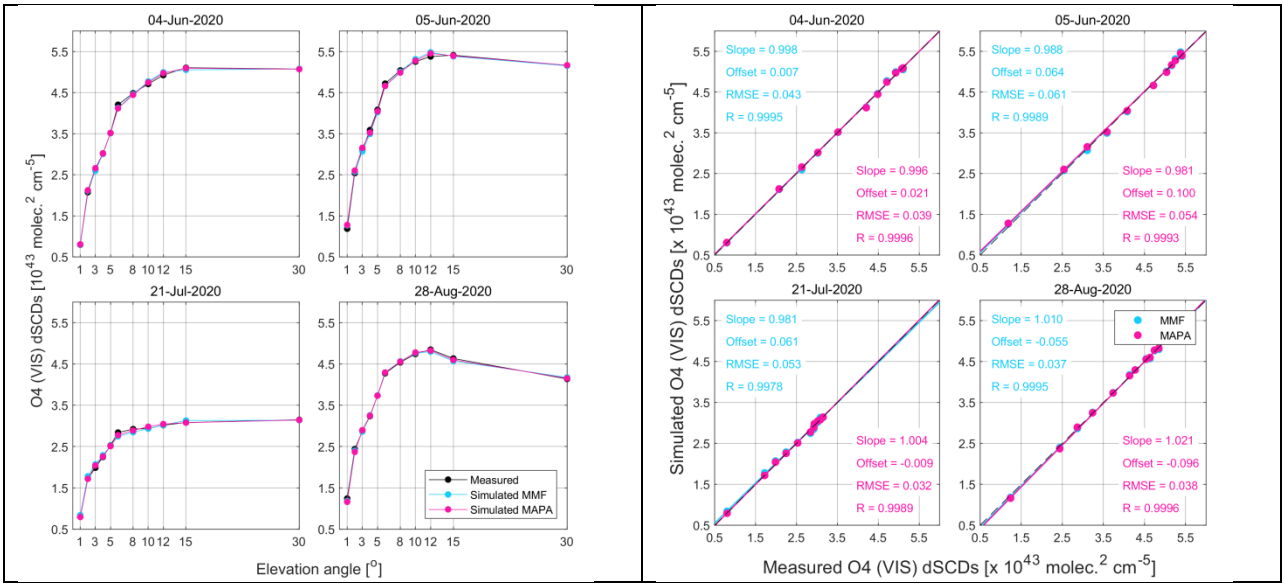

Figure A3: As you have mentioned, the error bars for the scaling factors are larger in winter than in summer.I was wondering if the number of data points in winter is large enough to show a mean daily variation for January (and compare with a similar curve from August)? Do these curves show a clear diurnal cycle?

In this study we do not investigate the diurnal cycle of the O_4 scaling factor, because no clear conclusion can be drawn, mainly due to insufficient number of data during winter, as the reviewer has pointed out. This study is based on approximately 1 year of data. It should be noted also that only the O_4 SFs for which $65^o < SZA < 75^o$ are presented. Nevertheless, despite of the large error bars, an apparent seasonal pattern is observed that could be further investigated when longer time series of MAX-DOAS measurements become available.

References

Beirle, S., Dörner, S., Donner, S., Remmers, J., Wang, Y., and Wagner, T.: The Mainz profile algorithm (MAPA), Atmospheric Measurement Techniques, 12, 1785–1806, https://doi.org/10.5194/amt-12-1785-2019, 2019.

Tirpitz, J.-L., Frieß, U., Hendrick, F., Alberti, C., Allaart, M., Apituley, A., Bais, A., Beirle, S., Berkhout, S., Bognar, K., Bösch, T., Bruchkouski, I., Cede, A., Chan, K. L., den Hoed, M., Donner, S., Drosoglou, T., Fayt, C., Friedrich, M. M., Frumau, A., Gast, L., Gielen, C., Gomez-Martín, L., Hao, N., Hensen, A., Henzing, B., Hermans, C., Jin, J., Kreher, K., Kuhn, J., Lampel, J., Li, A., Liu, C., Liu, H., Ma, 820 J., Merlaud, A., Peters, E., Pinardi, G., Piters, A., Platt, U., Puentedura, O., Richter, A., Schmitt, S., Spinei, E., Stein Zweers, D., Strong, K., Swart, D., Tack, F., Tiefengraber, M., van der Hoff, R., van Roozendael, M., Vlemmix, T., Vonk, J., Wagner, T., Wang, Y., Wang, Z., Wenig, M., Wiegner, M., Wittrock, F., Xie, P., Xing, C., Xu, J., Yela, M., Zhang, C., and Zhao, X.: Intercomparison of MAX-DOAS vertical profile retrieval algorithms: studies on field data from the CINDI-2 campaign, Atmospheric Measurement Techniques, 14, 1–35, https://doi.org/10.5194/amt-14-1-2021, 2021.