Reply to the comments from referee # 1

(In red our replies, in green the text added.)

We first kindly thank the referee for his time, useful comments, and constructive criticism. We used his suggestions to prepare a new version of the manuscript.

This paper discusses the air quality measurement campaign, AROMAT, and puts it into the context of if/how these measurements can assist in future validation efforts for satellites, such as Sentinel 5P TROPOMI. The significance of this work is within the scope of AMT and is key as the air quality community works toward validating satellites that measure urban air quality (e.g., TROPOMI) and there are some novel 'take-home' messages from this work that are worthy for publishing. From what is shared in the paper, the quality of the work appears valid however there is clarity needed in some areas. This paper also needs restructuring to improve the clarity of the take-home conclusions. For example: The paper lacks details about the campaign and information about the measurements are scattered throughout the paper.

We have added two sections to describe the campaign deployment and mention more intensively the supplement when appropriate, the supplement contains more details and measurements. We also added a schematic for the overall deployment.

Specific comments/questions:

• The title of this paper does not clearly reflect the contents of the paper. The current title would attract readers as an overview paper for the measurements during AROMAT, but this is not the purpose of this paper nor is there a detailed overview of the entire campaigns. If agreed by the authors, I suggest changing to title to something that reflects that AROMAT could be a concept model for validation campaigns of satellite retrievals.

This paper aimed to be an overview paper but we agreed we had put too much information in the supplement so we added section 2.4 and 2.5 which describe the campaign deployments, they were previously in the Supplement. We prefer to keep our title.

• Are conclusions made about validation only valid over Romania? Or can these lessons be extended beyond Romania? Please be clear in the paper which conclusions can be extended beyond the AROMAT region.

About NO2, since Bucharest is a relatively small source compared to other cities, many other urban areas could be used as target for satellite validation using airborne measurements, we have performed other airborne measurements in Berlin and in Belgium in the AROMAPEX and BUMBA campaign which are already in the references.

About SO2 and H2CO, our conclusions are valid for Romania only since there are larger sources in other parts of the world. We have already mentioned Serbia in the conclusions.

We added a sentence to extend the scope of the conclusions for NO2:

These conclusions for NO2 above Bucharest apply to other large polluted urban areas.

• It seems that the model for the conclusions is based on TROPOMI requirements. Please comment on if/how this extends to the requirements of the other planned missions or make the specific message in the paper that the conclusions that are made are specific to TROPOMI.

The qualitative part of the conclusion is also valid for future satellites which are in Table 1 but the quantitative accuracy we give is indeed based on TROPOMI characteristics. We rephrased

Our simulations, which are based on our measurements and TROPOMI characteristics, indicate that we can constrain the accuracy of the satellite NO2 VCDs within 37 or 28%, with and without information on the aerosol and NO2 profile, respectively.

• A weakness in the general analysis is the lack of discussion on temporal variation and the time of the airborne and ground based measurements and how this relates to the time of the satellite overpass, emissions inventories, etc. The authors should keep this in mind to address through revisions.

We do not fully agree with the comment. When the time difference could explain an observed discrepancy we have mentioned it, e.g for the AirMAP to MPIC Mobile-DOAS comparisons in Bucharest

Note that the systematic differences between AirMAP and the MPIC Mobile-DOAS at the eastern part of the ring road on 31 August 2015 were due to the time differences between both measurements.

About the satellite validation, we have assessed the temporal error (in Section 4.1.2) for the comparison between an airborne and a satellite instrument. This is anyway more visible in this new version of the manuscript since we have included in the main manuscript the figure from the supplement which illustrates how we quantified the temporal error.

About the emissions section, we have added a sentence in the new paragraph presenting this section to emphasize the fact that we compare instantaneous emissions with yearly emission inventories:

The comparisons with reported emissions should not be overintrepreted since we compare campaign-based flux measurements performed during a few days in daytime with reported emissions which represent yearly averages. Nevertheless, they give interesting indications about the operations of the FGD units of the power plants and possible biases in emission inventories.

• Section 2 should start by painting a picture of AROMAT 1 and 2 deployments and measurements that used in this analysis. While much of this information is in the supplement and scattered throughout the paper, the general reader enters Section 3

without the proper background to assess what is being discussed. Currently, it does not effectively communicate the needed details about the AROMAT campaign before moving into the results sections. To fix this, the authors could reorganize the section by moving 2.3 to before 2.1 and 2.2. Then there needs to be discussion (and maybe a Table) that summarizes each campaign. This table and/or discussion must include time periods of each deployment, location of each deployment, payloads for the aircraft and relevant details about ground measurements (in line with Tables 4-6) and could extend into partners and other details from 2.3 as seen fit.

We thank the referee for his suggestion but we do not want to move Section 2.3 before Section 2.1 and 2.2 because we think it is clearer to present the sites before the campaign deployment on these sites. About the tables for the campaign, we also prefer to let them in the supplement since they are quite large and include ancillary measurements (ground-based in-situ, ACSM…) that we do not use later on in the main manuscript. So we think it would be distracting to put them in the main article.

But for clarity, we have moved the sections which present the practical deployments in 2014 and 2015 from the supplement to the main manuscript. We have also added a schematic for the campaign. We present this new schematics in Sect 2.3.

Figure 4 illustrates the typical instrumental deployment during the campaigns. The set-up combined airborne and ground-based measurements to sample the 3-D chemical state of the lower troposphere above polluted areas.

Table 3 does not add substantial information to this paper and that space would be more effectively be used to summarize the campaigns themselves.

We agree with the comment and moved Table 3 to the Supplement.

Section 3 and 4 are hard to follow as its jumps between regions and different trace gas measurements. A suggestion would be to reorganize into sections focuses on specific trace gases. For example: Section 3 could just be about NO2. With the following sections.
3.1: similar for 3.1.3 with summarizing Bucharest observations
3.2: similar for 3.2.1 with summarizing The Jiu Valley observations
3.3: Relevant discussion from Section 4 about lessons about validation

We thank the referee for his advices but we prefer to keep our structure because it follows in Sect. 3 a geographical order with the two sites which corresponds to the geographical presentation of Sect.2. This is interesting for a reader interested in pollution sources in Romania. In Sect. 4 we draw conclusions for each molecule, which is more interesting for a reader coming from the AQ satellite validation community. We have written a new introduction for Sect. 4.4 which presents our flux estimates.

• Section 3.1.1 and 3.1.2 and their associate figures do not fit within the scope of the paper as separate sections. Any relevant discussion could fit in within the other trace gas sections, in the supplement, or omitted.

We agree with this comment. We have moved the figure to the Supplement.

The below of the comments are organized by trace gas.

• **NO2:**

Line 209: The statement about the datasets in Figure 7 appearing consistent is not valid, which is alluded to later in the paragraph. Please reword or omit that statement in the discussion.

We meant that the measurements were consistent for each given day, not in general. We have removed this statement which was indeed misleading.

Line 221: what is the difference in time between the two measurements?

The AirMAP/Cessna VCDs correspond to several flight lines recorded between 12:00 and 13:30 UTC. We have added this info to the text, which already included the CAPS/BN-2 time.

between 12:00 and 13:30 UTC

The statement about NO2 vmr at 300m being a proxy for NO2 VCD is not valid. It may be for that specific case but not overall. The results over the Jiu Valley even refute this statement.

Although we had written at the beginning of the sentence “along this portion of the flight, which was inside the plume but outside the city,” we agree that the scope of validity of this statement was not clear enough. We rephrased this paragraph for clarity:

This suggests that along this portion of the flight, which was inside the plume but outside the city, the NO2 VMR measured at 300 m a.s.l. may be used as a proxy for the NO2 VCD. Indeed, the BLH was about 1500m (Fig.S9 in the Supplement) during these observations. Assuming a constant NO2 250 VMR of 3.5 ppb in the boundary layer leads to a NO2 VCD of 1.4 x 1016 molec cm2. This estimate is close to the AirMAP NO2 VCD observed in the plume (Fig. 6). When measured at 300 m a.s.l., the NO2 VMR thus seems a good estimate of its average within the boundary layer. Note that this finding is specific to the configuration in Bucharest where we flew at 10 km from the city center and does not apply to our measurements in the exhaust plume of the Turceni power plant (Fig. 9). Future campaigns should include vertical soundings inside the Bucharest plume to further investigate its NO2 vertical distribution.

Line 229: Is the Avantes spectrometer the Bremen Nadir instrument from Table 7? Please make descriptions consistent.

Indeed, we rephrased:

from the IUP-Bremen nadir instrument

Not required but Figure S3 seems like a good candidate to move to the actual manuscript for comparing/contrasting with SO2. It could also be helpful to see how Figure 7 and other airborne figures translate to the TROPOMI pixels. When talking about the validation context.

We agree with these comments which improve the manuscript. We have moved Figure S3 to the main manuscript, merging it with the SO2 map. The figure S10 of the previous version of the Supplement, which illustrated the temporal variation, also shows how airborne measurements translate to hypothetical TROPOMI pixels. We have also moved it to the main manuscript.

Line 334: It seems that temporal variation could also lead to overestimation in the slope depending on how the NO2 is varying through time.

A systematic variation of the NO2 VCD through time would lead to a bias between reference (x) and satellite measurements (y) if the NO2 VCD, this bias could be either positive or negative.

But considering only a random variation added to the reference measurement as a noise, the dynamic range of this 'noisy' reference measurements is likely to increase, which could lead to an underestimation of the slope between x and y without this temporal noise. We have actually made some simulations to emphasize this effect and added a figure (Fig.12). Here is its legend:

Effect of an underestimation of the random error in a regression analysis simulating TROPOMI validation using airborne mapping as reference measurements of NO2 VCDs. The dynamic range (blue line) of the reference measurements increases with the applied random error. For the considered a priori random error (dashed vertical line, 1x10^15 molec.cm-2), this leads to an underestimation of the regression slope (red line). These simulations use the AirMAP data of 31 August 2015 (afternoon flight).

And the added text

Finally, it should be noted that these regression simulations assume a correct estimation of the temporal random error. Underestimating this error propagates in the fit of the regression slope. Figure 12 presents the possible effect of such an underestimation when the a priori random error of the reference measurements is set at 1x1015 molec cm-2, using again the AirMAP observations of Fig. 5 (right panel) as input data. As the dynamic range of the reference measurements increases with the applied error, the fitted slope decreases. For a true error of 4x1015, this leads for instance to an underestimation of the slope of about 5%. This effect is small but other sources of random error (e.g undersampling the satellite pixels) would add up in a real-world experiment. Wang (2017) observed such a systematic decrease of the regression slope when averaging MAX-DOAS measurements within larger time windows around the satellite overpass.

It should be noted that the temporal variation uncertainty quantified in this paper was specific to that area during that particular morning and more data would have to be analyzed to see if this is a typical value or not. These temporal variations are also

likely much different during the time TROPOMI overpasses (not in the early morning) and on different days. Though the technique for quantifying temporal variation using the airborne data is novel and would be interesting to extend to other datasets.

We agree with the comment. We had already emphasized that writing *Clearly, the NO2 VCD temporal variation depends on characteristics of a given validation experiments, such as the source locations and the wind conditions during the measurements* in Section 4.1.2 and in the conclusion *it varies with local conditions for a given experiment*.

We have further emphasized that it also depends on the time of the day and that our measurements were not at the TROPOMI overpass time.

The temporal variation also depends on the time of the day and we base our estimate here on measurements around 11:00 LT while TROPOMI overpass is at 13:30 LT.

It would also be helpful to add some more details in the writing or references about the exercise done in the first paragraph of section 4.1.3 so it can be recreated by others with similar datasets.

It seems to us that the important steps of our method are already described but we have added a reference to the section 4.1.1 describing the input data, to improve the clarity.

We simulated TROPOMI Cal/Val exercises with the spatially averaged AirMAP observations described in Sect. 4.1.1

Moreover, we have moved to the main manuscript the figure from the supplement which illustrates the quantification on the temporal error as it also shows the average of the airborne measurements at the resolution of TROPOMI so we think it improves the overall clarity of this section. This figure is described in the previous section :

Figure 11 illustrates our estimation of the temporal variation of the NO2 VCDs comparing consecutive AirMAP overpasses above Bucharest from the morning flight of 31 August 2015.

• H2CO and SO2

Line 239-240: Are there H2CO direct emissions in Bucharest? That seems to be the implication with the statement in line 239.

We do not know that. We clearly measured an enhancement of H2CO above Bucharest but we can not conclude on its origin. To investigate that, we would need 1) more measurements, in particular of the VOCs which are the main H2CO precursors, and 2) modelingstudies. Such work was done in particular by Johansson et al. (2014)

Johansson, J. K. E., Mellqvist, J., Samuelsson, J., Offerle, B., Moldanova, J., Rappenglück, B., … Flynn, J. (2014). Quantitative measurements and modeling of industrial formaldehyde emissions in the Greater Houston area during campaigns in 2009 and 2011. *Journal of Geophysical Research: Atmospheres*, *119*(7), 4303–4322. https://doi.org/10.1002/2013JD020159

But it is outside the scope of our study. We agree that our word 'source' is misleading as it can be understood as direct source. So we have rephrased:

The difference between NO2 and H2CO spatial patterns may be explained by the different origins of NOx compared to H2CO.

This is relevant to both SO2 and H2CO since they both have the conclusion that validation of satellite H2CO and SO2 is better suited with ground-based measurements.

Is this a recommendation only for Romania?

As replied above, this is valid for Romania. There are larger SO2 sources such as volcanoes or oil industry in the Persian Gulf which may be interesting for airborne validation. For large power plants outside Romania without FGDs, one should study the satellite data to verify that their signal-to-noise ratio enables their validation with airborne measurements in practice (e.g. considering the costs). Note that the signal-to-noise may improve with geostationary platforms.

Are there ground based measurements from AROMAT that can be discussed in terms of validation like the airborne data is? If so, add this to the discussion. If not, the conclusion that ground-based measurements would be better suited than airborne may not be valid.

We have discussed the interest of our Mobile Max-DOAS (Ground based) for H2CO in Section 4.2, and the interest of the temporary SO2 cameras (Ground based) for SO2 in Section 4.3. These findings motivate the installation of automatic and static instruments since 1) MAX-DOAS are already demonstrated for H2CO VCD validation 2) TROPOMI-derived fluxes are already demonstrated and their volcanic part is already used for SO2 validation. So we consider our conclusions valid. But we agree we missed a reference for H2CO. We added a reference to Desmedt (2015) in the H2CO section, and modified a sentence:

Indeed, long-term ground-based measurements at two sites would be useful to investigate seasonal variations of H2CO, as already demonstrated in other sites (De Smedt, 2015).

Its mentioned that the individual flights cannot always help in validation, which is true. But systematic measurements may help as discussed in the conclusions. Please say something about this within the sections themselves.

We modified a sentence in sect 4.2:

This limits the relevance of individual mapping flights for the validation of H2CO, yet systematic airborne measurements would improve the statistics.

And added one in sect 4.3

As for H2CO, systematic airborne measurements would improve the statistics.

Emissions section (section 4.4):

• This section lacks sufficient background on the methodology for computing fluxes and lacks the context on how this fits with the scope of the paper. Fluxes are not mentioned in the abstract, intro, nor is emission estimate validation within the requirements for validation of satellite products. Though emission estimations put into the context of satellite applications is important and evaluating that is very important scientifically from that perspective.

Options:
§ Omit this section.
§ Add sufficient details or references for emission flux calculations and put this into the context on how this helps with satellite product evaluation as alluded to in the latter part of section 4.3. (Though, this section with all details could potentially be a stand-alone manuscript).
• If kept, when comparing the emissions to inventories, be sure to consider variations in emissions from hourly/daily/seasonal timescales and the AROMAT measurements were only a small subset in time.

We agree that this section was not introduced enough in the abstract and introduction and we have modified both to mention this part of the study. We added in the abstract:

We also quantify the emissions of NOx and SO2 at the two sites.

We added a sentence in the introduction:

Eventually, we use the AROMAT measurements to derive NOx and SO2 fluxes from the two sites.

Several studies already present in detail the traverse method used to quantify the fluxes with the DOAS method.e.g:

Ibrahim, O., Shaiganfar, R., Sinreich, R., Stein, T., Platt, U., and Wagner, T.: Car MAX-DOAS measurements around entire cities: quantification of NO_x emissions from the cities of Mannheim and Ludwigshafen (Germany), Atmos. Meas. Tech., 3, 709–721, https://doi.org/10.5194/amt-3-709-2010, 2010.

Johansson, J. K. E., Mellqvist, J., Samuelsson, J., Offerle, B., Moldanova, J., Rappenglück, B., Lefer, B., and Flynn, J. (2014), Quantitative measurements and modeling of industrial formaldehyde emissions in the Greater Houston area during

campaigns in 2009 and 2011, *J. Geophys. Res. Atmos.*, 119, 4303– 4322, doi:10.1002/2013JD020159.

From AirMAP, this is also explained in the AROMAT AirMAP study

Meier, A. C., Schönhardt, A., Bösch, T., Richter, A., Seyler, A., Ruhtz, T., Constantin, D.-E., Shaiganfar, R., Wagner, T., Merlaud, A., Van Roozendael, M., Belegante, L., Nicolae, D., Georgescu, L., and Burrows, J. P.: High-resolution airborne imaging DOAS measurements of NO_2 above Bucharest during AROMAT, Atmos. Meas. Tech., 10, 1831–1857, https://doi.org/10.5194/amt-10-1831-2017, 2017.

We added an introduction with these references at the beginning of the section :

This section presents our estimates of the NOx and SO2 fluxes from Bucharest and the power plants in the Jiu Valley, combining our different 2014 and 2015 measurements. Campaign-based estimates of NOx emissions from large sources are relevant in a context of satellite validation since the high resolution of TROPOMI enables to derive such emissions on a daily basis Lorente(2019). Regarding SO2, as discussed in the previous section, the low signal-to-noise ratio of the satellite measurements implies averaging for several months to derive a SO2 flux (Fioletov-2019), yet campaign measurements are useful to select an interesting site and test the ground-based apparatus and algorithms.

The comparisons with reported emissions should not be overintrepreted since we compare campaign-based flux measurements performed during a few days in daytime with reported emissions which represent yearly averages. Nevertheless, they give interesting indications about the operations of the FGD units of the power plants and possible biases in emission inventories.

Our flux estimates are all based on optical remote sensing measurements. They involve integrating a transect of the plume along its spatial extent and multiplying the outcome by the plume speed, which may correspond to the stack exit velocity (camera pointing to the stack) or to the wind speed (Mobile-DOAS and imaging-DOAS). We refer the reader to previous studies for the practical implementations. Ibrahim (2010) presented the method we used for Bucharest, where we encircled the city with the Mobile-DOAS. Meier (2017) presented the AirMAP-derived flux estimations, while Johansson (2014) derived industrial emissions from a car-based Mobile-DOAS instrument as we did for the Turceni power plant. Constantin (2017) presented the fluxes based on the ULM-DOAS measurements. Regarding the SO2 Camera, we present hereafter the method and previous related works.

Other comments within the text:

• Line 27-28: Veefkind et al., 2012 doesn't reference the 3.5x5.5km resolution. Refer to the switch through the Readme file or another reference that talks about it: http://www.tropomi.eu/sites/default/files/files/publicSentinel-5P-Nitrogen-Dioxide-Level-2-Product-Readme-File.pdf

We agree with the comment, nevertheless Veefkind (2012) is a more complete reference for TROPOMI. So we kept it and added the readme in the references and a note after the reference to Veefkind:

the original TROPOMI resolution of 7x5.5 km2 was increased on 6 August 2019 MPC (2019))

• Line 45: Can it be made clear what small signals mean? Does this mean the small signal:noise ratios or more a reference to clean areas that don't have a lot of signal?

There are several aspects of the small signals, as Richter et al. (2013) point out:

An additional challenge is the small signal often obtained for tropospheric species,either because their abundances are small or because it is difficult to separate the tropospheric from the stratospheric signals. In many cases, the validation measurements themselves are also not as accurate and precise for these small signals as one would like, adding the uncertainty of the validation data to that of the satellite measurement.

We mention this paper from Richter et al. just before, L.42.:

Richter et al. (2014) have discussed the challenges associated with the validation of tropospheric reactive gases.

So we consider we have given the reader the useful reference to have more information on this aspect.

• Line 114: what are the European thresholds? Add a reference and values.
The reference (EEA,2019) is already given in the sentence right after, which also gives the typical value for yearly NO2 in the center of Bucharest, we added the EU limit value for yearly NO2 to compare with.

"For instance, the annual mean concentration of NO2 at the traffic stations was about 57 ug.m-3 in 2017 (EEA,2019), when the EU limit is 40 ug.m-3. "

• Table 1: change GEMS to launched instead of planned.
Corrected.

• Throughout the paper: Spatial resolution is in km and not km2. For example, 7 x 7km2 is not the same as 7km x 7km.

We do not agree with this comment, we think width x height in km2 is clear enough and shorter, thus better. This way of writing is largely used in other publications including by the TROPOMI science team. e.g. recently:

van Geffen, J., Boersma, K. F., Eskes, H., Sneep, M., ter Linden, M., Zara, M., and Veefkind, J. P.: S5P TROPOMI NO_2 slant column retrieval: method, stability, uncertainties and comparisons with OMI, Atmos. Meas. Tech., 13, 1315–1335, https://doi.org/10.5194/amt-13-1315-2020, 2020.

• Line 261: delete the word ‘those’

Here we think our sentence is clearer as it is since we do not show all the sonde measurements, neither a random selection of them, but only the ones which detected the plume. Grammatically, we think it is also fine to use “those” … “which” since the Nobel prize in literature Bertrand Russell used this structure in his essay ‘Our knowledge of the external world’ (1914)

Things are those series of aspects which obey the laws of physics.

• Line 396: change ‘As for’ to ‘Similar to’
Corrected.

• Line 399: Start a new paragraph with the sentence starting with ‘On the other hand’

Corrected.