Dear Ass. Editor,

Thank you for your time and effort to go through our revised manuscript. Herein we will try to answer to your comments.

"One important reason is that you have changed the data significantly with respect to the previous (reviewed) version. If data are changed significantly, the paper needs another round of review."

Thank you for pointing this issue, but in fact we did not change the dataset that was used for this validation work. We still use the same TROPOMI/S5P TCWV data and the same Level-2 data from AERONET stations as in the initial version of the paper, but we modified the comparison methodology, following well-justified and valuable suggestions from the reviewers. Namely, instead of comparing all CIMEL observations within ±30' from the satellite overpass time, in the revised version we considered only the nearest in time (within a ±30' time interval) ground-based observations. This change led to similar results concerning the relative bias but reduced the dispersion of the comparisons.

"Second, although your statistical analysis appears sound, no attempt is made to explain the results phenomenologically. For example, the patterns in Fig. 8 are not explained using meteorological or observational variability. In fact, in the description of Fig. 9 the latitudinal dependence of the relative difference is deemed non-existent ("no clear pattern in the dependency of the relative differences on latitude").

Validation studies can be a multi-phase activity, involving the interaction between the data producers and the validation team. It is common practice, as a first step, to use the comparisons to a dataset characterized as reference to investigate and identify issues related to the algorithm and observing geometry of the satellite product. This step uses global databases. Then at a second step, when certain instrumental issues are excluded, attempts are made to understand the origin of remaining differences, in collaboration with the algorithm teams. Indeed, to that respect, case studies and association of the observed differences with the variability of geophysical variables that affect the retrievals are extremely useful, but these will require more analysis which could be part of a follow up paper.

To explain better the patterns seen in Fig. 8, we changed the manuscript. Please see our reply to your comment #14, below. For the discussion on Fig. 9, please see our answers to your comments #17 and #18, below.

"Moreover, whereas the other TROPOMI TCWV retrieval validation papers are now referenced and the results compared with those from your own validation, similarities and differences between the algorithms are not discussed. These would be of use to understand why validation results agree - or not."

Concerning the differences between Chan et al. (2022) and the other two retrieval algorithms for TROPOMI TCWV products (Borger et al., 2020; Schneider et al., 2020

and 2022), they are now added to the manuscript (Section 1), as suggested. See also our answer to comment #2, below.

"Third, there are a lot of qualitative statements ("the performance of the TROPOMI/S5P TCWV retrieval algorithm (...) is currently adequate but could be improved further in the future."; "is a temporally stable product of high quality and precision..."). The findings of a validation study should be given in quantitative terms, e.g., "The product shows a stability of X%/decade, the uncertainty range is Y, and the mean bias Z; this is (or is not) in agreement with requirements / in agreement with previous studies."."

In such an extensive article, the use of qualitative terminology is indeed unavoidable at points during the discussion. We share the editor's view that numerical findings should be clearly stated in a paper, which is why we opted to include numerical results both in the discussion of the figures and statistical results as well as a bullet-point presentation in the conclusions.

According to the co-authors' experience in the field of satellite products' validation and the respective literature, apart from quantitative information, general statements are useful for a potential user of the satellite data to put these quantitative statements in perspective. Therefore, we indeed include a few such statements like the two that you point out. Specifically, the second sentence ("*is a temporally stable product of high quality and precision...*") is part of the paper's concluding remarks and follows one full page of conclusions (Section 5) which is rich in numbers and percentages.

"And fourth, the arguments in this manuscript are often not properly backed by evidence. Example: "The monthly mean relative bias (...) depends strongly on the ground-based instrument's operation and maintenance". I doubt it - but if it is true, there should be a reference to literature or other reliable source."

Specifically for this sentence we give extended answers in your comments #11 and #15, below.

After taking under consideration your valued questions, comments and suggestions, our manuscript was revised again, and a new version is submitted.

We would like to sincerely thank you for your help in the process of improving our manuscript and your constructive comments.

The authors

Replies to specific comments on the manuscript:

(Please note that the numbering of the lines seen below corresponds to the manuscript version that you used to provide your comments/questions and not the updated version of the manuscript that we submit after your comments.)

1. Abstract, line 28: "This is quite a significant change wrt the old version"

   Reply: Indeed, the mean relative bias was decreased by 1.3% with respect to the previous version, still within the 1-sigma statistical variability. This decrease is due to the change in the co-location methodology that was applied after the implementation of the reviewers' suggestions.

2. Introduction, lines 68-81: "Please describe the differences between Chan, Borger, and the other TROPOMI algorithm so that we can relate those to the agreements in validation results"

   Reply: The following sentences will be added at the end of this paragraph (line 75):
   "The methods of Borger et al. (2020) and Chan et al. (2022), are similar in principal but they differ in some important aspects such as: (i) Chan et al. (2022) fit for the 435-455nm spectral range, while Borger et al. (2020) use a slightly different wavelength range, 430-450nm; (ii) for the AMF calculation, the algorithm of Borger et al. (2020) assumes an exponential decay profile with empirical parameterization of the water vapor scale height, while Chan et al. (2020) use an a-priori profile from the statistical analysis of historical data and they dynamically pick the most appropriate one; (iii) for the surface albedo parameter, Chan et al. (2020) use the TROPOMI/S5P GE_LER which is derived at the same spectral fitting range (435-455nm), while Borger et al. (2020) use the OMI surface albedo retrieved at 442nm. The comparison of the two surface albedo products is extensively discussed in Chan et al. (2022)."

   The Schneider et al. algorithm is using a different spectral range, which is already mentioned in the manuscript (lines 76-77). The following sentence will be added after the respective paragraph (line 81):
   "Compared to Chan et al. (2022), the Schneider et al. (2020, 2022) algorithm employed a completely different technique and due to the differences in the spectral range of the measurement, the final water vapor product has a different vertical sensitivity."

3. Introduction, line 80: "Why did you make a change here? This is not a new result, right?"

   Reply: Actually, no, this is not a new result. The percentages given in the previous version of the manuscript referred to mid-latitude stations only. I deleted the phrase "for mid-latitude stations" and changed the percentages according to Schneider et al. (2020) to give a more general overview of the product's bias.

4. Section 2.1, line 126: "Please keep units consistent!"

   Reply: These are the units that Vaquero-Martinez et al. (2022) use in their work. But we will, of course, turn them into kg/m^2.

5. Section 2.2, line 154: "And should therefore not affect your results, since you are comparing clear-sky data only, right?"

Reply: Cimel will measure TCWV when the solar disk is clear, but this does not mean that there are no clouds within a radius of 10 km around the station. This is why the cloud fraction of the satellite measurements (Figure 11) goes up to 0.5. Therefore, we are not comparing only clear-sky data.

6. Section 2.2, line 165: "So which is it??? Can you combine all these numbers into a single number?"

   Reply: The uncertainty of the AERONET v3 algorithm is not assessed by any of these papers, but as it is written in line 166, it is expected to be better than 10%, which is the uncertainty of the v2 algorithm (line 165). As for the AERONET dry bias, it is summarized in the phrase: "-5 to -10% depending on the study and its reference" (line 458, Section 5). The following sentence will be added in this paragraph (line 162, before Campanelli et al.), as well:
   "An approximate mean bias for the AERONET TCWV product that results from all these studies, that are based on various stations, temporal coverages and reference measurements, is -5 to -10 %."

7. Section 2.2, line 171: "...use of a single retrieval algorithm..." That, and the consistency in instruments and calibration!

   Reply: Thank you for the suggestion, this sentence will be rephrased to:
   "This fact, in addition to the use of a single standardized retrieval algorithm and the consistency in instruments calibration, are strong advantages in favor of using the AERONET for this validation work."

8. Section 2.2, lines 180-183: "This is neither transparent, nor reproducible. You need to put quantitative information here. How many months of data were required, etc."

   Reply: The phrase will be changed to:
   "An in-house quality control based on the visual and statistical analysis of the available datasets per station, ensured that only stations with data that cover fully the time period of our study, or cover at least 20 out of the 32 months of the TROPOMI/S5P dataset, are contributing to the ground-based reference dataset."

9. Section 3, line 191: "Not a very scientific argument. Better would be: considering the scale of spatial variability of TCWV, the value detected at 10 km distance is expected to be very close."

   Reply: Thank you for the comment. In this sentence, we are only illustrating that we used the same criteria for the co-locations as other studies did. We think that what you suggest is already stated in the previous sentence.

10. Section 3, line 198: "I think you mean continuous. You are using Level-2 data, which is not at all instantaneous."

    Reply: Actually, no, the term "instantaneous" does not mean it is near-real-time (which, of course, Level-2 data are not). As we explain further in the sentence, we mean that these are not hourly or daily mean values. The record used here consists of all individual measurements performed every 15' during each day. To further clarify this issue, we will also add this information in line 177, where the term in traduced for the first time:

"The data files retrieved from AERONET are available in ASCII format in daily, monthly or instantaneous (i.e. measurements performed every 15') temporal analysis"

11. Section 3, line 216: "You argued before that the instruments are very similar (as a reason to use AERONET), so this cannot be right. Also, I wonder how you come to this conclusion."

    Reply: Thank you for the comment. You are right, this phrase does not communicate correctly the message that we had in mind. The sentence will be rephrased to:
    "The monthly mean relative biases per station (panels a) for the example stations shown here are within ±0.2 %, demonstrating the good agreement between satellite and ground-based observations, as well as a good temporal stability of both sources of measurement for the available dataset spanning 2.5 years. The variability of the biases, depicted as error bars, may be due to both the ground-based and space-born instrument observational accuracy as well as algorithm and/or meteorology-related effects."

12. Figure 3: "The figure has changed quite a bit"

    Reply: Yes, it has changed due to the different co-location methodology that was applied following the suggestions of the reviewers.

13. Section 3, lines 235 – 237: "And no TROPOMI data..."

    Reply: TROPOMI/S5P data are available in the South Pole. We did have co-locations for a two-month period during local summer of 2018-2019 (mid-November until mid-January), but it was decided to not use them in the analysis, due to their lack of representativeness.

14. Figure 8: "Are you going to explain these patterns?"

    Reply: The discussion on Fig. 8 was further developed and the patterns were linked to the latitudinal and temporal changes in TCWV content seen in Fig. 1 and Fig. 9. The final form of the paragraph is quoted after your comment # 18, below.

15. Section 4.1, line 346: "What does this mean? How do you know? Literature reference?"

    Reply: This was a wrong choice of words and thank you for noticing. The sentence will be rephrased as follows:
    "…since it is well known that some ground-based stations may overestimate or underestimate their observational constituent systematically due to the meteorological conditions occurring at the station site. Moreover, when such a station does not provide a continuous record of observations, there is a high possibility that it will introduce an artificial and non-representative bias to the validation. Most of these stations, that did not fully cover the time period of our study, were filtered out of the ground-truth database used in this work."

16. Section 4.1, line 350: "To me, co-location statistics are the number of overpasses, or similar. I don't think that's what's meant here."

    Reply: Thank you for the comment. The phrase will be changed as follows:
    "... the respective stations were considered with the remark that the statistics resulting from their co-locations should be interpreted with caution."

17. Section 4.1, line 351: "Yes there is - see Fig. 8"

Reply: Of course, you are right! This phrase ("Nevertheless, as shown above, ... overestimation close to the poles.") will be deleted. Moreover, since the dependency of the percentage differences on latitude is already discussed in the previous paragraph, discussing Fig. 9, the next two sentences ("Considering the uncertainties … in the future.") will be moved to the end of the previous paragraph. The final form of the paragraph is quoted after comment #18, below.

18. Section 4.1, lines 354-356: "I do not know what this conclusion is based on."

Reply: This conclusion is based on the previous sentence, that compares the measurement uncertainties to the mean biases per latitude bin showed in Fig. 9. To make it clearer and to also introduce the dependence of the product on surface albedo, the sentence will be changed as follows:
"The performance of the TROPOMI/S5P TCWV retrieval algorithm with respect to the surface albedo parameter which significantly changes with latitude is currently adequate but could be further improved in the future, as is also shown further on in this work (Figure 12, panel b)."

After all the changes that are introduced as replies to your comments/questions #14-18 above, the paragraphs discussing Figures 8 and 9, will be reformatted as follows:

The contour plots in Fig.8 show the mean relative percentage differences (panel a) and the respective standard deviations (panel b) of the satellite and ground-based co-locations, with respect to latitude and season. Panel a shows that for the mid-latitude winter months of each hemisphere, when the water vapor content of the atmosphere is below ~ 20 kg/m^2 (see Fig. 1, panel a), the mean relative bias for the respective stations is positive, between 0 and +15 or +20 % (January). During the summer months, when the highest values of water vapor occur for the mid-latitudes, up to 40 kg/m^2 (see Fig. 1, panel c), the mean relative bias is within ± 5 %. In the tropics, where the TCWV content is higher throughout the year (40 – 80 kg/m^2), the mean relative bias is constantly negative, ranging between -5 and -20 %. Panel b depicts the strong seasonality of the comparisons' standard deviations, i.e. of their variability, which is high during the mid-latitudes winter months of both hemispheres (up to ~ 90 %), and lower (10-30 %) during summer, when the number of ground-based AERONET measurements (i.e. the number of co-locations) and their accuracy is much higher (Fragkos et al., 2019). Additionally, as shown in Fig. 1, for latitudes higher than the tropics of both hemispheres the water vapor content is lower and has a stronger temporal variability, explaining the higher standard deviations of the relative differences. It is also interesting to see that the variability of the comparisons for the tropics (Fig. 8, 15° N-15° S) is much lower (up to 20 %) compared to the other latitude belts, showing that the negative mean relative bias of our comparisons is temporally invariable in this part of the globe, where the water vapor content is higher than ~ 40 kg/m^2.
To further investigate the latitudinal patterns of our comparisons, the mean relative percentage differences per station with available ground-based data are averaged in 10° latitude belts and are shown versus latitude in Fig. 9, panel (a). The same is also shown in panel (b), but the averaged parameter per latitude bin is the difference between satellite and

ground-based observations in kg/m^2. The overall mean relative percentage bias for the latitudinal dependency is -1.1 ± 6.1 % and has a mean standard error of 6.8 %. The agreement between satellite and ground-based observations remains within ± 10 % for individual belts of the NH and the belt northwards 50° of the SH. The latitude bins -30° S to 30° N form a U-shaped curve, showing that the satellite instrument reports lower TCWV up to ~10 % with respect to ground-based observations close to the equator and reaching ~0 % at ±30°. This result, which corresponds to a difference between satellite and ground-based observations up to -4 kg/m^2 (panel b), is in agreement with Chan et al. (2022), where the dry bias is attributed to albedo effects in the visible band over vegetation and to the presence of aerosol and/or clouds in the measurement field. For the NH high latitude stations, above 70° N, the discrepancy becomes positive up to 10 % and has a very large standard error due to the limited number of co-locations (panel a). In terms of difference (panel b), this percentage accounts for a small overestimation of less than 1 kg/m^2 by TROPOMI/S5P occurring close to the poles where the amount of water vapor is less than 20 kg/m^2 (see Fig. 1). Considering that the uncertainties of both types of measurement is ~10 %, the comparison of the satellite and ground-based observations is regarded as satisfactory. The performance of the TROPOMI/S5P TCWV retrieval algorithm with respect to the surface albedo parameter which significantly changes with latitude is currently adequate but could be further improved in the future, as is also shown further on in this work (Fig. 12, panel b).

The statistics per latitude belt in terms of mean difference (satellite – ground) in kg/m^2, mean relative bias ± standard deviation and mean standard error of the comparisons (in %), are shown in Table 2 for 15° latitude belts up to ± 30° N and S. The belts 0 to 15° N and 0 to 15° S represent the tropics. Above 30°, the binning is doubled because due to the low water vapor content and its low variability, the differences in the statistics between belts 60°N-75°N and 75°-90°N would be negligible.

It is worth noting that the mean relative bias of each latitude bin and the respective mean standard deviation, thus the variability (not shown in Fig. 9), should not be attributed to the satellite product only, since it is well known that some ground-based stations may overestimate or underestimate their observational constituent systematically due to the meteorological conditions occurring at the station site. Moreover, when such a station does not provide a continuous record of observations, there is a high possibility that it will introduce an artificial and non-representative bias to the validation. Most of these stations, that did not fully cover the time period of our study, were filtered out of the ground-truth database used in this work. Nevertheless, for some latitude bins, like 40° S to 50° S, where the station density or the temporal coverage is low, the respective stations were considered with the remark that the statistics resulting from their co-locations should be interpreted with caution.

---

19. Section 4.2, line 364: "Assorted variables???"

    Reply: The reviewers suggested that the term "variables" or "parameters" is used instead of "influence quantities". The phrase will be changed to "different variables".

20. Section 4.2, line 366: "But the AMF depends on cloud and surface information?!"

Reply: You are correct that the AMF depends on this input information, however it also depends on numerous other variables. In validation studies it is hence always examined on its own accord as well.

21. Figure 10: "Figures changed"

Reply: Panel (a): the only change here is the different color of some data-points to better indicate the solar zenith angle bins with a limited number of co-locations. This information is given in the 1st paragraph of Section 4.2.
Panel (b): Following the suggestion of one of the reviewers, the viewing zenith angle dependency plot was replaced with the one showing the dependency on the satellite pixel.

22. Section 4.2.2, line 399: "What instrument is this?"

Reply: The phrase will be changed to "that validated their TCWV product against Special Sensor Microwave Imager Sounder (SSMIS) measurements on board f16 and f17,..."

23. Section 4.2.2, line 403: "How different?"

Reply: The differences between the two algorithms will be added in Section 1 (see our answer in your respective question #2, above).

24. Section 4.2.2, line 405: "How is cloud albedo defined?"

Reply: According to https://glossary.ametsoc.org/wiki/Cloud_albedo
"(Cloud albedo is) The fraction of solar radiation reflected directly by clouds in the atmosphere."
The reference will be added to the manuscript, even though it is a rather common term.

25. Section 4.2.2, lines 412-414: "If your validation says it's OK, why not use them?" and "I don't understand this."

Reply: As we state in the previous sentences, the co-locations with cloud fraction ranging between 0.3 and 0.5 do not agree well with the ground-based data, having mean relative biases of +20 to 60 %. Nevertheless, they are very few in population so they don't change our validation results, since their contribution is not important. We think that the suggestion to not use these data for scientific studies of the TCWV is within the purposes of a validation paper.
The sentence will be changed to:
"Filtering-out the co-locations dataset for cloud fraction over 0.3 was also investigated, resulting to no major differences in the overall validation results due to the limited contribution of the relatively low number of co-locations with cloud fraction values in the range 0.3 – 0.5. Nevertheless, it is advisable to not use this small portion of TCWV data for future scientific studies"

26. Section 4.2.2, lines 427-428: "This sounds a bit like "They don't look too good, so we left them out". Either you show them, or you give a scientific argument why not (too few points?)"

Reply: The co-locations with surface albedo over 0.2 were left out of the new figure because this was a suggestion of one of the reviewers. She/He proposed to limit the x-axis of the figure up to 0.2 and make the binning of the co-locations with respect to surface albedo finer. We also mention in the same sentence that "very few co-locations have surface albedo above 0.2".

27. Figure 12: "How come these figures changed so much?"

Reply:
Panel (a): The overall number of co-locations changed. Therefore the data-points showing co-locations with surface pressure below 950 hPa, which were few in number in the previous version, became even fewer and changed the statistics of the respective averaging bins.
Panel (b): The above is valid here, too. Additionally, in this figure the x-axis range was changed as we explained in the previous comment (#26).

28. Competing interests: "What a strange statement. I mean, can't you just say "D.L. is a member of the editorial board"?"

Reply: According to the AMT instructions for this field:
"*If some authors are members of the editorial board of the journal, a sentence should be included for the sake of transparency: "Some authors are members of the editorial board of journal X. The peer-review process was guided by an independent editor, and the authors have also no other competing interests to declare.".*
We will also add the second part of the statement in our manuscript.