

Review: TROPOMI/S5P formaldehyde validation using an extensive network of ground-based FTIR stations by Vigouroux et al.

General comments

Vigouroux and co-authors present extensive validation of TROPOMI's formaldehyde retrievals (version 1.1.[5-7]) against ground-based FTIR retrievals from 25 stations around the world. Most of these stations belong to the Network for the Detection of Atmospheric Composition Change (NDACC). Their results indicate that TROPOMI satisfies pre-launch precision and accuracy requirements. TROPOMI overestimates HCHO columns (~26%) over locations with small HCHO levels while underestimates HCHO columns (~30%) over locations with high HCHO levels.

This paper provides an excellent example of careful and sound satellite validation using ground-based remote sensing observations. It provides a detailed description of the methods and datasets used. It is well written and provides clear descriptions of the most important results. The paper should be published with minimal changes since it provides a compelling case supporting the quality and capacity of S5P HCHO retrievals, its current biases and what users should expect to achieve with S5P.

Some minor questions are raised. The aim is to further improve the clarity of the text and the description of the methodology and results.

Specific comments

Abstract.

Page 2, line 4: "We observe that, at all sites, the TROPOMI accuracy is below the upper limit of the pre-launch requirements of 80%, and below the lower limit of 40% for 20 of the 25 stations." This sentence is confusing. What are the pre-launch requirements? If HCHO retrievals accuracy are below lower limit of 40% there are also below the upper limit of 80% why both are mentioned?

Introduction.

Page 3, line 2: Validation from aircraft has been expanded to multiple locations by Zhu et al., 2020 (<https://www.atmos-chem-phys-discuss.net/acp-2019-1117/>). Could we value to add it to the list of aircraft based validation efforts?

TROPOMI HCHO data.

The description of TROPOMI data and versions is very complete but after reading this section the question remains, of all the options (RPRO, OFFL and NRTI) which one has

been used? If several depending on the station and the period of time, that should also be explained?

Given the unprecedented TROPOMI spatial resolution, the surface elevation could play a bigger role while explaining biases for some locations with complicated topography. What is the source of TROPOMI surface elevation information?

Page 4, line 6. "All cross-sections have been pre-convolved", these cross-section include HCHO and interferers but that may be not clear to someone without a background on DOAS retrievals. Maybe worth explaining? How stable have been TROPOMI slit functions after launch? Is the algorithm correcting cross-sections for changes in the slit function?

Page 4, line 20. How is M_0 calculated? Is it the average of the AMFs of the slant columns considered in the calculation of $N_{(s,0)}$?

Ground-based FTIR HCHO data

Figure 1 caption could be expanded to provide some information about the spatial resolution of the averaged TROPOMI data shown. What kind of averaging algorithm was used to generate the background data?

Page 7, line 22: Maybe adding described by to "is 13% in the network of Vigouroux et al., (2018)" could be more precise "is 13% in the network described by Vigouroux et al., (2018)"

Page 7, line 25: Please clarify, it looks like if stations using the PROFFIT9 retrieval code can have bigger systematic uncertainty due to uncertainty on the channeling that is not taken into account yet in the SFIT4 code. If the SFIT4 code is not taking this channeling uncertainty in the budget it just means that is introducing a systematic error for those stations?

Page 8, line 3: Why the smoothing systematic uncertainty (on the total column) is significantly bigger for the 5 added sites?

Validation method

Collocation criteria

What is the effect of reducing/increasing the TROPOMI/FTIR collocation radius (currently set at 20km)? Is there a radius threshold/range where no improvement is achieved in the comparisons?

For each station, after co-adding, what is the median TROPOMI detection limit and random uncertainty? That will be an interesting fact to know

Building inter-comparable products

Equation 2 could have dimensions problem: a_s SP5 averaging Kernel is defined on the S5P vertical grid according to line 16 page 9 while x'_F and $x_{S,a}$ are defined on the FTIR vertical grid.

Validation results

As mentioned above, including a table showing the period of time each one of the products (RPRO, OFFL) has been used in the calculations will assure full reproducibility of the results shown.

TROPOMI observed BIAS and accuracy

Page 12, line 10: This sentence is confusing "... it is negative for higher levels and very consistent for the stations from 8.7 to 28.6 x 10¹⁵..." This is my interpretation "... it is negative and very consistent for stations with higher levels, ranging from 8.7 to 28.6 x 10¹⁵..." but maybe is the HCHO level what is 8.7 to 28.6 x 10¹⁵.

Page 12, line 10: Lower levels are defined in the abstract and below at page 12, line 21 as 2.5x10¹⁵ molec/cm². What is the meaning of 6.5x10¹⁵ molec/cm².

Do the authors suggestions on how to link/explain the constant and proportional biases to different instrumental, algorithm, or geophysical parameters?