TROPOMI/S5P formaldehyde validation using an extensive network of ground-based FTIR stations

Corinne Vigouroux[1], Bavo Langerock[1], Carlos Augusto Bauer Aquino[2], Thomas Blumenstock[3], Zhibin Cheng[4], Martine De Mazière[1], Isabelle De Smedt[1], Michel Grutter[5], James W. Hannigan[6], Nicholas Jones[7], Rigel Kivi[8], Diego Loyola[4], Erik Lutsch[9], Emmanuel Mahieu[10], Maria Makarova[11], Jean-Marc Metzger[12], Isamu Morino[13], Isao Murata[14], Tomoo Nagahama[15], Justus Notholt[16], Ivan Ortega[6], Mathias Palm[16], Gaia Pinardi[1], Amelie Röhling[3], Dan Smale[17], Wolfgang Stremme[5], Kim Strong[9], Ralf Sussmann[18], Yao Té[19], Michel van Roozendael[1], Pucai Wang[20], and Holger Winkler[16]

[1]Royal Belgian Institute for Space Aeronomy (BIRA-IASB), Brussels, Belgium
[2]Instituto Federal de Educaçao, Ciência e Tecnologia de Rondônia (IFRO), Porto Velho, Brazil
[3]Karlsruhe Institute of Technology (KIT), Institute for Meteorology and Climate Research (IMK-ASF), Karlsruhe, Germany
[4]German Aerospace Centre (DLR), Remote Sensing Technology Institute, Oberpfaffenhofen, 82234 Weßling, Germany
[5]Centro de Ciencias de la Atmósfera, Universidad Nacional Autónoma de México (UNAM), Mexico City, México
[6]Atmospheric Chemistry, Observations & Modeling, National Center for Atmospheric Research (NCAR), Boulder, CO, USA
[7]Centre for Atmospheric Chemistry, University of Wollongong, Wollongong, Australia
[8]Finnish Meteorological Institute (FMI), Sodankylä, Finland
[9]Department of Physics, University of Toronto, Toronto, Canada
[10]Institut d'Astrophysique et de Géophysique, Université de Liège, Liège, Belgium
[11]Saint Petersburg State University, Atmospheric Physics Department, St. Petersburg, Russia
[12]Observatoire des Sciences de l'Univers Réunion (OSU-R), UMS 3365, Université de la Réunion, Saint-Denis, France
[13]National Institute for Environmental Studies (NIES), Tsukuba, Ibaraki 305-8506, Japan
[14]Graduate School of Environment Studies, Tohoku University, Sendai 980-8578, Japan
[15]Institute for Space-Earth Environmental Research (ISEE), Nagoya University, Nagoya, Japan
[16]Institute of Environmental Physics, University of Bremen, Bremen, Germany
[17]National Institute of Water and Atmospheric Research Ltd (NIWA), Lauder, New Zealand
[18]Karlsruhe Institute of Technology, IMK-IFU, Garmisch-Partenkirchen, Germany
[19]LERMA-IPSL, Sorbonne Université, CNRS, Observatoire de Paris, PSL Université, 75005 Paris, France
[20]Institute of Atmospheric Physics, Chinese Academy of Sciences (CAS), Beijing, China

Correspondence to: C. Vigouroux
(corinne.vigouroux@aeronomie.be)

Abstract. TROPOMI (the TROPOspheric Monitoring Instrument), on-board the Sentinel-5 Precursor satellite, has been monitoring the Earth's atmosphere since October 2017, with an unprecedented horizontal resolution (initially 7x3.5 km^2, upgraded to 5.5x3.5 km^2 since August 2019). Monitoring air quality is one of the main objectives of TROPOMI, with the measurements of important pollutants such as nitrogen dioxide, carbon monoxide, and formaldehyde (HCHO). In this paper we assess the quality of the latest HCHO TROPOMI products (version 1.1.[5-7]), using ground-based solar-absorption FTIR (Fourier Transform Infrared) measurements of HCHO from twenty-five stations around the world, including high, mid, and low latitude sites. Most of these stations are part of the Network for the Detection of Atmospheric Composition Change (NDACC), and they provide a wide range of observation conditions from very clean remote sites to those with high HCHO levels from anthropogenic or biogenic emissions. The ground-based HCHO retrieval settings have been optimized and harmonized at all the stations, ensuring a consistent validation among the sites.

In this validation work, we first assess the accuracy of TROPOMI HCHO tropospheric columns, using the median of the relative differences between TROPOMI and FTIR ground-based data (BIAS). The pre-launch requirements of the TROPOMI HCHO accuracy are 40-80%. We observe that these requirements are well reached, with the BIAS found below 80% at all the sites, and below 40% at 20 of the 25 sites. The provided TROPOMI systematic uncertainties are well in agreement with the observed biases at most of the stations, except for the highest HCHO levels site where it is found to be underestimated. We find that, while the BIAS has no latitudinal dependence, it is dependent on the HCHO concentration levels: an overestimation (+26±5%) of TROPOMI is observed for very small HCHO levels ($<2.5\times10^{15}$ molec/cm^2), while an underestimation (-30.8%±1.4%) is found for high HCHO levels ($>8.0\times10^{15}$ molec/cm^2). This demonstrates the great value of such a harmonized network covering a wide range of concentration levels, the sites with high HCHO concentrations being crucial for the determination of the satellite bias at the regions of emissions, and the clean sites allowing a small TROPOMI offset to be determined. The wide range of sampled HCHO levels within the network allows the robust determination of the significant constant and proportional TROPOMI HCHO biases (TROPOMI=+ 1.10 (±0.05) $\times10^{15}$+ 0.64 (±0.03) × FTIR, in molec/cm^2).

Second, the precision of TROPOMI HCHO data is estimated by the median absolute deviation (MAD) of the relative differences between TROPOMI and FTIR ground-based data. The clean sites are especially useful to minimize a possible additional collocation error. The precision requirement of 1.2 $\times10^{16}$ molec/cm^2 for a single pixel is reached at most of the clean sites, where it is found that the TROPOMI precision can even be twice better (0.5-0.8 $\times10^{15}$ molec/cm^2 for a single pixel). However, we find that the provided TROPOMI random uncertainties may be underestimated by a factor of 1.6 (for clean sites) to 2.3 (for high HCHO levels). The correlation is very good between TROPOMI and FTIR data (R=0.88 for 3 hours-mean coincidences; R=0.91 for monthly means coincidences). Using about 17 months of data (from May 2018 to September 2019), we show that the TROPOMI seasonal variability is in very good agreement at all of the FTIR sites.

The FTIR network demonstrates the very good quality of the TROPOMI HCHO products which is well within the pre-launch requirements for both accuracy and precision. This paper advises for a refinement of the TROPOMI random uncertainty budget and of the TROPOMI quality assurance values for a better filtering of the remaining outliers.

1 Introduction

TROPOMI (the TROPOspheric Monitoring Instrument), on-board the Sentinel-5 Precursor (S5P) satellite, has been monitoring the column amounts of atmospheric constituents since October 2017, at the unprecedented horizontal resolution of 7x3.5 km^2, upgraded to 5.5x3.5 km^2 since August 2019. This huge amount of data, delivered to the public and the scientific community, represents a big step to improve our knowledge of chemical and dynamical processes in the atmosphere. It is crucial to validate the quality of these new satellite data to trust and benefit their scientific exploitation. This paper focuses on the first quality assessment of the latest publicly available TROPOMI HCHO data products (v.1.1.[5-7]).

In the past, the HCHO satellite products have been validated at a few locations only, mainly using aircraft (Martin et al., 2004; Zhu et al., 2016, 2020), MAX-DOAS (Multi-AXis Differential Optical Absorption Spectroscopy) technique over land (Wittrock et al., 2006; De Smedt et al., 2015) or ship-based (Peters et al., 2012; Tan et al., 2018) and FTIR (Fourier Transform Infra-Red) technique (Jones et al., 2009; Vigouroux et al., 2009; De Smedt et al., 2015). However, given the high spatial heterogeneity of HCHO concentrations due to its short lifetime (a few hours), there is a crucial need for a more extended world coverage to assess unambiguously the satellites' achieved accuracy and precision. Furthermore, increasing the number of ground-based locations is not sufficient: it is also important to harmonize the reference data obtained at all the stations, in order to facilitate the interpretation of the satellite validation by minimizing the site-to-site biases. In this view, and in particular in the framework of the TROPOMI Calibration and Validation (Cal/Val) activities, we have developed HCHO retrieval settings that are suitable for any ground-based FTIR site, which have been consistently applied in Vigouroux et al. (2018) at twenty-one FTIR stations, most of them affiliated with NDACC (Network for the Detection of Atmospheric Composition Change). Vigouroux et al. (2018) described in detail the retrieval settings and the HCHO harmonized time-series obtained at these stations which cover a large range of HCHO concentrations, from very clean Arctic and oceanic sites to high HCHO levels sites, such as polluted cities (e.g. Paris or Mexico City) and sites close to large biogenic emissions like the Amazon basin (Porto Velho).

This paper presents the validation of the TROPOMI HCHO product (v.1.1.[5-7]) using an updated network of twenty-five ground-based FTIR stations. In the first section, the TROPOMI HCHO data are introduced with their uncertainty budget and their quality flag criteria. The second section describes the ground-based FTIR HCHO network and the characterization of these reference data (uncertainties and averaging kernels). Then, the validation procedure (collocation criteria, smoothing technique, definition of the quantities to be used in the quality assessment) is explained in Sect. 4. Finally, Sect. 5 shows the validation results using comparisons between TROPOMI and FTIR ground-based network data, leading to an assessment of the TROPOMI HCHO accuracy and precision, and the observed TROPOMI bias.

2 TROPOMI HCHO data

TROPOMI, on the S5P platform, is in a low-Earth afternoon polar orbit with a swath of 2600 km resulting in a daily global coverage (Veefkind et al., 2012). Operational Level 2 (L2) products include vertical columns of O_3, SO_2, NO_2, HCHO, CO and CH_4, as well as O_3 profile, aerosol layer height, cloud information and aerosol index. The spatial resolution, originally of 3.5x7 km^2 has been increased to 3.5x5.5 km^2 on 6 August 2019.

The prototype algorithm of the formaldehyde product is being developed at the Royal Belgian Institute for Space Aeronomy (BIRA-IASB) and the corresponding operational processor is being developed at the Remote Sensing Technology Institute (IMF) of the German Aerospace Center (DLR). The product has been declared operational and released to the public at the end of 2018. At the time of writing this paper, the latest product versions 1.1.[5-7] provide a consistent time series of Reprocessed+Offline (RPRO+OFFL) data, covering the period between May 2018 up to (at least) December 2019 (last access). The detailed validation results shown in Sect. 5 are obtained using this consistent time-series (RPRO+OFFL, from 2018-05-14 to 2019-12-31). The version numbers and their dates of change are given in Table 1, and further details are given in the Readme file [1]. The Near-Real-Time (NRTI) product, for the same versions 1.1.[5-7], started in December 2018 up to December 2019 (last access). This product has also been validated, but the results being very similar to the RPRO+OFFL validation, we do not show them in details in this paper.

Table 1. TROPOMI RPRO+OFFL complete time-series (versions 1.1.[5-7]) used in the present work.

Date	Processor version	Relevant improvements (see Readme file [1]).
2018-05-14 to 2018-11-28	RPRO v.1.1.5	Alignment of the configuration for NRTI, OFFL and RPRO chains regarding the Chemistry
2018-11-28 to 2019-03-28	OFFL v.1.1.5	Transport Model input, leading to the same product quality.
2019-03-28 to 2019-04-23	OFFL v.1.1.6	- Surface classification climatology updated - Fixed a bug in the interpolation of the surface albedo climatology - Fixed a problem regarding the retrieved CLOUD product parameters being too close to the a-priori values. This might have affected the calculation of the HCHO in cloudy cases.
2019-04-23 to 2019-12-31	OFFL v.1.1.7	No changes (for HCHO) with respect to previous version.

The S5P HCHO retrieval algorithm is based on the DOAS method, and is directly inherited from the OMI QA4ECV product retrieval algorithm (https://doi.org/10.18758/71021031). It consists in a 3-step method (slant column retrieval, air mass factor calculation, and conversion to tropospheric column), fully described in De Smedt et al. (2018). The retrieval of the slant columns (N_s) is performed in the UV part of the spectra (in TROPOMI channel 3), in a fitting interval of 328.5-359 nm. The HCHO cross-section is from Meller and Moortgat (2000). Together with the HCHO cross-section, the absorptions of NO_2, BrO, O_3 (at two temperatures) and O_4 are fitted. A Ring cross-section and two pseudo-cross sections to account for non-linear O_3 absorption effects are also included in the fit. References are given in De Smedt et al. (2018). All cross-sections have been pre-convolved for every row separately with an instrumental slit function adjusted just after launch. The DOAS reference spectrum is updated daily with an average of Earth radiances selected in the Equatorial Pacific region on the previous day. The result of the fit is therefore a differential slant column, showing increases over continental sources compared to the remote background. The conversion from slant to tropospheric columns (N_v) is performed using a look up table of vertically resolved air mass factors (M) calculated at 340 nm with the radiative transfer model VLIDORT v2.6 (Spurr, 2008). Parameters for each ground pixel are the observation geometry, the surface elevation and reflectivity, including the clouds (that are treated as reflecting surfaces), and a priori tropospheric profiles. The surface albedo is taken from the monthly OMI albedo climatology (minimum Lambertian equivalent reflectivity, (Kleipool et al., 2008)) at the spatial resolution of $1°x1°$. A priori vertical profiles are specified using the TM5-MP daily forecast, at the same spatial resolution (Williams et al., 2017). Cloud properties are provided by the S5P operational product in its CRB mode (Cloud as Reflecting Boundary, Loyola et al. (2018)). A cloud correction based on the independent pixel approximation (Boersma et al., 2004) is applied for cloud fractions larger than 0.1. In order to correct for any remaining global offset and stripes, a background correction is applied based on HCHO slant columns from the 5 previous days in the Pacific Ocean ($N_{(s,0)}$), as described in De Smedt et al. (2018). Finally, the background vertical

[1] http://www.tropomi.eu/sites/default/files/files/publicSentinel-5P-Formaldehyde-Readme_20191213.pdf

column of HCHO, due to methane oxidation, is taken from the TM5 model in the reference region ($N_{(v,0)}^{CTM}$). The equation of the tropospheric HCHO vertical column can be written as follows:

$$N_v = \frac{(N_s - N_{(s,0)})}{M} + \frac{M_0}{M} \cdot N_{(v,0)}^{CTM}, \qquad (1)$$

with M_0 the average of the air mass factors M of the slant columns selected in the reference sector, the Pacific Ocean ($N_{(s,0)}$). Intermediate quantities and auxiliary data are all provided in the L2 files (http://www.tropomi.eu/sites/default/files/files/Sentinel-5P-Level-2-Product-User-Manual-Formaldehyde_v1.01.01_20180716.pdf).

Several diagnostic variables are provided together with the measurements. Quality assurance values (QA) are defined to perform a quick selection of the observations. QA>0.5 filters out most observations presenting an error flag or a solar zenith angle (SZA) larger than 70°, a cloud radiance fraction at 340 nm larger than 0.6 or an air mass factor smaller than 0.1. The product Readme file reports that in the current version, the QA values are not always correctly set over snow/ice regions or above 75° of SZA. They also need to be further checked over cloudy scenes. In the forthcoming S5P version 2, QA values will be refined, and will exclude data with surface albedo larger than 0.2 and snow/ice warning, and remaining SZA larger than 75°.

The tropospheric column uncertainty is divided into random (precision) and systematic components (accuracy), and is provided per pixel. It varies with the observation conditions. Over remote regions at moderate solar zenith angle, the precision of an individual observation is about 5x10^{15} molec/cm^2. This value agrees with the standard deviation of the columns in the same region for a particular day. The random uncertainty is dominated by the random error on the slant columns. The tropospheric column accuracy is the combined systematic uncertainty resulting from the slant column, the air mass factor and the background correction errors. It varies between 30 and 60% of the columns. The column averaging kernel and the a priori profiles are provided for every observation.

3 Ground-based FTIR HCHO data

We show in Fig. 1 a map of the ground-based FTIR stations used in this TROPOMI validation. The background image represents the global TROPOMI monthly mean tropospheric columns for September 2018, illustrating the different HCHO levels sampled by the ground-based network: from clean Arctic and oceanic sites to very high-concentrations sites such as Porto Velho, in the Amazon basin.

Table 2 lists the ground-based FTIR stations, their coordinates and altitude, the spectrometer type, the retrieval code, and the team involved in the measurements and/or the retrievals of HCHO. For more details on the monitoring of FTIR solar absorption spectra at these stations, we refer to Vigouroux et al. (2018) and references therein, and, for the FTIR retrieval principles, to e.g. Vigouroux et al. (2009).

The same retrieval settings are used at all the stations to avoid introducing possible bias in the HCHO total columns between the stations and inconsistent comparisons with the satellite. Details are given in Vigouroux et al. (2018). The main settings that might be responsible for internal biases within the network are the spectroscopic database and the fitted spectral windows, the spectroscopic parameters being the main source of the FTIR HCHO systematic uncertainties. The HCHO spectral signatures

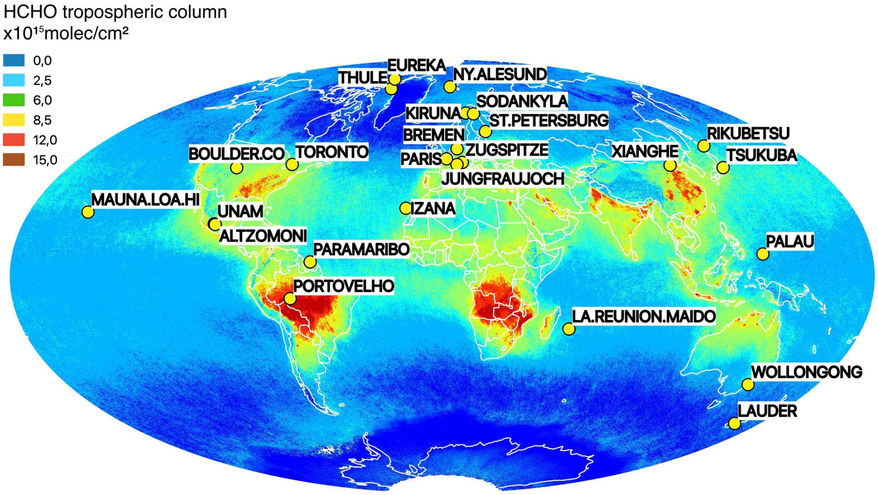

Figure 1. Network of ground-based FTIR stations providing HCHO total columns data. The background is the September 2018 monthly mean of TROPOMI HCHO tropospheric columns, averaged on a 0.2°x0.2° grid, using the HARP tool v.1.5 (https://atmospherictoolbox.org).

lie in the 3.6 μm region and belong to the ν_1 and ν_5 bands (fitted windows are: 2763.42 - 2764.17; 2765.65 - 2766.01 ; 2778.15 - 2779.1 ; 2780.65 - 2782.0, in cm^{-1}). The spectroscopic database used is the atm16 linelist by G. Toon (JPL), which can be found at http://mark4sun.jpl.nasa.gov/toon/linelist/linelist.html. This linelist is optimized for the main absorbing gases in the fitted windows (HDO, CH_4, O3, N_2O, CO_2) and is based on HITRAN 2012 (Rothman et al., 2013) for HCHO, which used the work of Jacquemart et al. (2010).

The retrieval codes used in the FTIR NDACC community are PROFITT9 (Hase et al., 2006) and SFIT4.0.9.4 (updated from SFIT2 (Pougatchev et al., 1995)), which are both based on the optimal estimation method (Rodgers, 2000). A past comparison exercise has shown a very good agreement between the retrieved products obtained with these two codes (Hase et al., 2004). Based on a priori profile information (from the model WACCM, Garcia et al. (2007)), and a L1 Tikhonov regularization matrix (Tikhonov, 1963), low vertical resolution profiles can be retrieved in principle, as well as total columns. However, as described in Vigouroux et al. (2018), the degrees of freedom for signal are very low for HCHO (median value of 1.1 for all FTIR sites), meaning that we essentially have one piece of information. The FTIR total column averaging kernel shows a decrease of the sensitivity at the surface, which is quite similar to the TROPOMI sensitivity. This can be seen in Fig. 2, as an example for the Maïdo station. We also show in Fig. 2 the FTIR a priori profile at Maïdo, which is based on a climatology (1980-2020) from the WACCM model calculated at Maïdo. A single profile is used for the whole time series at a specific station (Vigouroux et al., 2018), while TROPOMI uses daily a priori profiles from TM5 (Sect. 2). An example is shown in Fig. 2 for the 18th January 2019.

Table 2. FTIR stations that are contributing to the present work: location and altitude (in km a.s.l.), instrument type, retrieval code, team.

Station	Latitude	Longitude	Altitude	Instrument	Code	Team
Eureka	80.05° N	86.42° W	0.61	Bruker 125 HR	SFIT4	U. of Toronto
Ny-Ålesund	78.92° N	11.92° E	0.02	Bruker 120 HR	SFIT4	U. of Bremen
Thule	76.52° N	68.77° W	0.22	Bruker 125 HR	SFIT4	NCAR
Kiruna	67.84° N	20.40° E	0.42	Bruker 120/5 HR	PROFFIT	KIT–ASF ; IRF Kiruna
Sodankylä	67.37° N	26.63° E	0.19	Bruker 125 HR	SFIT4	FMI ; BIRA
St-Petersburg	59.88° N	29.83° E	0.02	Bruker 125 HR	SFIT4	SPbU
Bremen	53.10° N	8.85° E	0.03	Bruker 125 HR	SFIT4	U. of Bremen
Paris	48.97° N	2.37° E	0.06	Bruker 125 HR	PROFFIT	Sorbonne U.
Zugspitze	47.42° N	10.98° E	2.96	Bruker 120/5 HR	PROFFIT	KIT–IFU
Jungfraujoch	46.55° N	7.98° E	3.58	Bruker 120 HR	SFIT4	U. of Liège
Toronto	43.60° N	79.36° W	0.17	Bomem DA8	SFIT4	U. of Toronto
Rikubetsu	43.46° N	143.77° E	0.38	Bruker 120/5 HR	SFIT4	Nagoya U. ; NIES
Boulder	40.04° N	105.24° W	1.61	Bruker 125 HR	SFIT4	NCAR
Xianghe	39.75° N	116.96° E	0.05	Bruker 125 HR	SFIT4	CAS ; BIRA
Tsukuba	36.05° N	140.12° E	0.03	Bruker 125 HR	SFIT4	NIES ; Tohoku U.
Izaña	28.30° N	16.48° W	2.37	Bruker 120/5 HR	PROFFIT	AEMET ; KIT–ASF
Mauna Loa	19.54° N	155.57° W	3.40	Bruker 125 HR	SFIT4	NCAR
Mexico City (UNAM)	19.33° N	99.18° W	2.26	Bruker Vertex 80	PROFFIT	UNAM
Altzomoni	19.12° N	98.66° W	3.98	Bruker 120/5 HR	PROFFIT	UNAM
Palau	7.34° N	134.47° E	0.03	Bruker 120/5 M	SFIT4	U. of Bremen
Paramaribo	5.81° N	55.21° W	0.03	Bruker 120/5 M	SFIT4	U. of Bremen
Porto Velho	8.77° S	63.87° W	0.09	Bruker 125 M	SFIT4	BIRA
Maïdo (LA.REUNION.MAIDO)	21.08° S	55.38° E	2.16	Bruker 125 HR	SFIT4	BIRA
Wollongong	34.41° S	150.88° E	0.03	Bruker 125 HR	SFIT4	U. of Wollongong
Lauder	45.04° S	169.68° E	0.37	Bruker 120 HR	SFIT4	NIWA

The FTIR uncertainty budget is calculated following the formalism of Rodgers (2000) and is described in Vigouroux et al. (2018). It is separated into random and systematic components. The random uncertainty is dominated at all sites by the measurement noise uncertainty, which can vary from site to site depending on the spectrometer. The uncertainty on the retrieved FTIR total columns for individual sites is given in Vigouroux et al. (2018) for the 21 sites involved at that time. We obtain a median random uncertainty of 2.3 $\times 10^{14}$ molec/cm^2, with a large value of 11.1 $\times 10^{14}$ molec/cm^2 only at Mexico City where a lower resolution instrument is used (Vertex 80). The smoothing uncertainty on the total column has a non negligible random component (median value of 1.2 $\times 10^{14}$ molec/cm^2). With the inclusion of the smoothing error in the uncertainty budget, the

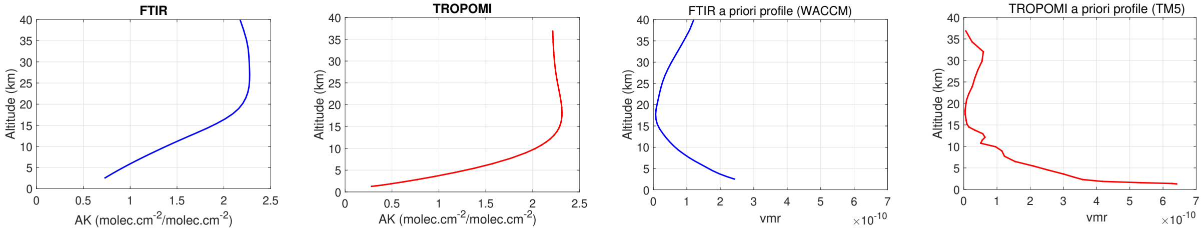

Figure 2. Left panels: typical total column averaging kernel (AK) from FTIR (blue) and TROPOMI (red) measurements at the Maïdo station (altitude: 2.2 km). Right panels: a priori profile used in the FTIR retrievals for the whole time series at Maïdo (blue), and an example of TROPOMI a priori profile at Maïdo (the 18th January 2019).

median total random uncertainty is 2.9 $\times 10^{14}$ molec/cm^2, which is very close to our empirical uncertainty estimation of 2.8 $\times 10^{14}$ molec/cm^2, based on the standard deviation of the differences between two FTIR individual subsequent measurements (within a maximum of 30 minutes interval), confirming our theoretical uncertainty calculation. Since the Vigouroux et al. (2018) paper, five more sites have joined the HCHO harmonized network. The mean random errors on individual FTIR measurements are (in molec/cm^2): 1.4 $\times 10^{14}$, 2.7 $\times 10^{14}$, 2.2 $\times 10^{14}$, 5.2 $\times 10^{14}$, and 5.6 $\times 10^{14}$, for Jungfraujoch, Rikubetsu, Tsukuba, Palau, and Xianghe, respectively.

The forward model parameters median systematic uncertainty on the HCHO FTIR total columns is 13% in the network described by Vigouroux et al. (2018). As already mentioned, the dominating systematic uncertainty sources are the spectroscopic parameters: the line intensities and the pressure broadening coefficients of the fitted HCHO absorption lines. We use 10% for the three parameters: the line intensity, and the air- and self- broadening coefficients. The systematic uncertainty can be larger (up to 21-26%) at the stations using the PROFFIT9 retrieval code, due to an assumed uncertainty on the channeling that is not taken into account yet in the SFIT4 code. However, this channeling uncertainty can also be negligible at some sites (it depends on each instrument), and more investigation is needed at each station to avoid its under- or over-estimation. The median smoothing systematic uncertainty is 3.4%. For the five added sites, the median total systematic uncertainty is 13% (Jungfraujoch, Tsukuba, Palau), or 14% (Rikubetsu, Xianghe), commensurate with the other sites.

4 Validation method

4.1 Collocation criteria

The precision of a single pixel TROPOMI HCHO measurement is expected to be below 1.2 $\times 10^{16}$ molec/cm^2 (pre-launch requirements) or even better, as 5 $\times 10^{15}$ molec/cm^2 for remote areas (after launch uncertainty analysis, see Sect. 2). These values are quite large compared to the measured levels of HCHO (around 1.5 $\times 10^{15}$ molec/cm^2 for very clean sites to e.g. around 9 $\times 10^{15}$ molec/cm^2 for a city such Paris). It is therefore necessary to average several pixels in order to reduce the random uncertainty of the TROPOMI mean HCHO data, improve the detection level and increase the TROPOMI sensitivity to day-to-day variability. For this reason, we choose to average the TROPOMI pixels located within 20 km from the FTIR station.

Once we filter out the TROPOMI pixels that do not reach the recommended quality criteria (QA flag > 0.5; see Sect. 2), we obtain a median value of 34 pixels to average. In cloudy conditions, this number can be smaller. A collocation pair is kept when at least 10 pixels can be averaged. Higher number of pixels can be averaged for Arctic stations (around 45-60), which is useful due to the very low HCHO levels to be detected there. At sub-tropical / tropical stations, the median number of pixels is around 20-29. The higher number of pixels in Arctic is due to the fact that each FTIR measurement is co-located to all S5P pixels that match the co-location criteria, even if these pixels originate from different orbits, with different overpass times. Before choosing the 20 km collocation criteria, we have tested several distances (10, 20, 30, 40, and 50 km). The 10 km criterion was discarded because of the poor remaining coincidences leading to less robust statistics. The 20 to 50 km criteria give similar biases between TROPOMI and FTIR. The standard deviations of the comparisons usually decrease slightly with increasing collocation distance due to a smaller TROPOMI random uncertainty (more pixels to average), except at the most polluted sites. However, the ratio between the standard deviations and the random uncertainty budgets is increasing with the collocation distance at all sites, pointing to an increased random error due to the collocation. We therefore choose the 20 km distance to reduce the random spatial collocation error.

The time coincidence criterion is set to ±3 hours. This choice is a compromise to obtain significant number of coincidences between TROPOMI and FTIR data, noting that the median FTIR measurement frequency is 5 per day (with a range of 3 to 10 depending on the station). A shorter time coincidence criterion decreases significantly the numbers of sampled collocated days and sometimes months, which is a limitation for checking the TROPOMI seasonality (sampled months: 267 for 1 h, and 305 for 3 h criteria). Note that a 6 h criterion would provide 20 additional sampled months: the critical stations are Mauna Loa, Altzomoni and Paramaribo, for which we would have coincidences back to May 2018. The standard deviations of the TROPOMI / FTIR comparisons are usually smaller with a longer time coincidence criterion, but this can be explained by the increased number of pixels (improved TROPOMI precision on the mean) in the 6 h collocation, mainly at Arctic sites with increased number of multiple orbits. Despite the smaller standard deviations usually obtained within a 6 h criterion, we finally choose 3 h to reduce the possible impact of some passing plumes and of the HCHO diurnal cycle on the comparisons. The diurnal cycle at most of the FTIR stations can be found in Vigouroux et al. (2018) and its Supplement. At many stations no significant diurnal cycle was observed but, in some cases, mainly polluted sites, we obtained a maximum around noon-1 p.m., close to the TROPOMI overpass time. At the Mexico City station, where the diurnal cycle amplitude is the greatest, the effect of collocation time (6 h vs 3 h) on the statistical bias is 4%.

4.2 Building inter-comparable products

Some manipulation of the original data products is needed before looking at the differences between TROPOMI and FTIR data. Both measurements provide total columns (for FTIR) or tropospheric columns (for TROPOMI) that have a lower sensitivity near the ground (see Fig. 2), and their retrievals use a priori profile information that have been chosen differently (TROPOMI: daily a priori profiles from TM5; FTIR: single a priori profile from climatology of WACCM). To correct for this, for each S5P individual pixel collocated with each FTIR measurement, we use the comparison method described in Rodgers and Connor (2003). First, the a priori substitution is applied, using the S5P a priori profile $\boldsymbol{x}_{S,a}$ as the common a priori profile. For this, the S5P a priori profile is regridded to the FTIR retrieval grid ($\boldsymbol{x}_{S,a/F}$) using a mass conservation algorithm (Langerock et al., 2015). In the rare situation where the satellite pixel elevation is above the FTIR site, the S5P a priori profile is extended to the FTIR instrument's altitude. The regridded S5P a priori $\boldsymbol{x}_{S,a/F}$ is then substituted following Rodgers and Connor (2003), and we finally use the corrected FTIR retrieved profile $\boldsymbol{x}'_F$ in the comparisons:

$$\boldsymbol{x}'_F = \boldsymbol{x}_F + (\mathbf{A}_F - \mathbf{I})(\boldsymbol{x}_{F,a} - \boldsymbol{x}_{S,a/F}), \tag{2}$$

where $\boldsymbol{x}_F$ is the original FTIR retrieved profile, $\mathbf{A}_F$ is the FTIR averaging kernel matrix, $\mathbf{I}$ is the unit matrix, and $\boldsymbol{x}_{F,a}$ is the FTIR a priori profile.

The next step, following Rodgers and Connor (2003), is to smooth the corrected FTIR profile with the S5P column averaging kernel $\boldsymbol{a}_S$. For that purpose we regrid the corrected FTIR profile $\boldsymbol{x}'_F$ to the S5P column averaging kernel grid ($\boldsymbol{x}'_{F/S}$) and apply the smoothing equation:

$$c_F^{smoo} = c_{S,a} + \boldsymbol{a}_S(\boldsymbol{x}'_{F/S} - \boldsymbol{x}_{S,a}) \tag{3}$$

with $c_{S,a}$ the S5P a priori column derived from the S5P a priori profile. We obtain a smoothed FTIR column c_F^{smoo} associated with a collocated TROPOMI pixel. In the case of mountain sites where the pixel altitude is below the instrument's height, the regridding of the FTIR profile $\boldsymbol{x}'_{F/S}$ is done such that the FTIR profile is extended with the S5P a priori profile (such an extension is invariant under the latter smoothing equation). Note that this FTIR regridding to the satellite grid has also the advantage that only the FTIR profile up to the altitude of the satellite product (which is only a tropospheric column) remains in the regridded column: we therefore finally compare tropospheric columns in both products.

Next, we need to take into account that, for mountain stations, the difference between satellite columns and the original ground-based columns can be significant. To bring both satellite and smoothed FTIR column c_F^{smoo} (which is calculated as a column valid at the satellite's pixel surface) values to the scale of the original FTIR columns, we apply a scaling factor f representative for the fraction of the partial column between the satellite pixel altitude and the FTIR station altitude. This scaling factor is derived from the satellite a priori profile and is defined as:

$$f = 1 - \frac{c_{S,a}^{\Delta z}}{c_{S,a}}, \tag{4}$$

where $c_{S,a}^{\Delta z}$ denotes the partial column derived from the S5P a priori profile between the pixel surface and the FTIR station. The TROPOMI column c_S and its random and systematic uncertainties are also scaled with the same factor, so that finally the collocated products are all expressed at the altitude of the FTIR site (and not of the pixel surface). For mountain stations, the scaling factor f, calculated for each satellite's pixel, can reach a minimum of 0.5 for stations located at about 2 km altitude from the satellite's pixel surface (Maïdo, Izaña, or Altzomoni), or even 0.3 at the higher sites Jungfraujoch and Zugspitze, while at sea-level sites it is of course close to 1.0. In the rare cases where the satellite pixel is above the FTIR station, we apply the conversion factor $f = 1 + c_{S,a}^{\Delta z}/c_{S,a}$, where the satellite a priori profile is extrapolated to the station surface in order to calculate the partial column of the a priori between both altitudes.

The final step is to average the individual smoothed and scaled FTIR columns $c_F^{smoo} \times f$ that are taken within 3 h, and the TROPOMI $c_S \times f$ individual pixel columns that are available within 20 km (which can belong to different orbits), to form the collocated pair FTIR$_i$ and TROP$_i$ used in the next section.

4.3 Estimation of the TROPOMI accuracy and precision

In Sect. 5.1, we assess whether the TROPOMI accuracy is compliant with pre-launch requirements (40-80%, as reported in the ESA official document S5P-RS-ESA-SY-164, 2014, Table 3, p. 19). The accuracy of the TROPOMI HCHO measurements will be estimated by deriving the median of the relative differences (BIAS) between the collocated TROP$_i$ and the reference FTIR$_i$ data at each station:

$$\text{BIAS} = med(\frac{(\text{TROP}_i - \text{FTIR}_i)}{\text{FTIR}_i}). \tag{5}$$

We can note that the applied scaling factor f (see previous section) does not affect the BIAS estimation, even at high mountains stations, because it cancels in the division.

For robust statistics, the median is preferred to the mean due to the presence of outliers (a few remaining TROPOMI outliers after the QA filter, and some very small FTIR values that give very large relative difference after the division in Eq. 5). The presence of TROPOMI outliers is minimized by using the median, but they should be ideally removed by the QA filter. An improvement of the QA value is foreseen in the next product version, which should improve, e.g., the filtering at Arctic sites (SZA>75°).

In the next section, we also compare the obtained BIAS with the systematic uncertainty on the difference σ_{syst}, to evaluate the TROPOMI uncertainty budget:

$$\sigma_{\text{syst}}^2 = (\sigma_{S,\text{syst}})^2 + \boldsymbol{a}_S^T \mathbf{S}_{F,\text{syst}} \boldsymbol{a}_S + \boldsymbol{a}_S^T (\mathbf{I} - \mathbf{A}_F) \mathbf{S}_{\text{var,syst}} (\mathbf{I} - \mathbf{A}_F)^T \boldsymbol{a}_S, \tag{6}$$

where $\sigma_{S,\text{syst}}$ is the systematic uncertainty of TROPOMI columns, as provided in the public release database (but scaled for altitude, see Sect. 4.2), $\boldsymbol{a}_S$ is the TROPOMI total column averaging kernel, and $\mathbf{S}_{F,\text{syst}}$ is the FTIR systematic covariance matrix provided in vmr^2 in the standardized GEOMS format converted in partial columns units. The last term is the impact of different low vertical resolution profile measurements (the smoothing error) on the comparisons (see Eq. 27 in Rodgers and Connor (2003)), where for the systematic uncertainty part, we account for possible bias on $\boldsymbol{x}_{S,a}$ by following von Clarmann (2014):
$\mathbf{S}_{\text{var,syst}} = (\boldsymbol{x}_{S,a} - <\boldsymbol{x}>)(\boldsymbol{x}_{S,a} - <\boldsymbol{x}>)^T$.

The $\boldsymbol{x}_{S,a} - <\mathbf{x}>$ is not known and we follow Vigouroux et al. (2018), with $\boldsymbol{x}_{S,a} - <\mathbf{x}>$=-50%, -20%, -10%, +10%, +8%, +5% for the ground-4 km; 4-8 km; 8-13 km; 13-25 km; 25-40 km; 40-120 km layers, respectively (expressed in molec/cm^2). The last term of Eq. 6 is found to be of the order of a few percent, therefore negligible in σ_{syst}. In practice, the systematic uncertainty on the difference σ_{syst} is dominated by the TROPOMI systematic uncertainty of about 40%, FTIR having a median systematic uncertainty of only 13% with a maximum of 26% (See Sect. 3).

Similarly, the precision of the TROPOMI HCHO products is estimated in Sect. 5.2, not with the usual standard deviation which is not robust in case of outliers, but with the median absolute deviation (MAD, see Huber (1981)) of the differences (DIFF_i=TROP_i-FTIR_i):

$$\mathrm{MAD} = k \times med(abs(\mathrm{DIFF}_i - med(\mathrm{DIFF}_i))), \tag{7}$$

where $k = 1.4826$ for a correspondence with the 1-σ standard deviation for normal distribution without outliers.

In Sect. 5.2, we compare the obtained MAD to the random uncertainty on the differences σ_{rand}, which is calculated following Rodgers and Connor (2003):

$$\sigma^2_{\mathrm{rand}} = (\sigma_{S,\mathrm{rand}})^2 + \boldsymbol{a}_S^T \mathbf{S}_{F,\mathrm{rand}} \boldsymbol{a}_S + \boldsymbol{a}_S^T (\mathbf{I} - \mathbf{A}_F) \mathbf{S}_{\mathrm{var,rand}} (\mathbf{I} - \mathbf{A}_F)^T \boldsymbol{a}_S, \tag{8}$$

where where $\sigma_{S,\mathrm{rand}}$ is the random uncertainty of TROPOMI columns, as provided in the public release database (but scaled for altitude, see Sect. 4.2), $\mathbf{S}_{F,\mathrm{rand}}$ is the FTIR random covariance matrix, and $\mathbf{S}_{\mathrm{var,rand}}$, to take into account the impact of low vertical resolution in the random part of the uncertainty, is the natural variability matrix chosen to be 50%, 50%, 40%, 35%, 30%, 30%, 10% for the ground-4 km; 4-8 km; 8-13 km; 13-25 km; 25-40 km; 40-120 km layers, respectively (expressed in molec/cm^2). As for the systematic uncertainty part, the random uncertainty on the difference is dominated by the TROPOMI random uncertainty (median of about 1.1 $\times 10^{15}$molec/cm^2 for TROP_i within 20 km), while FTIR_i has a median random uncertainty of 2.0 $\times 10^{14}$molec/cm^2. The last term of Eq. 8 is comparable to the FTIR one (median value of 2.4 $\times 10^{14}$molec/cm^2).

We can use MAD as an upper limit of the TROPOMI precision, since collocation in space and time of the sounded air-masses are never perfect. It is compared in the next section to the pre-launch precision requirement. The MAD estimation is influenced by the scaling factor f, which is important only for high altitude sites (Sect. 4.2). It should be interpreted as an estimation of the precision of a TROPOMI column that would be measured at the altitude of the FTIR site. The random uncertainty on the differences are also expressed at the altitude of the FTIR site, so that the comparison between MAD and σ_{rand} is always valid.

The observed BIAS between TROPOMI and the reference FTIR data is statistically significant if it exceeds its statistical error: $\mathrm{ERR_B} = 2 \times \mathrm{MAD}/\sqrt{n}$ (with n the number of coincidences).

5 Validation results

In this section, we provide a table and plots for the offline (RPRO+OFFL) HCHO TROPOMI product. We do not show detailed results for the near real time (NRTI) product (version 1.1.[5-7]) because they are very similar to the offline version. Numbers for the main conclusions will be given in the text for this NRTI product.

5.1 TROPOMI observed BIAS and accuracy

In Table 3, we provide, at each individual FTIR station, the mean of the FTIR HCHO total columns (mean FTIR), the obtained median of the relative differences BIAS (in % to compare with the pre-launch TROPOMI accuracy requirements of 40-80%, Eq. 5), the error on the BIAS ($\mathrm{ERR_B}$), and the number of collocated pairs n. The systematic uncertainty on a single difference is also given (in %, calculated from Eq. 6 where each term has been expressed in %, dividing by each instrument individual HCHO column).

We have ordered the stations, not in decreasing latitudes as in Table 2, but in increasing mean HCHO FTIR columns. The reason is that we observe a tendency of the BIAS between TROPOMI and FTIR: while the BIAS is always (with the exception of Eureka) positive or not significant (if $\mathrm{BIAS}<\mathrm{ERR_B}$) for very clean to clean sites with HCHO mean levels lower than 6.5×10^{15} molec/cm^2, it is negative and very consistent for the stations with higher HCHO levels, ranging from 8.7 to 28.6×10^{15} molec/cm^2 (-29 to -36 %) with small error on the bias (2 to 6 %). Note that the BIAS is also consistent at Paramaribo (-26%) but with larger error (14%), due to small number of collocations. This dependence of the TROPOMI bias on the HCHO concentration levels can be visualized in Fig. 3, where the BIAS at each station is plotted as a function of the mean FTIR columns. It is therefore not appropriate to use the median bias obtained using the data from all stations together (-10%), if one wants to correct the TROPOMI HCHO data in model inversion studies. If we calculate the median of the differences for HCHO FTIR columns >8.0×10^{15} molec/cm^2, we obtain a significant negative bias of $-30.8\pm1.4\%$. The detection of this bias is especially important for modeling studies that use satellite data to optimize the volatile organic compound emissions sources, as done in e.g. Fortems-Cheiney et al. (2012); Stavrakou et al. (2015) with OMI and GOME-2. The bias for clean HCHO levels (<2.5×10^{15} molec/cm^2) is significantly positive (+26 $\pm5\%$).

The validation results for the NRTI TROPOMI products give very similar results: a negative BIAS ($-31.7\pm1.8\%$) for the high HCHO levels (>8.0×10^{15} molec/cm^2) and a positive one (+22 $\pm7\%$) for low HCHO levels (<2.5×10^{15} molec/cm^2). The small differences are mainly due to the different sampling of the comparisons (NRTI data are since December 2018, while the OFFL data are since May 2018).

The different TROPOMI BIAS at different HCHO levels is pointing to the presence of two kinds of bias: a constant one and a proportional one. They can be obtained by using the scatter plot of the two instruments, shown in Fig. 4: the constant bias is the intercept of the linear relationship between TROPOMI and FTIR, while the proportional bias is given by its slope. But this has to be done carefully: a usual linear regression by ordinary least squares (OLS) is not statistically robust and can give spurious results in the presence of outliers and/or heteroscedasticity. We are confronted to both problems in our scatter plot: we do have outliers and the uncertainty is increasing with HCHO levels. Therefore, we use the robust Theil-Sen estimator (Sen, 1968) where the slope s of the scatter plot is the median of the slopes of the lines through all pairs of data points (TROP_j - $\mathrm{TROP}_i)/(\mathrm{FTIR}_j$ - FTIR_i), with $\mathrm{FTIR}_j \neq \mathrm{FTIR}_i$. The intercept b is then the median of ($\mathrm{TROP}_i - s\times\mathrm{FTIR}_i$). Using this robust estimator, we obtain the relation: TROP = 0.64 $\times$ FTIR + 1.10×10^{15} molec/cm^2. We have calculated the uncertainties in s and b using $2\times\mathrm{MAD}/\sqrt{n}$, with MAD the median absolute deviation of the slopes and intercepts of the pairs of data points, and n the numbers of pairs. We obtain an uncertainty of 0.03 and 0.05×10^{15} molec/cm^2 for s and b respectively. Therefore, both the constant ($1.10\pm0.05\times10^{15}$ molec/cm^2) and proportional ($0.64\pm0.03\%$) biases are significant.

Using the scatter plot to derive the constant and proportional biases is very sensitive to the range of observed values. As an example, if one would only use HCHO FTIR data >8.5×10^{15} molec/cm^2, one would obtain a slope of 0.51 and an intercept of 3.2×10^{15} molec/cm^2), which would point to a strong overestimation and underestimation of the constant and proportional

Table 3. Validation of TROPOMI RPRO+OFFL. Please note that the ordering of the sites is by increasing mean HCHO column. For each station: mean of the HCHO FTIR total columns (in molec/cm^2), median of the relative differences BIAS=med((TROP$_i$-FTIR$_i$)/FTIR$_i$) and its error ERR$_\mathrm{B}$ (in %, see text), number of collocated pairs n, systematic uncertainty on a single difference σ_syst (in %, Eq. 6), median absolute deviation (MAD, in molec/cm^2, Eq. 7), random uncertainty on a single difference σ_rand (in molec/cm^2, Eq. 8), and pre-launch TROPOMI precision requirements associated to the choice of 20 km around the station Requ=$1.2\times10^{16}/\sqrt{n}_\mathrm{pix}$ molec/cm^2, with n_pix the mean number of pixels averaged in the collocated TROPOMI data. The Pearson correlation coefficient R is given for individual coincidences (±3 h) and for monthly means of coincident data.

Station	mean FTIR	BIAS	n	σ_syst	MAD	σ_rand	Requ.	n$_{pix}$	R	R
	molec/cm^2	± ERR$_\mathrm{B}$ %		%	molec/cm^2	molec/cm^2	molec/cm^2		indiv.	monthly
Jungfraujoch	1.24E+15	19 ± 15	87	58	9.0E14	5.6E14	2.5E15	24	0.61	0.70
Zugspitze	1.36E+15	52 ± 10	184	59	7.8E14	5.0E14	2.1E15	33	0.71	0.86
Mauna Loa	1.60E+15	52 ± 22	52	54	1.3E15	8.8E14	2.5E15	23	-0.09	-0.05
Eureka	1.65E+15	-40 ± 11	114	97	1.1E15	5.3E14	1.8E15	45	0.22	0.43
Maïdo	1.86E+15	5 ± 9	155	43	1.0E15	7.1E14	3.0E15	16	0.45	0.53
Ny-Ålesund	1.90E+15	43 ± 20	47	41	1.2E15	4.9E14	1.7E15	52	0.35	0.38
Thule	2.06E+15	-3 ± 5	346	57	9.6E14	4.8E14	1.6E15	60	0.56	0.82
Izaña	2.07E+15	13 ± 10	97	83	8.9E14	6.4E14	2.5E15	24	0.47	0.79
Altzomoni	2.44E+15	66 ± 18	67	42	1.6E15	8.9E14	2.6E15	22	0.50	0.86
Kiruna	2.44E+15	50 ± 12	146	67	1.5E15	8.8E14	1.7E15	60	0.64	0.72
Lauder	2.54E+15	-11 ±14	225	78	2.6E15	1.3E15	2.1E15	33	0.38	0.65
Rikubetsu	3.16E+15	26 ± 40	16	50	2.8E15	1.0E15	1.9E15	41	0.45	0.60
Palau	3.80E+15	0 ±15	10	36	9.8E14	8.2E14	2.7E15	20	0.15	0.33
Sodankyla	4.15E+15	8 ± 7	307	51	2.5E15	1.2E15	1.7E15	48	0.51	0.69
Boulder	5.91E+15	-1 ± 9	103	50	2.2E15	1.3E15	2.2E15	31	0.79	0.90
St-Petersburg	6.21E+15	-4 ± 8	158	44	3.0E15	1.2E15	1.9E15	42	0.68	0.78
Wollongong	6.36E+15	9 ± 8	322	54	3.3E15	1.9E15	2.3E15	27	0.78	0.94
Tsukuba	7.05E+15	-23 ±12	34	44	3.1E15	1.2E15	2.2E15	31	0.68	0.51
Bremen	7.77E+15	-5 ± 12	46	39	3.2E15	1.4E15	1.8E15	43	0.63	0.68
Paramaribo	8.43E+15	-26 ±14	15	36	3.3E15	1.3E15	2.5E15	23	0.12	0.14
Paris	8.72E+15	-29 ±6	128	44	3.1E15	1.2E15	1.9E15	41	0.76	0.79
Toronto	1.06E+16	-34 ±4	251	38	3.7E15	1.6E15	2.3E15	29	0.73	0.95
Xianghe	1.43E+16	-33 ±2	384	38	5.5E15	2.1E15	2.0E15	36	0.86	0.97
Mexico City (UNAM)	1.92E+16	-27 ± 4	154	32	5.9E15	2.6E15	2.5E15	25	0.34	0.27
Porto Velho	2.86E+16	-36 ± 3	81	31	8.3E15	3.6E15	2.2E15	29	0.81	1.00
All stations BIAS	6.60E+15	-10 ±2	3529	48	2.4E15	1.2E15	2.1E15	34	0.87	0.91
Low HCHO BIAS FTIR<2.5×10^{15}	1.64E+15	+26 ±5	1321	52	1.3E15	7.7E14	2.2E15	31	0.40	
High HCHO BIAS FTIR>8.0×10^{15}	1.57E+16	-30.8 ±1.4	952	46	4.8E15	2.1E15	2.1E15	33	0.88	

biases, respectively. This would lead to erroneous prediction of TROPOMI overall bias outside the range of observed values. This demonstrates why such a ground-based network, covering very clean sites to high HCHO level sites, is crucial to provide a good estimate of both constant and proportional biases of TROPOMI.

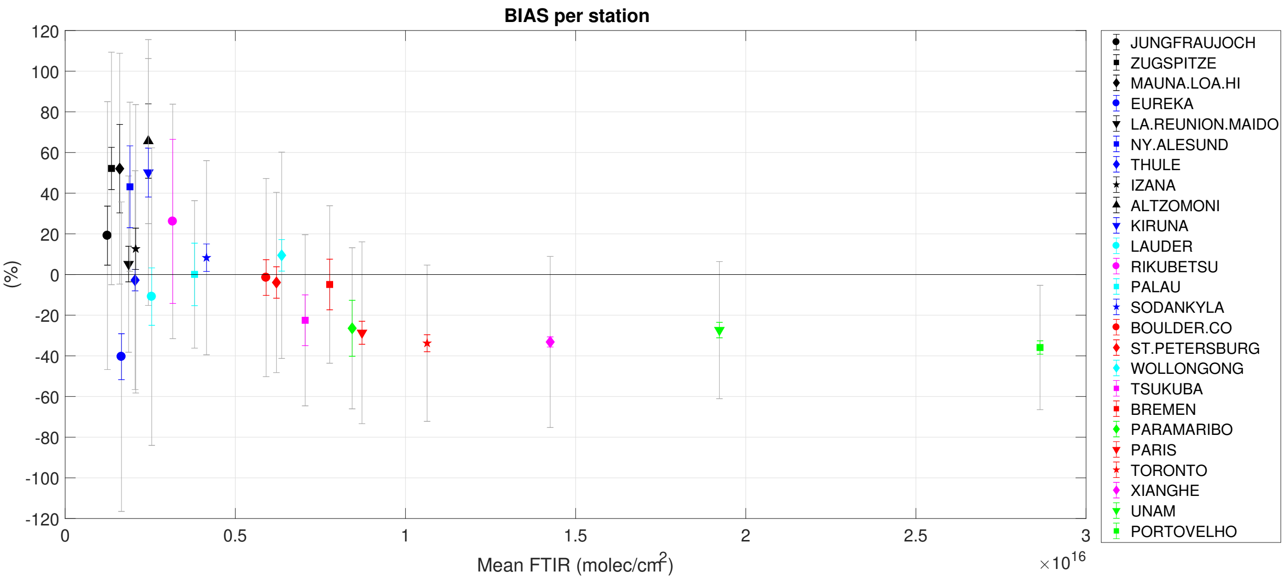

Figure 3. BIAS at each station (in %) as a function of the mean FTIR total columns (molec/cm^2). The gray bars are the systematic uncertainty on the differences σ_{syst} , and the colored error bars are the 2-σ error on the bias $\mathrm{ERR_B}$. Black markers are for mountains stations, blue for Arctic stations, cyan for Oceania/Australia/New Zealand, magenta for China/Japan, red for mid-latitude Europe/North America, and green for Central/South America.

The BIAS given in Table 3 are a combination of the constant and proportional biases, and can be use to statistically assess the TROPOMI HCHO overall accuracy. We can easily see from Table 3 that all BIAS values are within the upper limit of the pre-launch requirement of 80%, and they are within the 40% requirement lower limit for 20 of the 25 stations. The five stations exceeding a 40% BIAS are clean (Arctic or mountains) sites, with mean HCHO columns below 2.5×10^{15} molec/cm^2. But these are sites where the systematic uncertainty on the differences (see Table 3 and Eq. 6) are usually also the largest, leading to a good correspondence between observed higher BIAS and higher calculated uncertainty for 3 of these 5 stations (Zugspitze, Mauna Loa, and Kiruna).

Therefore, we can conclude that the TROPOMI HCHO accuracy satisfies the pre-launch requirements and that the systematic uncertainty budget is in very good agreement with observed bias, except at a very few stations (Ny-Ålesund 43>41%, Altzomoni 71>42%, and Porto Velho 36>31%). At most of the other stations, the reported systematic uncertainty tends to be larger than the BIAS. We find the same conclusions on TROPOMI accuracy when making comparisons with the NRTI products.

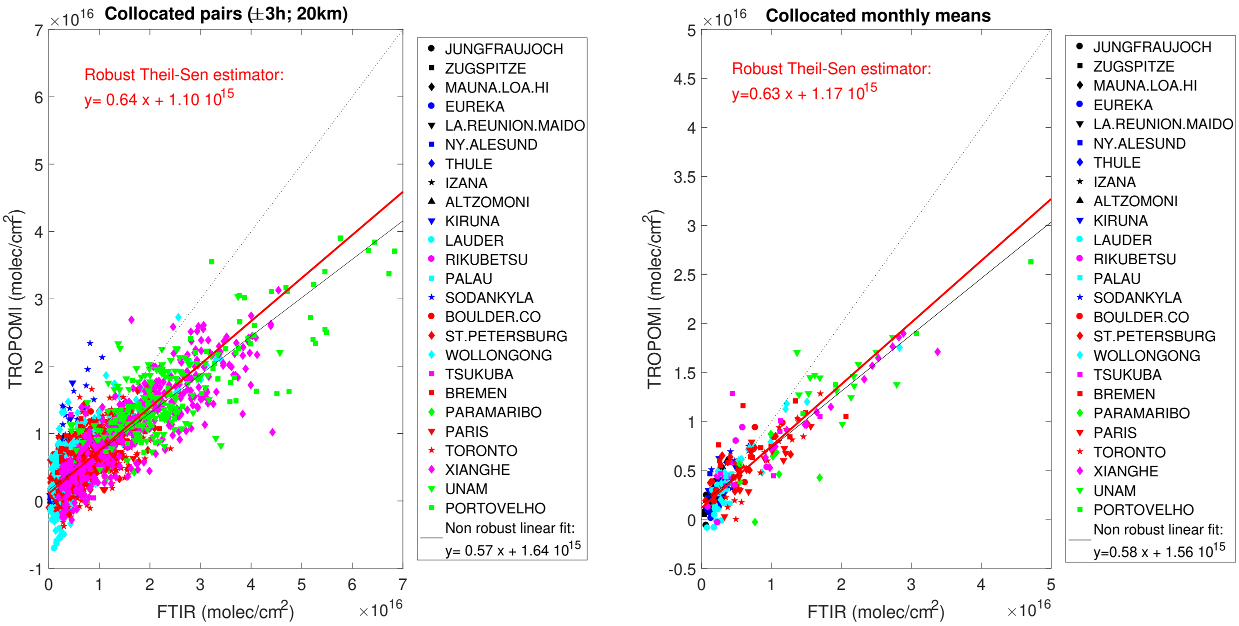

Figure 4. Scatter plots of TROPOMI versus FTIR data, for individual collocated pairs (±3 h; left panel), and for the monthly means of collocated data (right panel). The non robust OLS fit of the data is given in the legend and plotted as a black line, while the slope and intercept obtained with robust Theil-Sen estimator is given by the red line and text.

The systematic uncertainties leading to the observed constant and proportional biases of our study have been calculated as described in Sect. 3 of De Smedt et al. (2018). From the error propagation of the HCHO TROPOMI tropospheric columns (see Eq. 1), it can be found that the proportional bias is more likely due to air mass factor (M) uncertainties σ_M, while the constant bias is more likely due to the uncertainties of the slant columns uncertainties $\sigma_{N,S}$ and to the uncertainty of the background correction of the slant columns. This can be seen in Eq. 13 of De Smedt et al. (2018), where σ_M is proportional to $N_s - N_{(s,0)}$. We can list some known difficulties of the satellite product:

- The negative bias over high HCHO levels sites (biomass burning or mega-cities) could be due to aerosol effects. There is no plan to include a correction for aerosols in the operational product, but specific studies are foreseen to check its impact in a scientific product.
- The positive bias over clean polar sites could be due to the solar zenith angle (SZA) dependency of the slant columns fit results (because of spectral interferences with ozone and BrO). As explained in Sect. 2, the QA values need to be improved at large SZA, which is foreseen in the next version.
- The current albedo climatology is too coarse for TROPOMI, which could be especially a problem for polar, mountain or coastal sites. A climatology based on TROPOMI measurements is under development.

– It is also foreseen to test a regional model at higher spatial resolution for an improvement of the a priori HCHO profiles. This should improve the TROPOMI retrieved product, especially at polluted sites. However, the validation presented here is already taking the a priori information and averaging kernels into account. We therefore do not expect an important effect of the improved a priori profiles on the validation results.

5.2 Observed TROPOMI precision

For discussing the observed TROPOMI precision, we provide in Table 3, the MAD for each station (in absolute value to compare with the pre-launch precision requirement of 1.2×10^{16} molec/cm^2 for a single pixel, Eq. 7). The precision pre-launch requirement is provided at each site taking into account the mean number of pixels n$_{pix}$ involved in the collocated TROPOMI data (Requ.=1.2×10^{16} molec/cm$^2/\sqrt{n_{pix}}$). We see that for all the cleanest sites (<2.5×10^{15} molec/cm^2), where an additional collocation uncertainty is expected to be small, the MAD is well within the pre-launch requirements. The MAD for these cleanest sites has a median of 1.3×10^{15} molec/cm^2, and a minimum of 0.9×10^{15} molec/cm^2. This is a good estimate of the precision that TROPOMI can reach in remote conditions. For a single pixel, the TROPOMI best precision at remote conditions is therefore 5-8$\times10^{15}$ molec/cm^2.

It must be noted that the pre-launch HCHO precision requirements were chosen based on pre-launch requirements for the instrument signal to noise ratio (equivalent to OMI). The actual signal to noise of the measurements appears to be better than the requirements, especially in the HCHO wavelength fitting range. Furthermore, the good quality of the recorded spectra allowed to increase the size of the TROPOMI HCHO fitting spectral interval just after launch, further improving the precision of the slant columns. Indeed, as seen in Table 3, only at the three highest HCHO levels sites (Xianghe, Mexico City, and Porto Velho) the provided random uncertainties are as high as the pre-launch requirements. The actual provided random uncertainty are smaller, and we can see that, even for clean sites, the observed MAD is larger than the random uncertainty on the differences by a factor of 1.6. This factor increases up to 1.8 if we take into account all the stations, but this is expected due to a collocation uncertainty that should have more impact at high-levels sites (the factor rises up to 2.3 for high HCHO levels >8.0×10^{15} molec/cm^2). Our comparisons suggest that the TROPOMI random uncertainty is underestimated by at least a factor of 1.6 and up to maximum of 2.3 (if one would assume the collocation uncertainty to be smaller than the TROPOMI uncertainty). This underestimation could be due to the fact that currently the uncertainties associated to the air mass factor calculation and to the background correction step are assumed to be fully systematic. The discrimination between random and systematic part of the uncertainties might be refined in the future, based on such validation results.

5.3 Observed TROPOMI monthly variability

The Pearson correlation coefficient is very good for the collocated monthly means of TROPOMI and FTIR data (0.91, see Table 3 and Fig. 4), and is usually good for individual sites. However, Pearson correlation is not robust and can give a wrong conclusion when only few data are coincident, especially when outliers are present. We have 17 months of coincident TROPOMI and FTIR measurements in the best cases, while only 4 for the newest stations Palau and Porto Velho. We therefore verify that the TROPOMI precision allows the seasonal variability to be well captured, even at very clean sites which can be at the limit of the satellite detection, by plotting the individual monthly mean time-series in Fig. 5.

The seasonal variability, with a maximum in July-August, is well observed at all the Arctic sites (Eureka, Ny-Ålesund, Thule, Kiruna, and Sodankylä). The monthly mean correlation is better than 0.69, except at Eureka and Ny-Ålesund. It can be seen in Fig. 5 that Sept. 2019 is very high in TROPOMI data at Ny-Ålesund, and only 1 coincidence is found for this month. Removing this last outlier gives a 0.76 correlation coefficient at this station. The northern mid-latitude clean sites (mountains: Jungfraujoch, Zugspitze, Izaña) also display a seasonal variability in very good agreement, with correlation coefficients higher than 0.70. The Japanese clean site Rikubetsu shows poorer correlation (0.60) but only few data are in coincidence. The stations where we find the poorer correlations are the oceanic sites. The poorest one is Mauna Loa, but this is expected due to the very small seasonal variability there, associated to a small number of coincidences. A similar situation is observed at the other recent oceanic site Palau, where only 4 months of data are available. At the oceanic site Maïdo, we find a good agreement in most of the months but not in October-December, which are the predominant biomass burning months in the region so the collocation of the plumes might play a role there. Finally at Lauder, TROPOMI shows many negative values in the beginning of the period (May-Sept. 2018), which is responsible for a lower correlation (0.65) and for the negative bias there (although not significant), while other clean sites show usually positive ones (see Table 3).

The higher HCHO level sites show a TROPOMI seasonal variability in very good agreement with FTIR, with correlation larger than 0.90 for Boulder, Wollongong, Toronto, Xianghe, and Porto Velho. At Tsukuba, removing the outlier of 1 coincidence in November 2018 increases the correlation to 0.93. The poorest correlation (0.14) is found at the coastal site Paramaribo, where usually only one coincidence per month is found. Looking at the highest HCHO level sites, these monthly mean time-series also confirm that TROPOMI has more difficulty to reproduce the months with the highest enhancements, which is responsible for the significant negative bias (-31%) found in the previous section for high HCHO levels (>8.0×10^{15} molec/cm^2).

6 Conclusions

We have used a network of twenty-five FTIR stations, most of them affiliated to NDACC, to validate the latest TROPOMI HCHO tropospheric columns (v.1.1.[5-7]). This network covers a wide range of concentrations, from very clean Arctic, oceanic and mountain sites, with columns that can be lower than 10^{14} molec/cm^2, to high HCHO level sites such Mexico city or Porto Velho, near the Amazonian forest, where columns up to 7×10^{16} molec/cm^2 have been observed.

We found an overestimation (+26±5%) of TROPOMI OFFL products for very small HCHO columns (<2.5×10^{15} molec/cm^2) and an underestimation of TROPOMI of about -30.8% (±1.4%) for high HCHO columns (>8.0×10^{15} molec/cm^2), which can be used, e.g., to correct TROPOMI data near emissions sources. The results are very similar for NRTI products (+22 ±7% and -31.7±1.8% for small and high columns, respectively), and the differences are mainly due to the different period of available TROPOMI v.1.1.[5-7] products. Our wide range of HCHO levels and the use of the Theil-Sen method allow us to derive robust and significant constant (intercept) and proportional (slope) biases of TROPOMI (TROP=+ 1.10 (±0.05) ×10^{15}+ 0.64 (±0.03) × FTIR, in molec/cm^2). Such different BIAS for low/high target species concentration levels, due to the presence of

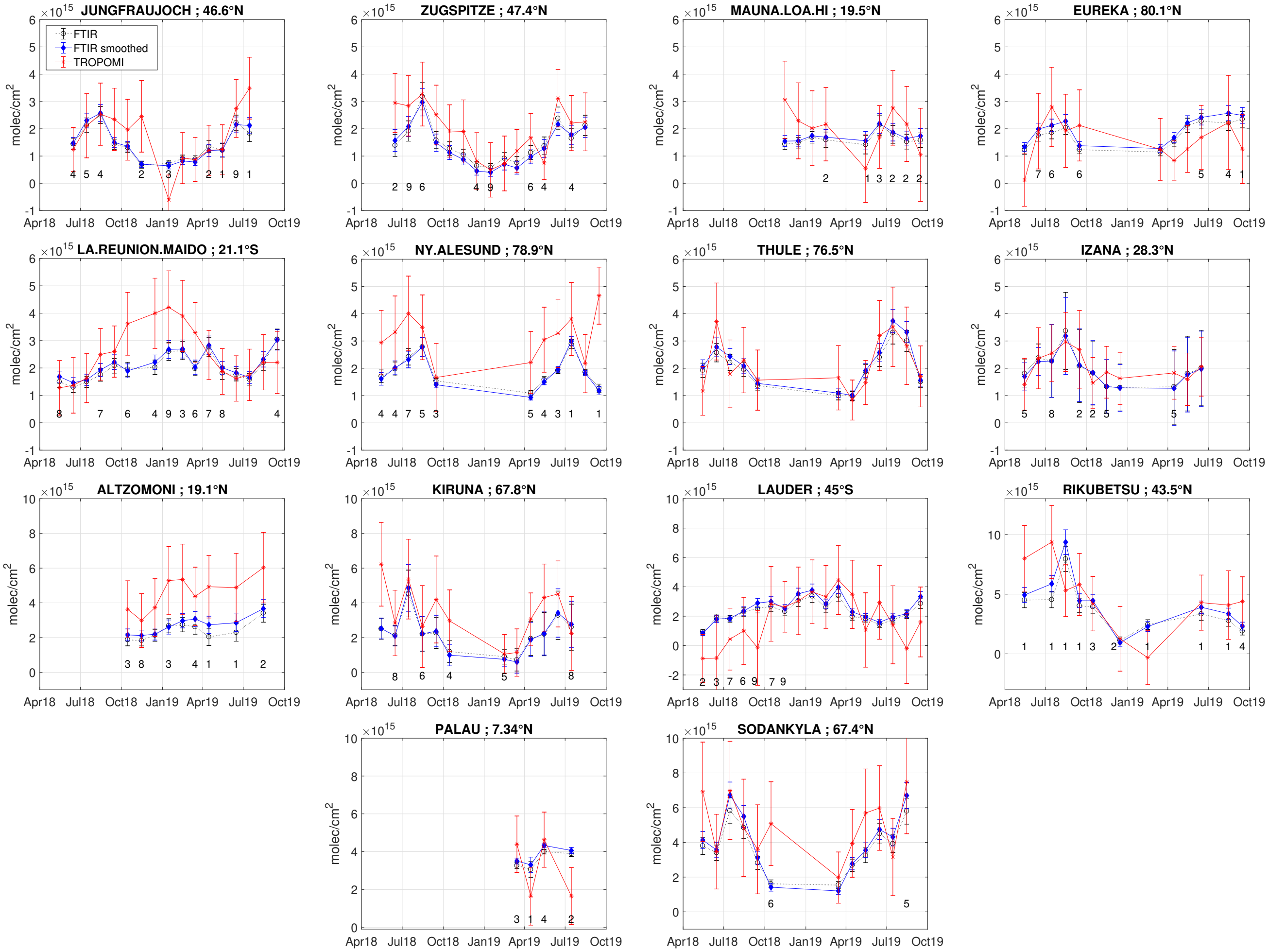

Figure 5. Monthly means time-series of FTIR raw data (black), FTIR data smoothed with the TROPOMI column averaging kernel (blue), and TROPOMI (red) at each site. Only data in coincidences are included in the monthly mean to avoid sampling bias. When the number of coincidences within one month is smaller than 10, it is written below the monthly mean.

both constant and proportional biases, was also recently observed (although with less FTIR sites involved) in another nadir satellite product, the formic acid observed by IASI (Supporting Information in Franco et al. (2020)). The NDACC FTIR network, which covers a large number of atmospheric species at wide ranges of concentrations, is a powerful source of reference data to detect such nadir satellites' biases.

Although significant, the observed overestimation and underestimation of TROPOMI are within the lower limits of the pre-launch requirements (±40%), as are the biases at individual sites for 20 of the 25 stations. The TROPOMI systematic uncertainty budget is in very good agreement with the observed bias, larger uncertainties being reported at stations where the bias exceeds the 40% requirements. Possible improvements in the TROPOMI biases could be achieved by taking into ac-

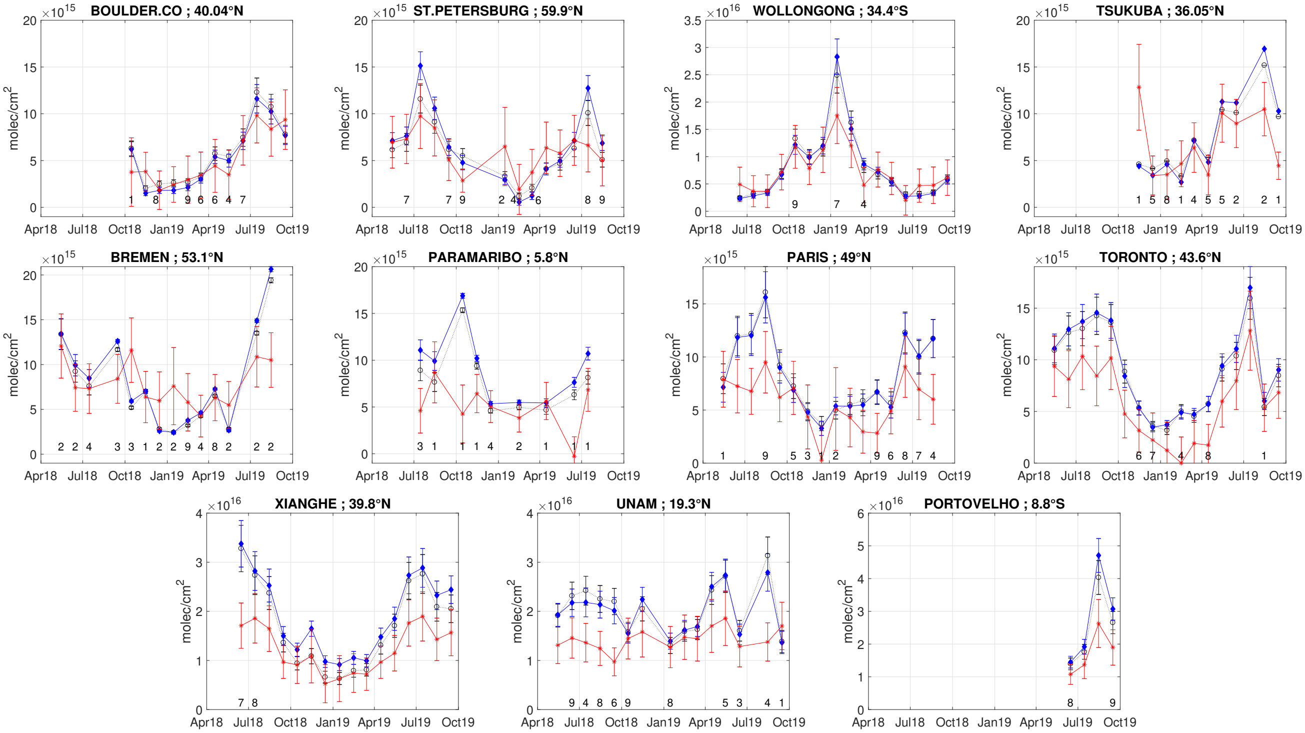

Figure 5. *Continued.*

count aerosol effects over polluted sites, improving the QA values at high SZA, and using an albedo climatology and a priori HCHO profiles at the TROPOMI spatial resolution. Except for the former, these improvements are foreseen in next versions of operational TROPOMI data.

The precision of TROPOMI OFFL products is estimated by the median absolute deviation (MAD) at the clean sites, where the collocation effect is expected to be small. For FTIR HCHO levels lower than 2.5×10^{15} molec/cm^2, the MAD is 1.3 ×10^{15} molec/cm^2, corresponding to a single pixel precision of 7×10^{15} molec/cm^2 (5 to 8×10^{15} at individual sites), which is well below the pre-launch precision requirements of 1.2×10^{16} molec/cm^2. However, the provided TROPOMI random uncertainties (after launch) were indeed found to be better than the pre-launch requirements, but they are too small by a factor of 1.6 compared to the MAD at the clean sites. There is a factor of 2.3 difference between MAD and the random uncertainty on the comparisons (dominated by TROPOMI random uncertainty) at the high-level sites, where an additional effect of collocation might take a role as well. The underestimation of the TROPOMI random uncertainty could be due to a random effect of the uncertainty associated to the air mass factor calculation that is not currently included in the budget. This would also explain a larger underestimation of random error at high-levels sites (factor 2.3 vs 1.6 at clean sites). Also, a systematic uncertainty component on a short-term (so not included in the TROPOMI random uncertainty) can have a random effect on our longer-term comparisons.

We have shown that the TROPOMI data capture very well the HCHO seasonal variability, even at very clean sites. The Pearson correlation coefficient for monthly mean coincident data is 0.91. Although we have found room for a refinement of the TROPOMI random uncertainty estimation and for an improvement of the QA values for a better filtering of the remaining few outliers and negative columns (exceeding the expected statistical distribution), this validation work has demonstrated the very good quality of the TROPOMI HCHO product, which is well within the pre-launch requirements for both accuracy and precision. This work has also shown the high value of the FTIR HCHO network, providing harmonized and well-characterized data covering a wide range of HCHO columns. These ground-based FTIR data are continuously extended by new measurements and will be used in the coming years for the routine S5P validation within the ESA dedicated validation server (https://mpc-vdaf-server.tropomi.eu/). The FTIR network will also be used in the near future for the validation of previous satellite missions such as OMI or GOME-2. New FTIR measurements are continuously performed and can be used in the coming years for the validation of new satellite generation, such as TEMPO, GEMS, Sentinel 5P, or Sentinel 4.

An extension of this TROPOMI HCHO validation with ground-based MAX-DOAS and Pandora instruments, especially at sites where both FTIR and UV-Visible techniques are available (e.g. Xianghe, Maïdo, Lauder,...) or at uncovered regions (Africa) would bring additional knowledge. However, there is first a need for a data product harmonization within the MAXDOAS network, as was done with the FTIR network used here. This work is ongoing as part of the ESA FRM4DOAS and Pandonia projects.

Data availability:

The TROPOMI HCHO data are publicly available at https://scihub.copernicus.eu. The access and use of any Copernicus Sentinel data available through the Copernicus Sentinel Data Hub is governed by the Legal Notice on the use of Copernicus Sentinel Data and Service Information which is given here:
https://sentinels.copernicus.eu/documents/247904/690755/Sentinel_Data_Legal_Notice.

The FTIR data sets can be provided in the public NDACC repository (ftp://ftp.cpc.ncep.noaa.gov/ndacc/station/, last access: January 2020) depending on each PI decision. Please pay attention to the NDACC data policy. The whole data set used in this publication can be provided upon request to Corinne Vigouroux (corinne.vigouroux@aeronomie.be) and data per station can be requested from the individual PIs.

Author contribution: Corinne Vigouroux and Bavo Langerock performed the validation using HCHO TROPOMI and FTIR data at all sites. They are also involved in the FTIR measurements at Maïdo and Porto Velho. Corinne Vigouroux analyzed the Maïdo, Porto Velho, Sodankylä and Xiangue data. Isabelle De Smedt is the TROPOMI HCHO product lead and participated in the paper (Section 2 and discussions). Michel Van Roozendael and Zhibin Cheng have also responsibility for the TROPOMI HCHO prototype algorithm and operational processor, respectively. Gaia Pinardi was involved in the validation method section through her expertise in validation using UV-Visible techniques, which is part of the projects TROVA and TROVA-2 that funded this work. All other co-authors provided the FTIR HCHO data for the station(s) they are responsible for.

Competing interests: The authors declare that they have no conflict of interest.

Acknowledgements. This study has been supported by the ESA PRODEX projects TROVA and TROVA-E2 funded by the Belgian Science Policy Office (Belspo). The measurements at Reunion Island have been also supported by the Université de La Réunion and CNRS (LACy-UMR8105 and UMS3365), and at Porto Velho by the BRAIN-pioneer project IKARE, funded by Belspo, with the collaboration of the Instituto Federal de Educaçao, Ciência e Tecnologia de Rondônia (IFRO). SPbU FTIR team has been supported by the Russian Foundation for Basic Research Project #18-05-00011. St. Petersburg FTIR measurements were carried out by the instrumentation of the Geomodel resource center of SPbU. The NDACC stations Bremen, Izaña, Ny-Ålesund and Paramaribo have been supported by the German Bundesministerium für Wirtschaft und Energie (BMWi) via DLR under grants 50EE1711A, B and D. We thank the EU-project STRATOCLIM for financial support to U. of Bremen. Measurements made at Lauder, by NIWA, are funded by New Zealand's Ministry of Business, Innovation and Employment through the Strategic Science Investment Fund. We thank the AWI Bremerhaven for logistical support and the station personnel in Ny Alesund. We acknowlegde the Meterologische Dienst van Suriname for logistical and on-site support in Paramaribo, Suriname. We acknowledge the AWI Potsdam and the Coral Reef Foundation for logistical and on-site support in Koror, Palau. ULiège has received support from the F.R.S. - FNRS, from the Fédération Wallonie-Bruxelles and from the GAW-CH programme of MeteoSwiss. We further acknowledge the International Foundation High Altitude Research Stations Jungfraujoch and Gornergrat (HFSJG, Bern) for supporting the facilities needed to perform the Jungfraujoch observations. E. Mahieu is Research Associate with the F.R.S. - FNRS. The National Center for Atmospheric Research is sponsored by the National Science Foundation. The NCAR FTS observation programs at Thule, GR, Boulder, CO, and Mauna Loa, HI are supported under contract by the National Aeronautics and Space Administration (NASA). The Thule work is also supported by the NSF Office of Polar Programs (OPP). We wish to thank the Danish Meteorological Institute for support at the Thule site and NOAA for support at the Mauna Loa. Financial support from grants by DGAPA-UNAM (07417 & 111418) and CONACYT (290589) are acknowledged as well as the University Network of Atmospheric Observatories (RUOA) for the maintenance and operation of the Mexican stations. The Paris TCCON site has received fundings from Sorbonne Université, the French research center CNRS, the French space agency CNES and Région Île-de-France. Eureka measurements were made at the Polar Environment Atmospheric Research Laboratory (PEARL) under the CANDAC and PAHA projects led by James R. Drummond, and in part by the Canadian Arctic ACE/OSIRIS Validation Campaigns, led by Kaley A. Walker. Funding was provided by AIF/NSRIT, CFI, CFCAS, CSA, ECCC, GOC-IPY, NSERC, NSTP, OIT, PCSP, and ORF. Logistical and operational support was provided by PEARL Site Manager Pierre Fogal, the CANDAC operators, and the ECCC Weather Station. Toronto measurements were made at the University of Toronto Atmospheric Observatory, supported by CFCAS, ABB Bomem, CFI, CSA, ECCC, NSERC, ORDCF, PREA, and the University of Toronto. FTIR operations of the Rikubetsu and Tsukuba sites are financially supported in part by the GOSAT series project.

The authors would like to thank D. Loyola for the operational TROPOMI HCHO data, and all the people responsible for the FTIR measurements and/or data analysis at the different sites: C. Hermans, N. Kumps, M. Zhou, BIRA-IASB; L. Gatti, INPE; Uwe Raffalski, IRF Kiruna; Omaira García and Eliezer Sepulveda, AEMet; C. Becker, Paramaribo; J. Robinson, Lauder; A.V. Poberovskii, H.H. Imkhasin, S.I. Osipov, SPbU; A. Bezanilla and C. Guarín, UNAM; P. Jeseck, Sorbonne Université; M. Rettinger, IMK-IFU; H. Nakajima, NIES; C. Servais: Université de Liège.

References

Boersma, K. F., Eskes, H. J., and Brinksma, E. J.: Error analysis for tropospheric NO2 retrieval from space, J. Geophys. Res., 109, doi:10.1029/2003JD003962, 2004.

De Smedt, I., Stavrakou, T., Hendrick, F., Danckaert, T., Vlemmix, T., Pinardi, G., Theys, N., Lerot, C., Gielen, C., Vigouroux, C., Hermans, C., Fayt, C., Veefkind, P., Müller, J.-F., and Van Roozendael, M.: Diurnal, seasonal and long-term variations of global formaldehyde columns inferred from combined OMI and GOME-2 observations, Atmos. Chem. Phys., 15, 12519-12545, doi:10.5194/acp-15-12519-2015, 2015.

De Smedt, I., Theys, N., Yu, H., Danckaert, T., Lerot, C., Compernolle, S., Van Roozendael, M., Richter, A., Hilboll, A., Peters, E., Pedergnana, M., Loyola, D., Beirle, S., Wagner, T., Eskes, H., van Geffen, J., Boersma, K. F., and Veefkind, P.: Algorithm theoretical baseline for formaldehyde retrievals from S5P TROPOMI and from the QA4ECV project, Atmos. Meas. Tech., 11, 2395–2426, https://doi.org/10.5194/amt-11-2395-2018, 2018.

Fortems-Cheiney, A., Chevallier, F., Pison, I., Bousquet, P., Saunois, M., Szopa, S., Cressot, C., Kurosu, T. P., Chance, K., and Fried, A.: The formaldehyde budget as seen by a global-scale multi-constraint and multi-species inversion system, Atmos. Chem. Phys., 12, 6699-6721, doi:10.5194/acp-12-6699-2012, 2012.

Franco, B., Clarisse, L., Stavrakou, T., Müller, J.-F., Taraborrelli, D., Hadji-Lazaro, J., Hannigan, J.W., Hase, F., Hurtmans, D., Jones, N., Lutsch, E., Mahieu, E., Ortega, I., Schneider, M., Strong, K., Vigouroux, C., Clerbaux, C., Coheur, P.-F.: Spaceborne Measurements of Formic and Acetic Acid: A Global View of the Regional Sources, Geophysical Research Letters, 47, e60094., https://doi.org/10.1029/2019GL086239, 2020.

Garcia, R. R., Marsh, D. R., Kinnison, D. E., Boville, B. A., and Sassi, F.: Simulation of secular trends in the middle atmosphere, 1950-2003, J. Geophys. Res., 112, D09301, doi:10.1029/2006JD007485, 2007.

Hase, F., Hannigan, J. W., Coffey, M. T., Goldman, A., Höpfner, M., Jones, N. B., Rinsland, C. P., and Wood, S. W.: Intercomparison of retrieval codes used for the analysis of high-resolution, ground-based FTIR measurements, J. Quant. Spectros. Radiat. Transfer, 87, 25-52, 2004.

Hase, F., Demoulin, P., Sauval, A. J., Toon, G. C., Bernath, P. F., Goldman, A., Hannigan, J. W., and Rinsland, C. P.: An empirical line-by-line model for the infrared solar transmittance spectrum from 700 to 5000 cm^{-1}, J. Quant. Spectros. Radiat. Transfer, 102, 450-463, 2006.

Huber, P.J., Robust Statistics, Wiley, New York, 1981.

Jacquemart, D., Laraia, A., Kwabia Tchana, F., Gamache, R. R., Perrin, A., and Lacome, N: Formaldehyde around 3.5 and 5.7 μm: measurement and calculation of broadening coefficients, J. Quant. Spectrosc. Radiat. Transfer, 111, 1209-1222, 2010.

Jones, N. B., Riedel, K., Allan, W., Wood, S., Palmer, P. I., Chance, K., and Notholt, J.: Long-term tropospheric formaldehyde concentrations deduced from ground-based fourier transform solar infrared measurements, Atmos. Chem. Phys., 9, 7131-7142, doi:10.5194/acp-9-7131-2009, 2009.

Kleipool, Q. L., Dobber, M. R., de Haan, J. F., and Levelt, P. F.: Earth surface reflectance climatology from 3 years of OMI data, J. Geophys. Res., 113, D18308, doi:10.1029/2008JD010290, 2008.

Langerock, B., De Mazière, M., Hendrick, F., Vigouroux, C., Desmet, F., Dils, B., and Niemeijer, S.: Description of algorithms for co-locating and comparing gridded model data with remote-sensing observations, Geosci. Model Dev., 8, 911–921, https://doi.org/10.5194/gmd-8-911-2015, 2015.

Loyola, D. G., Gimeno García, S., Lutz, R., Argyrouli, A., Romahn, F., Spurr, R. J. D., Pedergnana, M., Doicu, A., Molina García, V., and Schüssler, O.: The operational cloud retrieval algorithms from TROPOMI on board Sentinel-5 Precursor, Atmos. Meas. Tech., 11, 409–427, https://doi.org/10.5194/amt-11-409-2018, 2018.

Martin, R. V., Parrish, D. D., Ryerson, T. B., Nicks, D. K. Jr., Chance, K., Kurosu, T. P., Jacob, D. J., Sturges, E. D., Fried, A., and Wert, B. P.: Evaluation of GOME satellite measurements of tropospheric NO2 and HCHO using regional data from aircraft campaigns in the southeastern United States, J. Geophys. Res., 109, D24307, doi:10.1029/2004JD004869, 2004.

Meller, R. and Moortgat, G. K.: Temperature dependence of the absorption cross section of HCHO between 223 and 323 K in the wavelength range 225–375 nm, J. Geophys. Res., 105, 7089-7102, doi: 10.1029/1999JD901074, 2000.

Pougatchev, N. S., Connor, B. J., and Rinsland, C. P.: Infrared measurements of the ozone vertical distribution above Kitt Peak, J. Geophys. Res., 100, 16689–16697, 1995.

Peters, E., Wittrock, F., Großmann, K., Frieß, U., Richter, A., and Burrows, J. P.: Formaldehyde and nitrogen dioxide over the remote western Pacific Ocean: SCIAMACHY and GOME-2 validation using ship-based MAX-DOAS observations, Atmos. Chem. Phys., 12, 11179–11197, https://doi.org/10.5194/acp-12-11179-2012, 2012.

Rodgers, C. D.: Inverse methods for atmospheric sounding: Theory and Practice, Series on Atmospheric, Oceanic and Planetary Physics - Vol. 2, World Scientific Publishing Co., Singapore, 2000.

Rodgers, C. D. and Connor, B. J.: Intercomparison of remote sounding instruments, J. Geophys. Res. 108, 4116-4129, 2003.

Rothman, L. S., Gordon, I. E., Babikov, Y., Barbe, A., Benner, D. C., Bernath, P. F., Birk, M., Bizzocchi, L., Boudon, V., Brown, L. R., Campargue, A., Chance, K., Cohen, E. A., Coudert, L. H., Devi, V. M., Drouin, B. J., Fayt, A., Flaud, J.-M., Gamache, R. R., Harrison, J. J., Hartmann, J.-M., Hill, C., Hodges, J. T., Jacquemart, D., Jolly, A., Lamouroux, J., Roy, R. J. L., Li, G., Long, D. A., Lyulin, O. M., Mackie, C. J., Massie, S. T., Mikhailenk, S., Müller, H. S. P., Naumenko, O. V., and Nikitin, A. V.: The HITRAN2012 molecular spectroscopic database, J. Quant. Spectrosc. Ra., 130, 4-50, doi:10.1016/j.jqsrt.2013.07.002, 2013.

Sen, Pranab Kumar: Estimates of the regression coefficient based on Kendall's tau, Journal of the American Statistical Association, 63 (324): 1379–1389, doi:10.2307/2285891, 1968.

Spurr, R. J. D.: LIDORT and VLIDORT: Linearized pseudo-spherical scalar and vector discrete ordinate radiative transfer models for use in remote sensing retrieval problems, in Light Scattering Reviews, edited by: Kokhanovsky, A., 229-271, Berlin, 2008.

Stavrakou, T., Müller, J.-F., Bauwens, M., De Smedt, I., Van Roozendael, M., De Mazière, M., Vigouroux, C., Hendrick, F., George, M., Clerbaux, C., Coheur, P.-F., and Guenther, A.: How consistent are top-down hydrocarbon emissions based on formaldehyde observations from GOME-2 and OMI?, Atmos. Chem. Phys., 15, 11861-11884, doi:10.5194/acp-15-11861-2015, 2015.

Tan, W., Liu, C., Wang, S., Xing, C., Su, W., Zhang, C., Xia, C., Liu, H., Cai, Z., and Liu, J.: Tropospheric NO_2, SO_2, and HCHO over the East China Sea, using ship-based MAX-DOAS observations and comparison with OMI and OMPS satellite data, Atmos. Chem. Phys., 18, 15387–15402, https://doi.org/10.5194/acp-18-15387-2018, 2018.

Tikhonov, A.: On the solution of incorrectly stated problems and a method of regularization, Dokl. Acad. Nauk SSSR, 151, 501–504, 1963.

Veefkind, J.P., Aben, I., McMullan, K., Förster, H., De Vries, J., Otter, G., Claas, J., Eskes, H.J., De Haan, J.F., Kleipool, Q., Van Weele, M., Hasekamp, O., Hoogeveen, R., Landgraf, J., Snel, R., Tol, P., Ingmann, P., Voors, R., Kruizinga, B., Vink, R., Visser, H., and Levelt, P.F.: TROPOMI on the ESA Sentinel-5 Precursor: A GMES mission for global observations of the atmospheric composition for climate, air quality and ozone layer applications, Remote Sensing of Environment, 120, 70-83, doi:10.1016/j.rse.2011.09.027, 2012.

Vigouroux, C., Hendrick, F., Stavrakou, T., Dils, B., De Smedt, I., Hermans, C., Merlaud, A., Scolas, F., Senten, C., Vanhaelewyn, G., Fally, S., Carleer, M., Metzger, J.-M., Müller, J.-F., Van Roozendael, M., and De Mazière; M.: Ground-based FTIR and MAX-DOAS observations of formaldehyde at Reunion Island and comparisons with satellite and model data, Atmos. Chem. Phys., 9, 9523–9544, 2009.

Vigouroux, C., Bauer Aquino, C. A., Bauwens, M., Becker, C., Blumenstock, T., De Mazière, M., García, O., Grutter, M., Guarin, C., Hannigan, J., Hase, F., Jones, N., Kivi, R., Koshelev, D., Langerock, B., Lutsch, E., Makarova, M., Metzger, J.-M., Müller, J.-F., Notholt, J., Ortega, I., Palm, M., Paton-Walsh, C., Poberovskii, A., Rettinger, M., Robinson, J., Smale, D., Stavrakou, T., Stremme, W., Strong, K., Sussmann, R., Té, Y., and Toon, G.: NDACC harmonized formaldehyde time series from 21 FTIR stations covering a wide range of column abundances, Atmos. Meas. Tech., 11, 5049–5073, https://doi.org/10.5194/amt-11-5049-2018, 2018.

von Clarmann, T.: Smoothing error pitfalls, Atmos. Meas. Tech., 7, 3023-3034, https://doi.org/10.5194/amt-7-3023-2014, 2014.

Williams, J. E., Boersma, K. F., Le Sager, P., and Verstraeten, W. W.: The high-resolution version of TM5-MP for optimized satellite retrievals: description and validation, Geosci. Model Dev., 10, 721-750, doi:10.5194/gmd-10-721-2017, 2017.

Wittrock, F., Richter, A., Oetjen, H., Burrows, J. P. , Kanakidou, M., Myriokefalitakis, S., Volkamer, R., Beirle, S., Platt, U., and Wagner, T.: Simultaneous global observations of glyoxal and formaldehyde from space, Geophys. Res. Lett., 33, L16804, doi:10.1029/2006GL026310, 2006.

Zhu, L., Jacob, D. J., Kim, P. S., Fisher, J. A., Yu, K., Travis, K. R., Mickley, L. J., Yantosca, R. M., Sulprizio, M. P., De Smedt, I., González Abad, G., Chance, K., Li, C., Ferrare, R., Fried, A., Hair, J. W., Hanisco, T. F., Richter, D., Jo Scarino, A., Walega, J., Weibring, P., and Wolfe, G. M.: Observing atmospheric formaldehyde (HCHO) from space: validation and intercomparison of six retrievals from four satellites (OMI, GOME2A, GOME2B, OMPS) with SEAC^{4}RS aircraft observations over the southeast US, Atmos. Chem. Phys., 16, 13477-13490, doi:10.5194/acp-16-13477-2016, 2016.

Zhu, L., González Abad, G., Nowlan, C. R., Chan Miller, C., Chance, K., Apel, E. C., DiGangi, J. P., Fried, A., Hanisco, T. F., Hornbrook, R. S., Hu, L., Kaiser, J., Keutsch, F. N., Permar, W., St. Clair, J. M., and Wolfe, G. M.: Validation of satellite formaldehyde (HCHO) retrievals using observations from 12 aircraft campaigns, Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2019-1117, in review, 2020.

Reply to Anonymous Referee #1

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 this stations belong to the Network for the Detection of Atmospheric Composition Chance (NDACC).They 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. 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 publish 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.

We thank the referee for their very positive review and for their work that is helping us to improve the manuscript.

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?

The TROPOMI accuracy pre-launch requirements are given as a range: "40-80%". We have distinguished between the two limits of the ranges because at all sites the 80% requirements are reached (but this is an upper limit for the expected TROPOMI accuracy), and at 20 of the 25 sites the lower limit of the range (40%) is reached. Therefore, at 5 sites, we have a bias between 40 and 80%. To avoid any confusion, we have rewritten it as follows:

"The pre-launch requirements of the TROPOMI HCHO accuracy are 40-80%. We observe that these requirements are well reached, with the BIAS found below 80% at all the sites, and below 40% at 20 of the 25 sites."

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 valuable to add it to the list of aircraft based validation efforts?

Indeed. This reference has been added in the manuscript.

TROPOMI HCHO data.

The description of TROPOMI data and versions is very complete but after reading this section the question remains, off 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?

The text in our AMTD version was:
"At the time of writing this paper, the latest product versions 1.1.[5-7] provide a consistent time series of Reprocessed+Offline (RPRO+OFFL) data, covering the period between May 2018 up to (at least) December 2019 (last access). The Near-Real-Time (NRTI) product, for the same versions 1.1.[5-7], started in December 2018. Details are found in the Readme file (http://www.tropomi.eu/sites/default/files/files/publicSentinel-5P-Formaldehyde-Readme.pdf; doi: 10.5270/S5P-tjlxfd2)."

Indeed, the referee is right: it is not clear in this TROPOMI section which products are used in this paper (RPRO + OFFL, or NRTI). Actually, we performed the validation on the two sets of data. But, in this paper the tables and figures focus on the RPRO+OFFL data set. The NRTI validation results are so similar that we preferred avoiding giving details on them. We only give a summary of the NRTI biases in Sect. 5.1.

At all sites, the TROPOMI data set that we used is a combination of RPRO and OFFL products, from v.1.1.5 to 1.1.7, the versions 5 to 7 being consistent retrieved HCHO products. Indeed, the number of version corresponds to different period of time, but we did not find relevant to detail them since the products are consistent among these versions. The details of the dates are in the Readme file (more precisely in its Table 2) for which we gave the reference. For the referee and readers convenience, we provide them here, and *we will repeat them in a Table in the next version of the manuscript:*

- *From 2018-05-14 to 2018-11-28 : RPRO v.1.1.5*
- *From 2018-11-28 to 2019-03-28 : OFFL v.1.1.5*
- *From 2019-03-28 to 2019-04-23 : OFFL v.1.1.6*
- *From 2019-04-23 to present : OFFL v.1.1.7*

The validation of the NRTI products (results only summarized in one sentence in Sect. 5.1) is using:

- From 2018-12-05 to 2019-04-04 : NRTI v.1.1.5
- From 2019-04-04 to 2019-04-30 : NRTI v.1.1.6
- From 2019-04-30 to present : NRTI v.1.1.7

We have also repeated in the new table (on request of referee#2), the information on the differences between the versions that is in the ReadMe file.

We have added to the manuscript (in italic):
"At the time of writing this paper, the latest product versions 1.1.[5-7] provide a consistent time series of Reprocessed+Offline (RPRO+OFFL) data, covering the period between May 2018 up to (at least) December 2019 (last access). *The detailed validation results shown in Sect. 5 are obtained using this consistent time-series (RPRO+OFFL, from 2018-05-14 to 2019-12-31). The version numbers and their dates of change are given in Table 1, and further details are given in the Readme file* (http://www.tropomi.eu/sites/default/files/files/publicSentinel-5P-Formaldehyde-Readme.pdf; doi: 10.5270/S5P-tjlxfd2). The Near-Real-Time (NRTI) product, for the same versions 1.1.[5-7], started in December 2018 *up to December 2019 (last access).* *This product has also been validated, but the results being very similar to the RPRO+OFFL validation, we do not show them in details in this paper."*

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?

Yes, we agree that topography could play a significant role if not taken into account carefully, both for the quality of the product, and for the comparison between satellite and ground-based quantities. However, we considered it in both cases.
For S5P L2 products, the digital elevation map is from GMTED2010 (Danielson et al., 2011), and an average over the ground pixel area is considered. Furthermore, as explained in the HCHO the Algorithm Theoretical Basis Document (ATBD, De Smedt et al. 2018): "To reduce the errors associated to topography and the lower spatial resolution of the model compared to the TROPOMI 3.5x7 km2 spatial resolution, the a priori profiles need to be rescaled to effective surface elevation of the satellite pixel. The TM5 surface pressure is converted by applying the hypsometric equation and the assumption that temperature changes linearly with height"
Finally, as described in Sect.4.2, the different elevation between the altitude of the ground-based station and the surface elevation of the satellite pixel is taken into account. We believe that the positive bias usually observed at mountain stations is related to the constant bias of TROPOMI for small HCHO columns, because it is also observed at clean sites that have an altitude close to sea level (Kiruna, Ny-Alesund).

Danielson, J.J., and Gesch, D.B.: Global multi-resolution terrain elevation data 2010 (GMTED2010): U.S. Geological Survey Open-File Report 2011–1073, 26 p, 2011.

De Smedt, I., Theys, N., Yu, H., Danckaert, T., Lerot, C., Compernolle, S., Van Roozendael, M., Richter, A., Hilboll, A., Peters, E., Pedergnana, M., Loyola, D., Beirle, S., Wagner, T., Eskes, H., van Geffen, J., Boersma, K. F., and Veefkind, P.: Algorithm theoret10 ical baseline for formaldehyde retrievals from S5P TROPOMI and from the QA4ECV project, Atmos. Meas. Tech., 11, 2395–2426, https://doi.org/10.5194/amt-11-2395-2018, 2018.

Page 4, line 6. "All cross-sections have been pre-convolved", these cross-sections 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?

Together with the HCHO cross-section, the absorptions of NO2, BrO, O3 (at two temperatures) and O4 are fitted. A Ring cross-section and two pseudo-cross sections to account for non-linear O3 absorption effects are also included in the fit. References are given in De Smedt et al. (2018). *This more detailed description has been added in the new manuscript.*

The operational algorithm does not have the capability to fit directly the slit functions, it has to be done offline. Up to now, the TROPOMI slit functions have been stable. No update of the pre-convolved cross-sections are planned, but this is monitored.

Page 4, line 20. How is M0 calculated? Is it the average of the AMFs of the slant columns considered in the calculation of N(s,0)?

Yes; M0 is an average of the air mass factors (M) of the slant columns selected in the reference sector, the Pacific Ocean (N(s,0)).
This has been added in the text.

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?

The spatial resolution used for this map is 0.2°x0.2°. We use the HARP v1.5 tool, which can be found at https://atmospherictoolbox.org. This information has been added in the Fig.1 caption, as suggested by the referee.

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)”

Done, as suggested.

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?

The channeling is due to (possible) imperfections in the instrument that may (or may not) lead to artefacts in the interferogram. This error is included in the PROFFIT9 code, and not yet in SFIT4. However, at present the fact that there is or not a channeling in the spectra at each station (it is obviously depending on each instrument) has not been measured at each site. Such an exercise has been initiated after the Vigouroux et al. (2018) paper for a set of stations (by T. Blumenstock, KIT, co-author of the present paper), but has not been done at each site systematically. For the sites that have been tested, we found that a non-negligible channeling is indeed present at some sites, but not at all sites. Therefore, introducing such an additional error in the theoretical calculation without knowing if it is indeed present may also lead to an overestimation of the systematic uncertainty. In the next update version of SFIT4, the random and systematic error on the target species due to channeling will be included, but its correct estimation would be possible only at the sites where the channeling itself is estimated. This is an on-going work within the IRWG (InfraRed Working Group) of NDACC.

In the present validation, the systematic bias between TROPOMI and FTIR stations are very consistent among the stations (see Fig. 3), except for Eureka which is the only clean site with a negative TROPOMI BIAS. However, Eureka was one of the sites participating on the channeling exercise, and the channeling was found very small for this instrument. So the channeling error is not explaining the different bias there. For the other stations, the good consistency of the TROPOMI BIAS at the different stations (which depends on the HCHO levels, and not on individual sites), shows that the BIAS is dominated by the TROPOMI systematic error, and that the channeling one should have a smaller impact.

To clarify that the channeling is not always under-estimated in the SFIT4 stations, and can be over-estimated in some PROFFIT4 stations, we have adapted the text:

“The systematic uncertainty can be larger (up to 21-26%) at the stations using the PROFFIT9 retrieval code, due to an assumed uncertainty on the channeling that is not taken into account yet in the SFIT4 code. However, this channeling uncertainty can also be negligible at some sites (it depends on each instrument), and more investigation is needed at each station to avoid its under- or over-estimation.”

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

We think the referee has misinterpreted the sentence. The 13% and 14% for the 5 added sites, are for the total systematic uncertainty (dominated by the spectroscopy), and not for the smoothing part only. To avoid the confusion, we have changed the sentence to :

“For the five added sites, the median total systematic uncertainty is 13% (Jungfraujoch, Tsukuba, Palau), or 14% (Rikubetsu, Xianghe), commensurate with the other 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?

Before choosing the 20km collocation radius, we have indeed tested several distances: 10, 20, 30, 40, and 50 km. We provide in this discussion a plot of the median relative differences (bias) at each station (Fig.1) for the different collocation distances. Please, note that the numbers are not the same as in the AMTD paper, because this work on collocation distances were made in the course of the project (not at the time of writing the paper), so the time-series were shorter, and the collocated time was 6h (now it is set to 3h). We see in Fig. 1, usually similar biases for the 20 to 50 km criteria, especially for mid- HCHO levels sites. For clean sites, we observe usually slightly smaller biases with the 30km criteria than with the 20km one. For the most polluted sites, UNAM (Mexico City) and Porto Velho, the bias is increasing with the distance. The 10km collocation leads to more than twice less coincidences (at some stations, even 5 times less).

Therefore, the median biases obtained with this criterion were less robust, and the 10km choice was discarded.

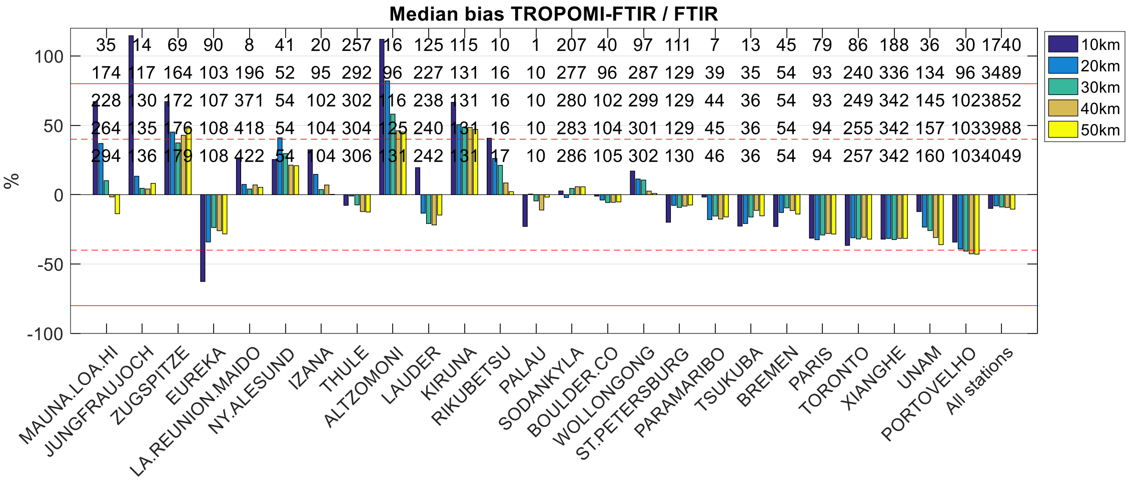

Figure 1: Median bias at each station for the different collocation distances. The numbers in black are the number of coincidences, from the 10km criterion (top) to the 50km criterion (bottom).

The median biases, being usually similar using the different collocation distances, were not so useful to determine our choice of collocation. We therefore looked at the MAD (median absolute deviation, see Eq. 6 for complete definition) to help for the choice. Figure 2 shows the MAD at each station for the different collocation distance.

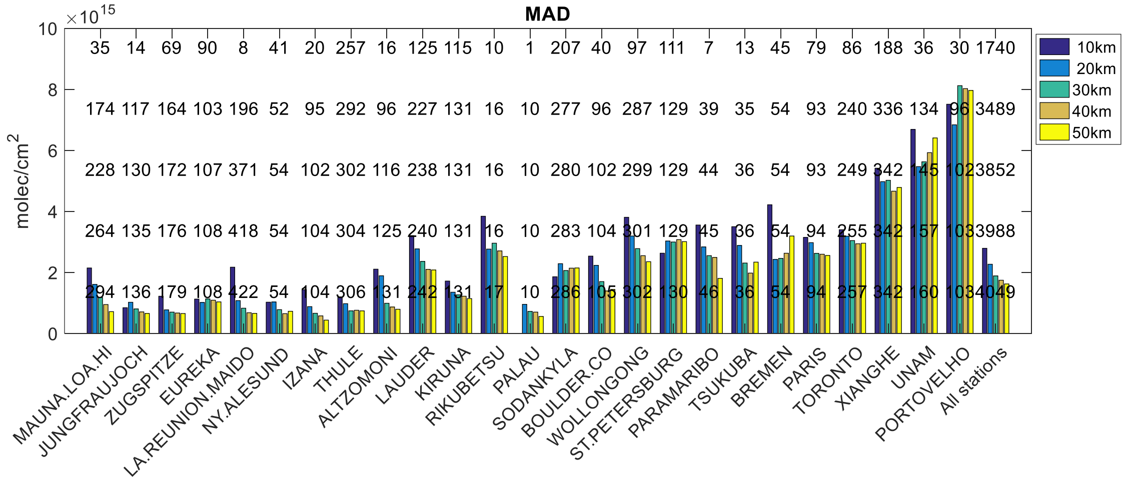

Figure 2: MAD at each station for the different collocation distances. The numbers in black are the number of coincidences, from the 10km criterion (top) to the 50km criterion (bottom).

We see from the figure that usually the MAD is decreasing with the distance increasing, except at a few cases (the polluted cases as expected: Porto Velho, UNAM=Mexico City,…). However, we cannot conclude that the comparisons are “improved”: indeed, while the MAD is decreasing due to the averaging of more TROPOMI pixels, the random uncertainties of the comparisons are also decreasing. In a world where the random error would be perfectly determined, we would have a constant ratio MAD / RandErr (no dependence on the collocation distance), equal to 1 if there is no collocation effect (so expected to be 1 at clean sites). If we plot this ratio (Figure 3), we see that it is increasing with the distance, pointing to an additional random error due to the collocation.

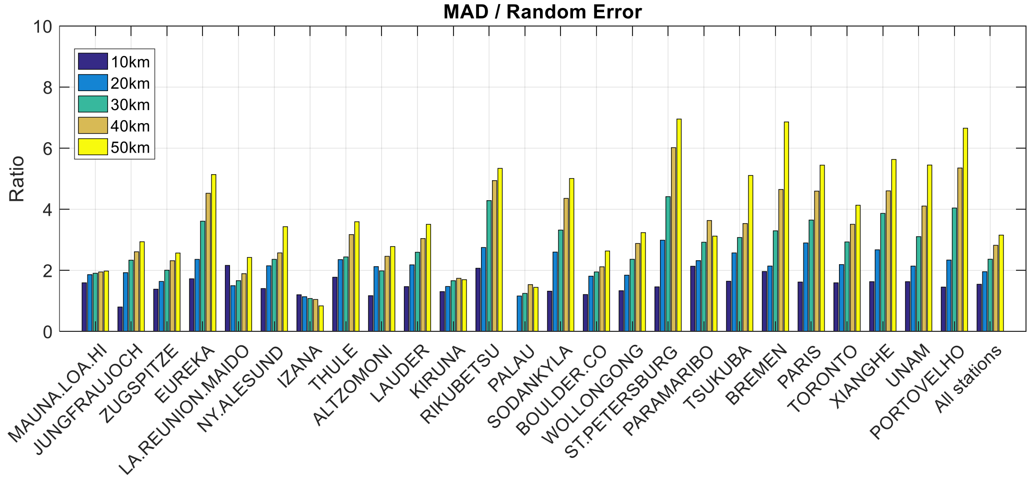

Figure 3: The ratio between the MAD and the random uncertainty on the differences between TROPOMI and FTIR.

As no clear threshold provides an improvement of the comparisons, we therefore decided to use the 20km collocation choice, a good compromise between the number of coincidences, and the best correspondence between MAD and random uncertainty budget. It also avoids an increasing MAD over the highest HCHO level sites (UNAM, Porto Velho).

In the new manuscript, we summarize this study by adding the following text:
“Before choosing the 20 km collocation criteria, we have tested several distances (10, 20, 30, 40, and 50 km). The 10 km criterion was discarded because of the poor remaining coincidences leading to less robust statistics. The 20 to 50 km criteria give similar biases between TROPOMI and FTIR. The standard deviations of the comparisons usually decrease slightly with increasing collocation distance due to a smaller TROPOMI random uncertainty (more pixels to average), except at the most polluted sites. However, the ratio between the standard deviations and the random uncertainty budgets is increasing with the collocation distance at all sites, pointing to an increased random error due to the collocation. We therefore choose the 20~km distance to reduce the random spatial collocation error.”

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

In Table 2 of the AMTD paper, we give σ_{rand} for each station. This value is the random uncertainty on the differences between TROPOMI and FTIR. It is fully defined by Eq. 7. In the text (Sect 4.3), we explain that since the other terms of Eq.7 are much smaller, σ_{rand} is dominated by the TROPOMI random error budget $\sigma_{S,rand}$. Therefore, the σ_{rand} in Table 2 is in first approximation the number that the referee is asking (~TROPOMI random uncertainty, $\sigma_{S,rand}$). Please note that there was an error in AMTD version in Eq. 7: the matrix for FTIR random uncertainty was called $S_{s,rand}$ instead of $S_{F,rand}$. It is now corrected.

Then, the detection limit is usually defined as 3*$\sigma_{S,rand}$, so it is easily determined at each station from Table 2, by approximating $\sigma_{S,rand}$ with the provided σ_{rand}, and multiplying by 3. For all stations together, we obtain $3.6\text{x}10^{15}$ molec/cm^2 as the TROPOMI detection limit (for an average of about 34 pixels), so $2.1\text{x}10^{16}$ molec/cm^2 for a single pixel.

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.

Actually, we said in the text above Eq. 2, that x'_F has been regridded to the satellite grid before applying Eq. 2. But the referee is right that this is not clear enough because we kept the same name for x'_F and $x_{S,a}$ in both grids (to try to have a small number of variable names). So, for clarity, we now introduce different names for different grids: we call now $x_{S,a}$ the S5P a priori on the original satellite grid and keep x'_F the FTIR profile on original FTIR grid, and we call $x_{S,a/F}$ the S5P a priori profile regridded to the FTIR grid, and $x'_{F/S}$ the FTIR profile regridded to the satellite grid.

The new text becomes:

"*First, the a priori substitution is applied, using the S5P a priori profile as the common a priori profile. For this, the S5P a priori profile* $\underline{x_{S,a}}$ *is regridded to the FTIR retrieval grid* $\underline{(x_{S,a/F})}$ *using a mass conservation algorithm (Langerock et al., 2015). In the rare situation where the satellite pixel elevation is above the FTIR site, the S5P a priori profile is extended to the FTIR instrument's altitude. The regridded S5P a priori* $\underline{x_{S,a/F}}$ *is then substituted following Rodgers and Connor (2003), and we finally use the corrected FTIR retrieved profile* x'_F *in the comparisons:*

$$x'_F=x_F+(A_F-I)(x_{F,a}-\underline{x_{S,a/F}}),$$

where …"

And also below:

"For that purpose we regrid the corrected FTIR profile x'_F *to the S5P column averaging kernel grid* $\underline{(x'_{F/S})}$ *and apply the smoothing equation:*

$$c_F^{smoo}=c_{S,a}+a_S(\underline{x'_{F/S}-x_{S,a}}) \qquad (2)$$

with $c_{S,a}$ *the S5P a priori column derived from the S5P a priori profile. We obtain a smoothed FTIR column* c_F^{smoo} *associated with a collocated TROPOMI pixel. In the case of mountain sites where the pixel altitude is below the instrument's height, the regridding of the FTIR profile* $\underline{x'_{F/S}}$ *is done…"*

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.

We followed the referee's suggestion by adding such a Table (now Table 1).

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 10^{15}..."This is my interpretation "...it is negative and very consistent for stations with higher levels, ranging from 8.7 to 28.6 x 10^{15}..." but maybe is the HCHO level what is 8.7 to 28.6 x 10^{15}.

Page 12, line 10: Lower levels are defined in the abstract and below at page 12, line 21 as 2.5x10^{15} molec/cm^2. What is the meaning of 6.5x10^{15} molec/cm^2.

We meant that the biases were always negative above 6.5x10^{15} (including Tsukuba and Bremen), and that they are consistent "only" above 8.7 x10^{15} (because the bias at Bremen, -5%, is lower). The 6.5x10^{15} limit was appearing in Table 2 (AMTD version) as a limit between positive/non significant trends (below) and always negative trends (above). However, because the limit of 8.0 x10^{15} was chosen for the "high levels" median bias calculation, we did not put a separation line at 6.5x10^{15}, which seems to be source of confusion. We decided to simplify the sentence as suggested by the referee.

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

This validation exercise could not identify a specific problem in the instrument itself or in the satellite retrieval algorithm. We will add the following text to the new manuscript (end of Sect. 5.1) in order to give some possible explanations to the observed biases (that are, however, in agreement with the systematic uncertainty budget).

The systematic uncertainties leading to the observed constant and proportional biases of our study have been calculated as described in Sect. 3 of De Smedt et al. (2018). From the error propagation of the HCHO TROPOMI columns (equation of N_v, in Sect. 2 of our AMTD paper, now numbered Eq.1 in the new manuscript), it can be found that the proportional bias is more likely due to air mass factor (M) uncertainties σ_M, while the constant bias is more likely due to the uncertainties of the slant columns uncertainties $\sigma_{N,S}$ and to the uncertainty of the background correction of the slant columns. This can be seen in Eq. 13 of De Smedt et al. (2018), where σ_M is proportional to N_s-$N_{s,0}$.

We can list some known difficulties of the satellite product:

- The negative bias over high HCHO levels sites (biomass burning or megacities) could be due to aerosol effects. There is no plan to include a correction for aerosols in the operational product, but specific studies are foreseen to check its impact in a scientific product.
- The positive bias over clean polar sites could be due to the solar zenith angle (SZA) dependency of the slant columns fit results (because of spectral interferences with ozone

and BrO). As explained in the paper, the QA values need to be improved at large SZA, which is foreseen in the next version.

- The current albedo climatology is too coarse for TROPOMI, which could be especially a problem for polar, mountain or coastal sites. A climatology based on TROPOMI measurements is under development.
- It is also foreseen to test a regional model at higher spatial resolution for an improvement of the a priori HCHO profiles. This should improve the TROPOMI retrieved product, especially at polluted sites. However, the validation presented here is already taking the a priori information and averaging kernels into account. We therefore do not expect an important effect of the improved a priori profiles on the validation results.

In the conclusion, we have added the following summary:

Possible improvements in the TROPOMI biases could be achieved by taking into account aerosol effects over polluted sites, improving the QA values at high SZA, and using an albedo climatology and a priori HCHO profiles at the TROPOMI spatial resolution. Except for the former, these improvements are foreseen in next versions of the operational TROPOMI data.

Reply to Anonymous Referee #2

General comments

Although, the main finding are very well described, my main concern with the paper is the missing discussion on the reasons of main difference between TROPOMI and FTIR formaldehyde BIAS for some stations (large offsets) and also difference in seasonal cycle (e.g. Paramaribo, Paris, UNAM…) (See Figure 5.).

The topic of this work fits well within the scope of AMT. Although the paper is well structured, the text needs to be carefully revised in order to be more precise in some sections. I recommend acceptance to AMT after addressing the comments above and few minor comments below.

We thank the referee for their work and useful comments.

We first answer on the main remarks above and reply then following the minor comments below.

We have added some possible reasons for the observed TROPOMI bias in the revised version of the manuscript. To avoid repetition, we refer to our reply to referee#1 (last page) who had the same concern as referee#2 on missing discussion on the observed biases.
It should be noted that even if the offsets are large, they are within the accuracy requirements of the satellite (which were based on previous validation studies of HCHO satellite measurements), meaning that such large biases were expected.

The TROPOMI and FTIR seasonal cycles are usually in agreement. However, as pointed out by the referee, this is not the case for Paramaribo. But, as can be seen in Fig.5, the sampling (number of coincidences) is bad there with often only one coincidence per month. Then, if TROPOMI has a remaining outlier, it has a strong influence on the plotted seasonal cycle (e.g. June 2019 shows a negative TROPOMI value). With an improved QA value as expected for the next TROPOMI versions, the comparisons should also improve. For Paris, UNAM (Mexico City) and usually all polluted sites, the TROPOMI and FTIR seasonal cycles show similar features, but the amplitude is smaller with TROPOMI due to its proportional bias that leads to more under-estimation for high HCHO levels (so more under-estimation during the maximum of the FTIR seasonal cycle).

Page 2, line 5, confusing sentence, "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", it does not make sense to write that HCHO TROPOMI retrievals are below lower and upper limits. Please clarify it.

We have clarified the text:
"The pre-launch requirements of the TROPOMI HCHO accuracy are 40-80%. We observe that these requirements are well reached, with the BIAS found below 80% at all the sites, and below 40% at 20 of the 25 sites."

Page 3, line 1, is there any study of validation of satellite HCHO observation with ship-based measurements?

Indeed. We have added two references as example of such studies (Peters et al., 2012; Tan et al., 2018).

Page 3, line 8, please define what is "TROPOMI Cal/Val"

Done.

Page 3, line 15, would you please mention what are the differences among versions from v.1.1.5 to v.1.1.7?

In the AMTD paper, we referred to the ReadMe file (https://sentinel.esa.int/documents/247904/3541451/Sentinel-5P-Formaldehyde-Readme.pdf) for details on the differences in the versions because they have minor impacts on the HCHO TROPOMI time-series. However, as both referees ask that all is included in our paper, we have included a Table (Table 1 in the updated version) repeating the information about the different versions (dates and changes). (see also reply to referee#1)

Page 4, line 12, why to use OMI albedo climatology?

The OMI albedo climatology is the best product existing at 340 nm. The spatial resolution is indeed too coarse for TROPOMI. We are waiting for a climatology based directly on TROPOMI, but it is not yet available.

Page 4, line 13, "(Kleippol et al., 2008)".

Done (changed to Kleipool).

Page 4, line 20, please define all the quantities of the equation (e.g., M and M0)

All quantities have been defined in the text above the equation, except M0, which is an average of the air mass factors (M) of the slant columns selected in the reference sector, the Pacific Ocean (N(s,0)). We have added its definition in the new manuscript.

Page 6, line 6, what is the main difference between PROFITT9 and SFIT4.0.9.4?

Both codes are very similar. They are both line-by-line models for infrared solar transmittance spectra, including a radiative transfer model (FSCATM and KOPRA for SFIT4 and PROFITT9, respectively), and based on the optimal estimation method (Rodgers, 2000). They both allow for the Tikhonov regularization as well. Differences are minor, mainly lying in the different options that are available (but not used in the present work) in PROFFIT9: e.g. possibility to retrieve the temperature profiles,… The only relevant difference for the present work is the calculation of channeling error that is not yet included in SFIT4.
The use of different codes within the InfraRed Working Group (IRWG) of NDACC is historical. To certify a good homogenization in the delivered FTIR products in the NDACC database,

harmonization in the retrieved parameters (spectral micro-windows, a priori profiles, spectroscopic database,…) is required for all NDACC target species, and has also been done for the HCHO products presented here (Vigouroux et al., AMT, 2018). A comparison exercise of the two codes has been performed for four species (Hase, JQSRT, 2004) and an agreement within 1% has been found in the retrieved columns. Further details on both codes are available in this latter reference, which has been added in the next version of our manuscript.

Page 7, line 7, please be consistent between names used in the text "Maïdo" and used in the figure 1.

The figures are automatically generated using the name provided by the PIs in the geoms file. Indeed, it might confuse the reader to see two names, maybe not so much for Maïdo / LA.REUNION.MAIDO; but for Mexico City (used in Tables and text) and UNAM (used in automated figures). To help the reader, we have explicitly added the two possible names in Table 1 (Table 2 in the new version). We prefer to keep using both names because a geoms data user would find the "UNAM" name for the station, while a "Mexico City" name makes the information clearer for a simple reader that the station is in the city center of Mexico. Note that this situation is also there for other stations, but with clear signification (Izaña / IZANA; Mauna Loa / MAUNA.LOA.HI;…).
Because, the correspondence is less clear for Mexico City / UNAM, we have also repeated the two names in Table 2 (Table 3 in the new version).

Page 8, line 3,what are the reasons for the lowest smoothing systematic uncertainties in the 5 added sites.

The provided 3.4% number for median smoothing systematic uncertainty is the one given in Vigouroux et al., AMT, 2018. The 13% and 14% for the 5 added sites, are for the total systematic uncertainty (dominated by the spectroscopy), and not for the smoothing part only. To avoid the confusion, we have changed the sentence to:
"For the five added sites, the median total systematic uncertainty is 13% (Jungfraujoch, Tsukuba, Palau), or 14% (Rikubetsu, Xianghe), commensurate with the other sites."

Page 8, line 25, please remove "so"

Done.

Page 11, line 29, would be nice if you include one or two sentences describing the main differences between OFFL, RPRO and NRTI products. Are they different at all.

OFFL, RPRO and NRTI share the same algorithm (for the versions used in the paper). Changes of version numbers refer to changes in other components of the operational processor. However, slight differences come from auxiliary data. A priori profiles used for NRTI are from TM5 forecast model, while they are from TM5 analysis for OFFL/RPRO (this makes almost no difference since HCHO is not assimilated).
For the reprocessing (RPRO), data have been processed using 7-days parallelization (in order to speed up the reprocessing). It means that the slant columns used for the background correction are always at least 7 days older, while for OFFL and NRT, the gap is only 1 day. It results in stripes slightly more pronounced in the RPRO product than in the other versions.

We do not give these details on the OFFL / RPRO / NRTI data because they have negligible impact on the satellite data and validation results. But we have added a Table (Table 1 in the new version) with the date of the different versions, and all details can be found in the Readme file given as a reference in the manuscript.

Page 16, line 31, would you please clarify how the collocation plays a role in Maïdo? Fire emissions are included in the calculation of the a-priori profiles? Could fire emissions enhanced the HCHO amounts? What is the effect of changing the collocation radius in this station?

Maïdo is usually a clean site. During the biomass burning period, some plumes (mainly coming from Madagascar for the short lifetime species HCHO) can cross over Reunion Island. Since the overestimation of TROPOMI at Maïdo is larger during the biomass burning months, we suggest that this could be due to plumes that would be present in the 20km circle around the station covered by TROPOMI but not in the line of sight of the FTIR measurements for the collocated days. Unfortunately the collocation effect could not be confirmed at Maïdo: the 10km radius criterion lead to very few coincidences at Maïdo (see Fig. 1 of our Reply to referee#1), and none are during the biomass burning season.
This was only a suggestion from our side (we wrote “the collocation of the plumes *might* play a role there”), and could be investigated when more data are available in a future work.